# Table of Contents

1. **Introduction.** ................................................................. 1  
   1. Overview. ................................................................. 1  
   2. General notation. ...................................................... 2  
   3. Functions and relations. ............................................. 3  
   4. The number systems. .................................................. 4  
   5. Set theory. .............................................................. 5  

2. **Universal algebra.** ...................................................... 8  

3. **Orders.** ........................................................................ 15  

4. **Groups.** ....................................................................... 21  
   1. Basic definitions. ......................................................... 21  
   2. Normal Subgroups. ....................................................... 23  
   3. An example of a group. ............................................... 23  
   4. Homomorphism theorems. .......................................... 24  
   5. Product of groups. ...................................................... 25  

5. **Permutation groups.** .................................................... 27  

6. **Rings.** .......................................................................... 31  
   1. Basic definitions. ......................................................... 31  
   2. Ideals. ................................................................. 31  
   3. Commutative rings. .................................................... 32  
   4. Fractions. ............................................................... 35  
   5. Ordered commutative rings. .................................. 37  
   6. Chinese remainder theorem. ..................................... 37  
   7. Möbius inversion. ...................................................... 38  

7. **Polynomials.** ............................................................. 42  
   1. Single variable polynomials. .................................... 42  
   2. Interpolation. ........................................................... 45  
   3. Multi-variable polynomials. .................................... 47  

8. **Modules.** ..................................................................... 52  
   1. Basic definitions. ......................................................... 52  
   2. Multilinear maps. ....................................................... 54  
   3. Algebras. ............................................................... 55  
   4. Generation of modules. ........................................... 55  
   5. Noetherian modules. ................................................ 56  
   6. Module of fractions. ................................................ 56  
   7. Linear algebra of modules. .................................... 57  
   8. Modules over principal ideal domains. ..................... 58  

9. **Fields.** ......................................................................... 62  
   1. Basic definitions. ......................................................... 62  
   2. Algebraic extensions. ................................................ 62  
   3. Splitting fields. ........................................................ 63  
   4. Galois extensions. ...................................................... 63  
   5. Galois theory. .......................................................... 64  
   6. Finite fields. ............................................................. 65  

10. **Linear algebra.** .......................................................... 68
1. Basic definitions. ................................................................. 68
2. Dual space. ........................................................................... 70
3. Representations. ................................................................. 71
4. Rational canonical form. ..................................................... 71
5. Jordan canonical form. ......................................................... 72
6. Invariants. ............................................................................. 74
7. Bilinear forms. ...................................................................... 74
8. Orthogonal sums. ................................................................. 76
9. Witt’s theorem. ...................................................................... 78
10. Forms and operators. .......................................................... 79
11. Projection operators. ........................................................... 81
12. Factorizations. ..................................................................... 82
13. Compound matrices. ........................................................... 85

11. Model theory. ....................................................................... 88
1. Syntax and semantics. ........................................................... 88
2. Validity. ................................................................................ 91
3. Basic definitions. ................................................................. 95
4. Ideals and filters. ................................................................. 97
5. Ultraproducts. ...................................................................... 99

12. Computability. .................................................................... 101
1. Introduction. ......................................................................... 101
2. Equivalence of three models. ................................................. 102
3. Basic definitions. ................................................................. 105
4. Undecidability. ..................................................................... 106
5. Subrecursion. ....................................................................... 107
6. Decidability of theories. ....................................................... 110
7. NP completeness of satisfiability. ......................................... 117
8. Hilbert’s tenth problem. ....................................................... 118
9. Word problem for groups. ................................................... 122
10. Details for goto step simulation. ......................................... 123
11. Details for Turing machine step simulation. ......................... 124
12. Details for a universal Turing machine. ............................... 124
13. Proof sketch for theorem 13. .............................................. 124

13. Category theory. ................................................................. 126
1. Basic definitions. ................................................................. 126
2. Natural transformations. ...................................................... 129
3. Universal arrows. ............................................................... 130
4. Adjoint functors. ............................................................... 132
5. Limits. ................................................................................ 135
6. Monics and Epics. ............................................................. 139
7. Preservation of limits. ......................................................... 140
8. Galois adjunctions. ........................................................... 141
9. Normals and conormals. .................................................... 142
10. Images and coinages. ......................................................... 142
11. Biproducts. ........................................................................ 144
12. Abelian categories. ............................................................ 144
20. **Further topics in rings and fields.** .......................................................... 239
   1. Algebraic closure. ....................................................................................... 239
   2. Transcendence bases. ................................................................................ 239
   3. Local rings. .................................................................................................. 240
   4. Absolute values. .......................................................................................... 241
   5. Valuations. .................................................................................................... 243
   6. Integral extensions. ....................................................................................... 244
   7. Ring homomorphisms and ideals. ............................................................... 246
   8. The radical of an ideal. ............................................................................... 248
   9. Primary decomposition. .............................................................................. 249
  10. The Nullstellensatz. ..................................................................................... 250
  11. Dedekind domains. ..................................................................................... 251
  12. Complete fields. .......................................................................................... 253
  13. Real algebras. .............................................................................................. 255
  14. Lagrange’s four square theorem. ............................................................... 258

21. **Lattice orders.** .......................................................................................... 262
   1. Varieties of lattices. ..................................................................................... 262
   2. Complete lattices. ....................................................................................... 266
   3. Algebraic lattices. ...................................................................................... 268
   4. Stone Duality. ............................................................................................ 270
   5. Continuous lattices. .................................................................................... 275
   6. Exponential correspondence. ...................................................................... 279

22. **Convexity.** ............................................................................................... 283
   1. Basic facts. .................................................................................................. 283
   2. Polyhedra and linear inequalities. ............................................................... 286
   3. Linear programming. ............................................................................... 288
   4. Polytopes. ................................................................................................... 291
   5. Krein-Rutman theorem in Euclidean space. ............................................ 292

23. **Point lattices.** .......................................................................................... 295
   1. Basic facts. .................................................................................................. 295
   2. Lebesgue measure. ...................................................................................... 296
   3. Minkowski’s theorem. ............................................................................... 299
   4. Hermite and Smith normal form. ............................................................... 300
   5. Reduced bases. .......................................................................................... 303
   6. Root systems. ............................................................................................. 304
   7. The wallpaper groups. ............................................................................... 311
   8. Continued Fractions. .................................................................................. 314

24. **Algebra and Topology.** ............................................................................. 319
   1. Topological structures. ............................................................................... 319
   2. Topological groups. .................................................................................... 321
   3. Topological vector spaces. .......................................................................... 322
4. Additional facts about metric spaces. .................................................. 324
5. Normed linear spaces. ................................................................. 325
6. Complete normed linear spaces. ..................................................... 329
7. Inner product spaces. ................................................................. 332
8. Hilbert spaces. ........................................................................... 333
9. Bundles. ......................................................................................... 334
10. Sheaves. ......................................................................................... 338
11. General manifolds. ......................................................................... 340
12. Tensors on manifolds. ................................................................. 341
13. Homotopy. ....................................................................................... 342
14. Homology of simplicial complexes. ................................................ 345

25. Algebraic geometry. ......................................................................... 355
1. Introduction. ..................................................................................... 355
2. Noetherian topological spaces. ........................................................ 355
3. Affine algebraic sets. ......................................................................... 357
4. Noether normalization. ....................................................................... 360
5. Dimension. ......................................................................................... 360
6. Projective algebraic sets. .................................................................... 363
7. Hilbert functions. ................................................................................ 366
8. Bezout’s theorem. ............................................................................... 369
9. Prime spectrum of a ring. ................................................................... 370
10. Schemes. .......................................................................................... 371
11. Groebner bases. ............................................................................... 373

26. Algebraic number theory. ............................................................... 378
1. Basic facts. ......................................................................................... 378
2. Discriminant. ...................................................................................... 379
3. Extensions of real valuations. ............................................................ 380
4. Valuations on Dedekind Domains. ..................................................... 382
5. Decomposition and inertia groups. .................................................... 384
6. Ideal norm. ......................................................................................... 385
7. Ideal class group. ............................................................................... 386
8. Gauss’ lemma. ................................................................................... 387

27. Lie groups. ......................................................................................... 389
1. Definition. .......................................................................................... 389
2. Differentiation on manifolds. ............................................................ 390
3. Zariski closed subgroups of GL(n). ................................................... 391
4. Lie algebras. ....................................................................................... 392
5. Lie algebra of a Lie group. ............................................................... 394
6. Covering projections. ......................................................................... 396

28. Some algorithms. ................................................................................ 400
1. Gaussian elimination. ......................................................................... 400
2. Coset enumeration. ............................................................................ 400
3. The Schreier-Sims algorithm. ............................................................ 402
4. Decision procedure for Presburger arithmetic. ................................. 403
5. Basic algorithms for polynomials. ...................................................... 405
6. Completeness of real closed fields. .................................................... 409
1. Introduction.

1. Overview. Abstract algebra is a subject whose scope and applicability continue to increase. Physical theories and engineering methods have come to make use of concepts from algebra to simplify their developments, and in some cases make essential use of them. Algebra has become indispensable throughout mathematics, and an increasingly rich subject in its own right. In the last twenty years, algebra has become increasingly relevant to computer science.

This text is intended as an introduction to abstract algebra for advanced undergraduates. Much of it is accessible to advanced third year students, and the first nine chapters can be covered even earlier. There are many introductions to abstract algebra, so a new one should have some distinguishing characteristics. The main distinguishing characteristics of this text are the coverage of universal algebra and category theory, and their use in the introductory presentation of topics in algebra; and the coverage of various topics outside the main line of classical abstract algebra, which are of interest. These include topological spaces, model theory, computability theory, point lattices, and some algorithms.

The prerequisites of the text are high school mathematics, and a willingness to read. Familiarity with the usual “fundamentals”, namely sets, relations, functions, the axiomatic method, and fundamental properties of the integers and rational, real, and complex numbers is assumed. Occasionally a theorem from elementary calculus will be needed, including basic properties of infinite sequences and series. Various background facts will be summarized below, following a description of the contents.

Chapter 2 presents the basic definitions and theorems of universal algebra. This “quantifies” the observation that the basic theories of the basic structures considered in abstract algebra involve various analogous facts. Indeed, basic facts about a particular type of structure can be proved as corollaries of facts proved in chapter 2. This serves to better organize the material, and make its presentation shorter. Further, the student is introduced as soon as possible to unifying themes that have been discovered among the branches of mathematics.

Chapter 3 is also a preliminary chapter, covering the basic theory of orders. In addition to its use throughout mathematics, the study of orders is a branch in its own right. Basic facts about algebraic structures can be better understood in terms of concepts from order theory.

Chapters 4 to 9 cover the classic introductory topics of abstract algebra, namely groups, permutation groups, rings, polynomials, modules, and fields. Results from chapters 2 and 3 are used to put various facts in a general perspective.

Chapter 10 begins the “second half” of the text, which attempts to build on the first 9 chapters to cover further topics in abstract algebra in an advanced undergraduate course. It presents a comprehensive introduction to linear algebra.

Chapter 11 covers model theory. This has various uses in algebra, indeed a “type of algebraic structure” is usually the models of some set of axioms. The material covered is sufficient to consider many of these uses. For similar reasons, chapter 12 covers basic computability theory; this allows giving formal discussions not only of decidability, but also of various algorithms which have been devised in the last twenty-five years by algebraists and computer scientists. Hilbert’s tenth problem is also covered.
Chapter 13 gives a thorough treatment of basic category theory, which has been the medium of much unification in mathematical thought since its introduction in the 1940’s. Students of algebra should master basic category theory as early as possible, and chapter 13 is intended to facilitate this. Later chapters assume that the student has mastered it. Material covered includes the basic facts concerning Abelian categories.

Chapters 14 to 16 return to traditional topics in abstract algebra, covering a variety of topics including solvability and representation of groups, and general topics such as decomposition in modular lattices.

Chapter 17 covers topological spaces. Topology and algebra are intertwined throughout mathematics, and a self-contained introduction to topology is useful in an algebra text. Using category theory, a robust introduction can be kept brief.

Chapters 18 and 19 cover tensor algebra and homology, adding to the student’s mastery of category theoretic machinery. Chapter 20 covers further topics in rings and fields. Some basic topics from “commutative algebra” are covered, for example the Nullstellensatz. Chapters 21 to 23 cover topics which are not usually included in introductions to algebra, and consequently not introduced as early as their utility warrants, namely lattice orders, convexity, and point lattices.

Chapters 24 to 26 provide brief introductions to topics in advanced branches of mathematics (algebra and topology, algebraic geometry, algebraic number theory), giving examples of how basic algebra is used. Chapter 27 gives an introduction to Lie group theory, which has applications throughout the sciences. Finally, Chapter 28 contains some algorithms for solving computational problems in algebra.

Appendix 1 covers some topics from basic set theory needed for some proofs. Appendix 2 covers the basic linear algebra of finite dimensional vector spaces over a field, including determinants; this is assumed in chapter 10, and sometimes in earlier sections. Appendix 3 covers basic facts about graphs, which are occasionally useful in algebra.

This book was typeset using the “tetex” release of TeX, which includes various files such as eplain, psfig, and AMS fonts. There are various online books which cover material as in chapters 19, 20, 24, and 26, and other chapters, [Ash] and [Milne] for example.

2. General notation. The abbreviations “iff” for “if and only if”, and “w.l.g.” for “without loss of generality” are used in the text. The propositional connectives $\neg, \land, \lor, \Rightarrow$ (not, and, or, implies) are used; and the quantifiers $\forall, \exists$ (for all, there exists). These are discussed in chapter 11. The notation used for sets is as follows.

- $x \in S$ denotes that $x$ is an element of the set $S$.
- $x \notin S$ denotes that $x$ is not an element of $S$.
- $\{x : \ldots x\ldots\}$ denotes the set of $x$ such that $\ldots x\ldots$ is true; this has some common variations, such as $\{x : x \in S : \ldots x\ldots\}$ to denote $\{x : x \in S \land \ldots x\ldots\}$, or $\{f(x) : \ldots x\ldots\}$ to denote $\{y : \exists x(y = f(x) \land \ldots x\ldots)\}$.
- $S = T$ denotes that the sets $S$ and $T$ are equal, that is, contain the same elements.
- $\emptyset$ denotes the empty set, the set containing no elements. A set is called empty or nonempty according to whether it equals $\emptyset$.
- $S \subseteq T$, or $T \supseteq S$, denotes that $x \in S \implies x \in T$; we say $S$ is a subset of $T$, or $T$ is a superset of $S$.
- $S \subset T$, or $T \supset S$, denotes that $S \subseteq T$ but $S \neq T$; we say $S$ is a proper subset of $T$.
- $S \cup T$ denotes $\{x : x \in S \lor x \in T\}$, the union of $S$ and $T$. More generally if $C$ is a collection of sets, $\bigcup C$ denotes $\{x : \exists S \in C : x \in S\}$; note that $\bigcup \emptyset = \emptyset$.
- $S \cap T$ denotes $\{x : x \in S \land x \in T\}$, the intersection of $S$ and $T$. More generally if $C$ is a nonempty collection of sets, $\bigcap C$ denotes $\{x : \forall S \in C : x \in S\}$. If there is a “universe” $U$ such that for all sets under consideration $S \subseteq U$, then $\bigcap \emptyset$ can be taken as $U$. 

2
- If there is a universe $U$ such that for all sets under consideration $S \subseteq U$, then the complement $S^c$ of $S$ equals $\{x \in U : x \notin S\}$.

- Sets $S$ and $T$ are called disjoint if $S \cap T = \emptyset$.

A collection of sets may be denoted $\{S_\alpha \}$ where $\alpha$ ranges over an “index set”; this is sometimes more convenient than using $\{S\}$ to denote it. As an example of its use, given $\{S_\alpha \}$ the disjoint union is defined to be the set of pairs $(\alpha, x)$ such that $x \in S_\alpha$. Thus, each element in the union is “tagged” by the set from which it came (by the index of the set, although the set itself could equally well be used).

An ordered $n$-tuple $(x_1, \ldots, x_n)$ is a list of $n$ elements, in order, with repetitions allowed; $x_i$ is called the $i$th component. Two ordered $n$-tuples are equal exactly if their $i$th components are equal for all $i$. Given sets $S_1, \ldots, S_n$, the set $(\{x_1, \ldots, x_n\} : x_i \in S_i, 1 \leq i \leq n)$ is called the Cartesian product of $S_1, \ldots, S_n$, and denoted $S_1 \times \cdots \times S_n$. If the $S_i$ are all the same set $S$ we write $S^n$ for the Cartesian product.

3. Functions and relations. An $n$-ary relation or predicate on a set $S$ is a subset of $S^n$. If $R$ is an $n$-ary relation we write $R(x_1, \ldots, x_n)$ as a synonym for $(x_1, \ldots, x_n) \in R$. An $n$-ary function from $S$ to $T$ is a subset $f \subseteq S^n \times T$ such that for each $(x_1, \ldots, x_n) \in S^n$ there is a unique $y$ such that $(x_1, \ldots, x_n, y) \in f$; we write $f(x_1, \ldots, x_n) = y$ for this $y$, which is called the value of $f$ at $x_1, \ldots, x_n$.

Although set-theoretically an $n$-ary function is always a set of $n+1$-tuples, to emphasize the fact the set of pairs of a binary function is sometimes called the “graph” of the function.

Infix notation may be used for binary relations and functions, that is, we may write $xfy$ or $xRy$ rather than $f(x, y)$ or $R(x, y)$. Parentheses may be used to indicate the order of evaluation of complex expressions involving infix function symbols. We mention one other point regarding symbols. The same symbol is often convenient than using $\sigma$ for the Cartesian product.

The notation $f : S \rightarrow T$ denotes that $f$ is a function from $S$ to $T$; $S$ is called the domain of $f$, and $T$ the codomain. Note that an $n$-ary function from $S$ to $T$ is essentially the same as a function from $S^n$ to $T$. For $S' \subseteq S$ $f[S']$ denotes $\{y \in T : f(x) = y \text{ for some } x \in S'\}$; $f[S']$ is called the image of $S'$, and the image of $S$ is called simply the image, or the range. The notation $f^{-1}[T']$ for $T' \subseteq T$ is used to denote $\{x \in S : f(x) \in T'\}$, which is called the inverse image of $T'$. We assume familiarity with the properties of these operations, which are given in exercise 1.

The arrow notation may be used to avoid giving a name to a function. We may write $x \mapsto E$ where $E$ is some expression involving $x$, to denote the function $f$ where $f(x) = E$ for all $x$.

Another notation for application of functions, the “exponential” notation, is occasionally used in the text. If $\sigma : S \rightarrow T$ and $x \in S$, $x^\sigma$ may be used to denote $\sigma(x)$; and for $R \subseteq S R^\sigma$ denotes $\sigma[R]$.

Functions or relations are said to be equal if they are equal as sets. If $f : S \rightarrow T$ and $g : T \rightarrow U$ their composition $g \circ f : S \rightarrow U$ is the function such that $g \circ f(x) = g(f(x))$ for all $x \in S$. The requirement that the codomain of $f$ be given and equal the domain of $g$ can be relaxed; $f[S] \subseteq T$ suffices to define the composition. In algebra the codomain of a function is often given; further remarks on this point are made in chapter 13. Composition of functions is associative; that is, if $h : S_1 \rightarrow S_2$, $g : S_2 \rightarrow S_3$, and $f : S_3 \rightarrow S_4$, then $f \circ (g \circ h) = (f \circ g) \circ h$. The identity function $\iota_S : S \rightarrow S$ on $S$ is defined by $\iota_S(x) = x$ for all $x \in S$. Clearly $f \circ \iota_S = \iota_T \circ f = f$.

The function $f : S \rightarrow T$ is called injective if $f(x) = f(y)$ only if $x = y$; surjective if its image is $T$; and bijective if it is both injective and surjective. The composition of injections is an injection; of surjections a surjection; and of bijections a bijection. A bijection is also called a one to one correspondence.

**Theorem 1.** Suppose $f : S \rightarrow T$; then the following are true.
a. $f$ is injective iff $S = \emptyset$ or there is a function $f^L : T \to S$ such that $f^L \circ f = \iota_S$; such an $f^L$ is called a left inverse of $f$.

b. $f$ is surjective iff there is a function $f^R : T \to S$ such that $f \circ f^R = \iota_T$; such an $f^R$ is called a right inverse of $f$.

c. $f$ is bijective iff there is a function $f^{-1} : T \to S$ such that $f^{-1} \circ f = \iota_S$ and $f \circ f^{-1} = \iota_T$. In this case there is a unique such $f^{-1}$; it is a bijection, and is called the inverse of $f$.

**Proof:** For part a, suppose $S \neq \emptyset$. If $f^L$ exists, suppose $f(x) = f(y)$ and apply $f^L$ to both sides; this shows $f$ is injective. If $f$ is injective, for each $y$ such that $f(x) = y$ for some $x$ let $f^L(y) = x$ and define $f^L(y)$ arbitrarily otherwise. For part b, if $f^R$ exists, consider $f^R(y)$ for $y \in Y$; this shows $f$ is surjective. If $f$ is surjective, one may choose an arbitrary $x$ with $f(x) = y$ for $f^R(y)$. To prove part c, observe that

$$f^L = f^L \circ \iota_T = f^L \circ f \circ f^R = \iota_S \circ f^R = f^R$$

if $f^L, f^R$ are left and right inverses; in particular the inverse is unique (the case $S = \emptyset$ follows also).

Suppose $S' \subseteq S$, $T' \subseteq T$. Given a function $f : S \to T$, its restriction to $S'$ is defined to be $f \cap (S' \times T)$; we denote this $f \mid S'$. If $f' : S' \to T$ is given and equals $f \mid S'$ we call $f$ an extension to $S$ of $f'$. If $f[S] \subseteq T'$ we call $f \cap (S \times T')$ the corestriction of $f$ to $T'$.

A partial function from $S$ to $T$ is a set of pairs such that for each $s \in S$ there is at most one pair whose first element is $s$. It is thus a function (possibly empty) $f : S' \to T$ where $S' \subseteq T$; $S'$ is the domain of $f$. Considered as sets of pairs, the partial functions $\phi : S \to T$ are partially ordered by inclusion; $\phi_1 \subseteq \phi_2$ iff $\operatorname{Dom}(\phi_1) \subseteq \operatorname{Dom}(\phi_2)$, and for $x \in \operatorname{Dom}(\phi_1)$ $\phi_2(x) = \phi_1(x)$. This order on the partial functions is called the approximation order. The union of a chain of partial functions is readily verified to be a partial function.

The characteristic function of an $n$-ary relation $R$ is the $n$-ary function whose value is 1 if $R$ is true, and 0 if $R$ is false. The codomain might be, for example, a ring where $0 \neq 1$. A useful binary function on $S$ is the The delta function, $\delta_S(x, y)$, or $\delta(x, y)$ if $S$ is clear, which equals 1 if $x = y$, else 0. This is the characteristic function of the equality relation on $S$. It is also customarily denoted $\delta_{xy}$, and called the “Kronecker delta function”.

The transpose $R^t$ of a binary relation $R$ is defined to be the binary relation such that $R^t(x, y)$ iff $R(y, x)$. A binary relation $R$ on a set $S$ is called

- reflexive if $xRx$;
- symmetric if $yRx$ whenever $xRy$;
- transitive if $xRz$ whenever $xRy$ and $yRz$;
- an equivalence relation if it is reflexive, symmetric, and transitive.

If $\equiv$ is an equivalence relation the set $\{y : y \equiv x\}$ is called the equivalence class of $x$, and denoted $[x]$.

A partition of a nonempty set $S$ is a collection of nonempty subsets $P_\alpha$, called the parts of the partition, such that $S = \cup_\alpha P_\alpha$ and $P_\alpha, P_\beta$ are disjoint for all $\alpha \neq \beta$.

**Theorem 2.** Suppose $S$ is nonempty. The equivalence classes of an equivalence relation on $S$ are the parts of a partition of $S$. Conversely given a partition of $S$, the relation of belonging to the same part is an equivalence relation. These maps between equivalence relations and partitions are inverse to each other.

**Proof:** Exercise.

Given an equivalence relation on a set $S$, or equivalently a partition of $S$, by a system of representatives is meant a subset $R \subseteq S$ such that $R$ contains exactly one element from each equivalence class, or part of the partition.

**4. The number systems.** We use $\mathcal{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ to denote the natural numbers or nonnegative integers, integers, rational numbers, real numbers, and complex numbers respectively. We assume familiarity
with the basic properties of these structures. We also use $\mathbb{R}^\neq$ for the nonzero reals, $\mathbb{R}^\geq$ for the nonnegative reals, and so on. The notation $R^\neq$ may be used in any ring (rings are defined in chapter 6; both $\mathbb{Z}$ and $\mathbb{R}$ are rings, so the notation $\mathbb{Z}^\neq$ is another example). The notation $R^*$ is often used, especially for a field, but the asterisk has so many uses in algebra that this notation is less confusing. $\mathbb{R}^\geq$ may be used in any ordered commutative ring.

The natural numbers have the property that any nonempty set of natural numbers contains a least element. A consequence of this is the principle of mathematical induction, which states that if $S \subseteq N$, $0 \in S$, and $n+1 \in S$ whenever $n \in S$, then $S = N$. One may strengthen the induction hypothesis “whenever $n \in S$” to “whenever $k \in S$ for $k \leq n$”. Another way of stating induction is, if $n \in S$ whenever $k \in S$ for $k < n$ then $S = N$, since for $n = 0$ the hypothesis is vacuously true.

A function $f : N \times S \rightarrow S$ may be defined “recursively” from functions $g : S \rightarrow S$ and $h : N \times S \times S \rightarrow S$ by the equations $f(0, x) = g(x)$, $f(n+1, x) = h(n, x, f(x, n))$, in the sense that there is a unique such function satisfying these equations. This is a fact of elementary set theory, and we will assume it.

A fundamental fact about $\mathbb{Z}$ is the division law, which states that for $p, d \in \mathbb{Z}$, $d > 0$, there are unique integers $q, r$, the quotient and remainder respectively, such that $p = qd + r$ and $0 \leq r < d$. If $r = 0$ $d$ is said to divide $p$, for which we write $d|p$. The remainder when $p$ is divided by $d$ will be denoted $p \mod d$. The division law can be proved from a suitable set of axioms for $\mathbb{Z}$.

The factorial function is that mapping the nonnegative integer $n$ to $n! = 1 \cdot 2 \cdots n$. In a ring we let the empty product equal 1; thus $0! = 1$. The quantities
\[
\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots1}
\]
for nonnegative integers $n, k$ with $0 \leq k \leq n$ are called the binomial coefficients. These are readily verified to satisfy
\[
\binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}
\]
where $0 < k < n$. From this, the binomial coefficients are readily seen to be positive integers.

A set $S$ is called finite if there is a bijection from $S$ to \{1, 2, …, $n$\} for some $n$; in this case $n$ is unique, and is called the cardinality of $S$, and denoted $|S|$. Given finite sets $S$ and $T$, $|S| \leq |T|$ iff there is an injection from $S$ to $T$.

The real numbers are sometimes extended with symbols $\pm \infty$, which obey some algebraic rules as follows, where $r$ denotes an ordinary real number.
- $-\infty < r < +\infty$.
- $r + \pm \infty = \pm \infty + r = \pm \infty$.
- if $r > 0$ then $r \cdot \pm \infty = \pm \infty \cdot r = \pm \infty$.
- if $r < 0$ then $r \cdot \pm \infty = \pm \infty \cdot r = \mp \infty$.
- $\pm \infty + \pm \infty = \pm \infty$.
- $\pm \infty \cdot \pm \infty = \mp \infty \cdot \pm \infty = -\infty$.

Other algebraic structures are sometimes extended in this way, such as an arbitrary field; in some cases only one of the two infinities is needed.

The “floor” function $\lfloor x \rfloor$ mapping $\mathbb{R}$ to $\mathbb{Z}$ maps $x$ to the largest integer $n$ such that $n \leq x$. The “ceiling” function $\lceil x \rceil$ maps $x$ to the smallest integer $n$ such that $n \geq x$. These functions obey various identities which follow immediately from the definitions; see [Knuth] for some of these.

5. Set theory. A set is called infinite if it is not finite. In algebra many facts about finite sets can be generalized to infinite sets. Introductory texts often omit the infinite case, to avoid appealing to the
necessary set theory. However, this can be kept to a minimum, and we will frequently give the infinite generalizations, assuming the necessary facts from set theory. Some discussion of these is given here, and further discussion in appendix 1.

The notion of an infinite Cartesian product is an example. If $I$ is a set, we can consider the notion of a sequence $(x_i)$ of elements from some set $S_i$ indexed by (the index set) $I$; $x_i$ is called the $i$th component. Set theoretically this is just a function from $I$ to $S$. Given a set $S_i$ for each $i \in I$, the Cartesian product $\times_i S_i$ of the sets is the collection of sequences $(x_i)$ where $x_i \in S_i$. Set theoretically, this is the collection of functions $f : I \rightarrow \bigcup_i S_i$ such that $f(i) \in S_i$ for all $i \in I$. If $S_i = S$ for all $i$ we write $S^I$ for the Cartesian product, or what is the same thing the set of functions from $I$ to $S$. An $n$-tuple is the special case where $I$ has $n$ elements. Technically, the earlier characterization is not the same thing as a sequence on an $n$-element index set, but there is an obvious correspondence which can almost always be ignored.

It is intuitively clear that the Cartesian product of nonempty sets is nonempty. This may be shown using the principle of higher set theory known as the axiom of choice; in fact it is equivalent to it. The axiom of choice states that given any collection of nonempty sets there is a function (from the collection to its union) which selects one member from each set.

The collection of subsets of a set $S$ is called the power set of $S$; in this text it is denoted $\text{Pow}(S)$. Collections of subsets of $S$ are often considered; these are just subsets of $\text{Pow}(S)$.

Let $S$ be a collection of subsets of some set. A maximal set in $S$ is one which is not properly contained in any other set in the collection. A chain in $S$ is a collection $C$ of sets from $S$, such that for any $A$ and $B$ in $C$, either $A \subseteq B$ or $B \subseteq A$. $S$ is said to be inductive if for every chain $C$ in $S$ there is a set in $S$ which contains every element of $C$ (often in applications this is the union of the sets in $C$). The maximal principle states that an inductive collection of subsets contains a maximal element.

The maximal principle is quite useful in handling infinite sets in algebra. For example as we will see it can be used to show that every vector space has a basis, which need not be finite in the general case.

The maximal principle is another example of a principle of set theory which, given the other axioms of set theory, is equivalent to the axiom of choice. In chapter 3 a generalization, called Zorn’s lemma, will be given. Many arguments in algebra involving infinite sets can be given using the maximal principle or Zorn’s lemma, but occasionally a more general method, called transfinite induction, is required. A discussion of transfinite induction is given in appendix 1, together with proofs of some theorems from algebra.

The cardinality $|S|$ of an infinite set can be defined. Indeed, the proper handling of this notion was fundamental to Cantor’s development of set theory. The cardinality of a set is something called a “cardinal number” in set theory. Basic algebra can often avoid having to consider these; we may define $|S| \leq |T|$ to hold if there is an injection from $S$ to $T$, and $|S| = |T|$ if there is a bijection. The Bernstein-Cantor-Schröder theorem of set theory states that if $|S| \leq |T|$ and $|T| \leq |S|$ then $|S| = |T|$; a proof is given in appendix 1. If there is a surjection from $S$ to $T$ then $|S| \geq |T|$, because there is an injection from $T$ to $S$. Note also that if $S \subseteq T$ and $|S| < |T|$ (i.e., $\neg(|S| \geq |T|)$), then $S \subset T$.

Occasionally a fact about cardinality is required which requires additional machinery to prove. For one example of such, if $S$ is an infinite set then $|S| = |S \times S|$. Various additional facts can be proved using this. For example, if $S$ is an infinite set and $T$ is the set of finite subsets of $S$ then $|T| = |S|$. Additional discussion can be found in appendix 1.

**Exercises.**

1. Suppose $f : S \rightarrow T$; $S', S_1, S_2 \subseteq S$; and $T', T_1, T_2 \subseteq T$. Show the following.
   a. If $S_1 \subseteq S_2$ then $f[S_1] \subseteq f[S_2]$. $f[\emptyset] = \emptyset$, $f[S_1 \cup S_2] = f[S_1] \cup f[S_2]$, and $f[S_1 \cap S_2] \subseteq f[S_1] \cap f[S_2]$.
   b. If $T_1 \subseteq T_2$ then $f^{-1}[T_1] \subseteq f^{-1}[T_2]$. $f^{-1}[\emptyset] = \emptyset$, $f^{-1}[T] = S$, $f^{-1}[T_1 \cup T_2] = f^{-1}[T_1] \cup f^{-1}[T_2]$, and $f^{-1}[T_1 \cap T_2] = f^{-1}[T_1] \cap f^{-1}[T_2]$.

6
c. \( f[S'] \subseteq T' \) iff \( S' \subseteq f^{-1}[T'] \).
d. \( f[f^{-1}[T']] = f[S] \cap T' \); thus, \( f[f^{-1}[T']] \subseteq T' \), and equality holds if \( f \) is surjective.
e. \( S' \subseteq f^{-1}[f[S']] \), and equality holds if \( f \) is injective.

2. Show that two finite sets have the same cardinality iff there is a bijection between them. You may assume that there is no bijection between \( \{0, \ldots, m-1\} \) and \( \{0, \ldots, n-1\} \) for \( m \neq n \).

3. Show that the divisibility relation on \( \mathbb{Z} \) has the following properties.
   - \( x|y \) iff \( x((-y)) \) iff \( (-x)|y \) iff \( (-x)|(-y) \)
   - \( x|x \)
   - if \( x|y \) and \( y|z \) then \( x|z \)
   - if \( x|y \) and \( y|x \) then \( y = \pm x \)
   - \( x|0 \)
   - \( 1|x \)
   - if \( x|y \) and \( x|z \) then \( x|y+z \)
   - if \( x|y \) and \( w \) is any integer than \( x|wy \)
   - if \( x, y > 0 \) and \( x|y \) then \( x \leq y \)
2. Universal algebra.

A structure consists of a set, with certain members, called constants, singled out; certain functions defined on the set, each having some number of arguments; and certain relations defined on the set, each having some number of arguments. For example in the case of the real numbers the constants are 0 and 1; the functions the two place functions + and ×; and the relations the two place relations = and ≤. The set is called the universe or domain of the structure.

A specification of what are to be the constants, functions, and relations, given say as a list of names together with the numbers of arguments, is called a type. An essentially identical notion is that of a language, which will be defined in chapter 11. Usually in algebra the type is finite, although it need not be. A structure of a given type is a set \( S \), together with elements, functions, and relations as specified by the type. Note that two constant names might refer to the same element, and similarly for functions and relations.

A convenient notational device allows using \( S \) to denote a structure, where \( S \) is the domain. Given a structure of some type, we let \( c \) denote a constant, either the name or the element; the context will make clear which. Similarly \( f \) denotes a function, and \( R \) a relation. In the notation \( f(x_1, \ldots, x_n) \) it is understood that \( f \) is \( n \)-ary. In algebra these conventions usually suffice, although when necessary we may denote a structure as \( \langle S, c_1, \ldots, f_1, \ldots, R_1, \ldots \rangle \) or even \( \langle S, c_1^S, \ldots, f_1^S, \ldots, R_1^S, \ldots \rangle \).

If there are no constants, the question arises of whether the empty set should be considered to be a structure. In universal algebra it is convenient to admit it, whereas in model theory it is omitted. In this chapter we admit it; theorem 1 for example then becomes true without proviso. The question will be mentioned again later in the text.

There are a variety of constructions in algebra, and many of them can be given for structures. In this chapter we will consider substructures, homomorphisms, and products. In later chapters we may appeal to the “universal” definitions, and basic facts which may be proved in the general setting.

If \( T \subseteq S \) and \( f : S^n \rightarrow U \) let us define its restriction to \( T \) to be \( f \cap (T^n \times U) \). Similarly if \( R \subseteq S^n \) we define its restriction to \( T \) to be \( R \cap T^n \). Given structures \( S, T \) of the same type, \( T \) is a substructure of \( S \) if \( T \subseteq S \) (note the use of the above mentioned notational device); and \( T \) contains the constants, and the functions and relations of \( T \) are the restrictions to \( T \) of those of \( S \).

A subset \( T \subseteq S \) is said to be closed under a function \( f \) if \( f(x_1, \ldots, x_n) \in T \) whenever \( x_1, \ldots, x_n \in T \). A substructure must clearly contain the constants and be closed under the functions. On the other hand, given a subset \( T \subseteq S \) which contains the constants and is closed under the functions, there is a unique substructure of \( S \) with domain \( T \), namely that where the constants are as in \( S \), and the functions and relations are the restrictions of those of \( S \). Clearly, if \( S_1 \subseteq S_2 \subseteq S \) where \( S_1, S_2 \) are substructures of \( S \) then \( S_1 \) is a substructure of \( S_2 \).

**Theorem 1.** If \( C \) is a nonempty collection of substructures of \( S \) then \( \bigcap C \) is a substructure of \( S \).

**Proof:** Since the constants are in every member of \( C \) they are in \( \bigcap C \). Suppose \( x_1, \ldots, x_n \in \bigcap C \); then \( x_1, \ldots, x_n \in S \) for every \( S \) in \( C \), so \( f(x_1, \ldots, x_n) \in S \) for every \( S \in C \), so \( f(x_1, \ldots, x_n) \in \bigcap C \).

Let \( T \) be an arbitrary subset of \( S \). Let \( C \) be the collection of substructures of \( S \) which contain \( T \). \( C \) is nonempty since it contains \( S \); \( \bigcap C \) is thus defined. We call it the structure generated by \( T \); it is the smallest substructure containing \( T \), in the sense that any substructure containing \( T \) contains the substructure generated by \( T \). The substructure generated by \( T \) is also called the hull. This construction appears throughout algebra, and in model theory also.

The hull can also be described as those elements which have an expression of the form \( f(\cdots) \) where the arguments are constants, elements of \( T \), or expressions of the same form. This is sometimes described as...
follows. Let $H_0$ be $T$, together with all constants. Given $H_k$, let

$$H_{k+1} = H_k \cup \{ f(x_1, \ldots, x_n) : f \text{ a function, } x_i \in H_k, 1 \leq i \leq n \}.$$ 

Let $H = \bigcup_k H_k$; then $H$ is the hull.

That $H$ is the hull should be clear, but it is worthwhile giving some details of a proof. First note that $H_i \subseteq H_j$ if $i \leq j$. Second, if $x_1, \ldots, x_n \in H$ then $x_i \in H_{k_i}$ for some $k_i$, for each $i$, and so $x_1, \ldots, x_n \in H_k$ where $k$ is the largest of the $k_i$. Thus, if $x_1, \ldots, x_n \in H$ then $x_1, \ldots, x_n \in H_k$ for some $k$, and so $f(x_1, \ldots, x_k) \in H_{k+1} \subseteq H$, proving that $H$ is closed under the functions and so is a substructure containing $T$. To complete the proof it suffices to show that any substructure containing $T$ contains all of the $H_k$. This is clear for $k = 0$, and it follows for $H_{k+1}$, assuming it for $H_k$.

We use the notation $h : S \rightarrow T$ when $h$ is a function from the domain of a structure $S$ to the domain of a structure $T$ of the same type. In this case $h$ is called a homomorphism if

- for each constant $c$, $h(c) = c$;
- for each function $f$, $h(f(x_1, \ldots, x_n)) = f(h(x_1), \ldots, h(x_n))$ for all $x_1, \ldots, x_n \in S$; and
- for each relation $R$, $R(x_1, \ldots, x_n)$ implies $R(h(x_1), \ldots, h(x_n))$, for all $x_1, \ldots, x_n \in S$.

We say $h$ preserves $c$, $f$, or $R$. Note that since no confusion arises, we have omitted the superscript on $c, f, R$.

If in the third requirement “implies” is replaced by “iff” the homomorphism is called strong. Homomorphisms are always strong if there are no relations; this is the most important case, but we will prove various theorems more generally, and these are occasionally useful.

A homomorphism $h$ from $S$ to $T$ is called

- a monomorphism or embedding if $h$ is injective;
- an epimorphism if $h$ is surjective;
- an isomorphism if $h$ is strong and bijective;
- an endomorphism if $T = S$; and
- an automorphism if $T = S$ and $h$ is strong and bijective.

We use the notation $\text{Hom}(S, T)$ for the set of homomorphisms, $\text{End}(S)$ for the set of endomorphisms, and $\text{Aut}(S)$ for the set of automorphisms. In some cases these sets might be understood to have additional structure, as will be seen.

**Theorem 2.** Suppose $h : S \rightarrow T$ is a homomorphism. If $S' \subseteq S$ is a substructure then the restriction of $h$ to $S'$ is a homomorphism, and $h[S']$ is a substructure of $T$. If $T' \subseteq T$ is a substructure then $h^{-1}[T']$ is a substructure of $S$. If $G \subseteq S$ generates $S'$ then $h[G]$ generates $h[S']$.

**Proof:** The restriction of $h$ to $S'$ clearly preserves the constants and functions. Certainly $h[S']$ contains the constants. To see that it is closed under the functions of $T$, if $y_1, \ldots, y_n \in T$ are in $h[S']$ there are $x_1, \ldots, x_n \in S'$ so that $h(x_i) = y_i$, $1 \leq i \leq n$, and so $f(y_1, \ldots, y_n) = f(h(x_1), \ldots, h(x_n)) = h(f(x_1, \ldots, x_n))$, which is in $h[S']$. Since $h(c) = c$ $h^{-1}[T']$ contains the constants. If $x_1, \ldots, x_n \in h^{-1}[T']$ then $h(x_i) \in T'$, $1 \leq i \leq n$, so $f(h(x_1), \ldots, h(x_n)) \in T'$, so $h(f(x_1, \ldots, x_n)) \in T'$, so $f(x_1, \ldots, x_n) \in h^{-1}[T']$. If $t = h(s)$ where $s \in S'$ and $s$ is given by an expression involving some elements $g_i \in G$, then $t$ is given by the same expression, in the elements $h(g_i)$.

**Theorem 3.** If $S, T, U$ are structures of the same type and $h : S \rightarrow T$, $k : T \rightarrow U$ are homomorphisms then $k \circ h$ is a homomorphism. If $k$ and $h$ are strong then so is $k \circ h$.

**Proof:** First, $k \circ h(c) = k(h(c)) = k(c) = c$. Second,

$$k \circ h(f(x_1, \ldots, x_n)) = k(h(f(x_1, \ldots, x_n))) = k(f(h(x_1), \ldots, h(x_n))) = f(k(h(x_1)), \ldots, k(h(x_n))) = f(k \circ h(x_1), \ldots, k \circ h(x_n)).$$
Third,
\[ R(x_1, \ldots, x_n) \implies R(h(x_1), \ldots, h(x_n)), \quad \text{implies} \]
\[ R(k(h(x_1)), \ldots, k(h(x_n))) \iff R(k \circ h(x_1), \ldots, k \circ h(x_n)); \]
for strong homomorphisms, “implies” may be replaced by “iff”.

**Theorem 4.** If \( h \) is an isomorphism from \( S \) to \( T \), then \( h^{-1} \) is a strong homomorphism, and so an isomorphism from \( T \) to \( S \).

**Proof:** If \( y_i = h(x_i), 1 \leq i \leq n \), then \( f(y_1, \ldots, y_n) = h(f(x_1, \ldots, x_n)) \) and so \( h^{-1}(f(y_1, \ldots, y_n)) = f(x_1, \ldots, x_n) \), that is \( h^{-1}(f(y_1, \ldots, y_n)) = f(h^{-1}(y_1), \ldots, h^{-1}(y_n)) \). The defining property for predicates is similarly verified.

Note also that a homomorphism \( h \) is an isomorphism if it is bijective and \( h^{-1} \) is a homomorphism. Two structures are said to be isomorphic if there is an isomorphism between them. From one point of view isomorphic structures are essentially the same. To say that there is a unique such and such usually means that all such and such’es are isomorphic. On the other hand in some contexts one might be concerned with a particular isomorphic copy, especially of substructures of a structure.

A congruence relation in a structure \( S \) is defined to be an equivalence relation where
- for each function \( f \) if \( x_1 \equiv y_1, \ldots, x_n \equiv y_n \) then \( f(x_1, \ldots, x_n) \equiv f(y_1, \ldots, y_n) \).
If also
- for each predicate \( R \) if \( x_1 \equiv y_1, \ldots, x_n \equiv y_n \) then \( R(x_1, \ldots, x_n) \) holds if and only if \( R(y_1, \ldots, y_n) \) does, the congruence relation is said to be strong.

Given a congruence relation \( \equiv \) on \( S \), a structure may be defined on the equivalence classes. This structure, called the quotient of \( S \) by \( \equiv \), and denoted \( S/\equiv \), is defined as follows. The domain of \( S/\equiv \) is \( \{ [x] : x \in S \} \), the set of equivalence classes. To define \( f \) in \( S/\equiv \), simply let \( f([x_1], \ldots, [x_n]) = [f(x_1, \ldots, x_n)] \); i.e., given argument classes simply pick an element (called a representative) of each, apply the function of \( S \), and consider the class of the result. By the defining properties of a congruence relation the class of the result will not depend on the representatives chosen from the argument classes. A constant is interpreted as the class containing it. \( R([x_1], \ldots, [x_n]) \) for a relation \( R \) is defined to hold if \( R(x_1, \ldots, x_n) \) does for some representatives \( x_1, \ldots, x_n \); if the congruence relation is strong then this will be so iff it holds for every choice of representatives.

If \( h : S \to T \) is a function between sets the relation \( \equiv \) defined on \( S \) by \( x \equiv y \) if and only if \( h(x) = h(y) \) is readily verified to be an equivalence relation. We call this the kernel of \( h \). Note that \( h \) is injective iff the kernel is the equality relation on \( S \).

**Theorem 5.** If \( \equiv \) is a congruence relation on \( S \) then the map \( e : S \to S/\equiv \) defined by \( e(x) = [x] \) is an epimorphism whose kernel is \( \equiv \). If \( \equiv \) is strong then so is \( e \).

**Proof:** Clearly \( e \) is surjective. For a constant \( c \) \( e(c) = [c] \) and \([c] \) is by definition the constant in the quotient. For a function \( f \) \( f([x_1], \ldots, [x_n]) = [f(x_1, \ldots, x_n)] \) by definition, and this is just the required property for functions. For a relation \( R \), if \( R(x_1, \ldots, x_n) \) then \( R([x_1], \ldots, [x_n]) \), and this is the required property for relations; if \( \equiv \) is strong then the converse implication also holds. Finally \( e(x) = e(y) \) iff \( [x] = [y] \) iff \( x \equiv y \).

The map \( e \) is called the canonical epimorphism. Starting with a congruence relation, we define the quotient structure and the canonical epimorphism from the structure onto the quotient. If we start with a homomorphism \( h : S \to T \), if \( h \) is injective then \( h[S] \) is isomorphic to \( S \) only if \( h \) is strong. More generally we have the following.
Theorem 6. If \( h : S \rightarrow T \) is a homomorphism the kernel \( \equiv \) is a congruence relation in \( S \). The map \( i : S/\equiv \rightarrow T \) defined by \( i([x]) = h(x) \) is a monomorphism. If \( h \) is strong so are \( \equiv \) and \( i \).

Proof: If \( x_1 \equiv y_1, \ldots, x_n \equiv y_n \), that is, \( h(x_1) = h(y_1), \ldots, h(x_n) = h(y_n) \), then
\[
  h(f(x_1, \ldots, x_n)) = f(h(x_1), \ldots, h(x_n)) = f(h(y_1), \ldots, h(y_n)) = h(f(y_1, \ldots, y_n));
\]
this shows that \( f(x_1, \ldots, x_n) \equiv f(y_1, \ldots, y_n) \), so \( \equiv \) is a congruence relation. If \( h \) is strong a similar computation shows that \( \equiv \) is. Next we verify that \( i \) is well defined, that is, that \( i([x]) \) does not depend on the representative \( x \). But this is clear; if \( [x] = [y] \) then \( h(x) = h(y) \) because \( \equiv \) is the kernel of \( h \). Next we verify that \( i \) is a homomorphism. For a constant \( c \), \( i([c]) = h(c) \), and \( [c] \) and \( h(c) \) are the constant in \( S/\equiv \) and \( T \) respectively. For a function \( f \),
\[
i(f([x_1], \ldots, [x_n])) = i(f(x_1, \ldots, x_n)) = h(f(x_1, \ldots, x_n)) = f(h(x_1), \ldots, h(x_n)) = f(i([x_1]), \ldots, i([x_n])).
\]
For a relation \( R \),
\[
R([x_1], \ldots, [x_n]) \text{ implies } R(x_1, \ldots, x_n) \text{ implies } R(h(x_1), \ldots, h(x_n)) \text{ implies } R(i([x_1]), \ldots, i([x_n])),
\]
provided the representatives are chosen correctly; if \( h \) is strong “implies” may be replaced by “iff”, for any choice of representatives. Finally, \( i([x]) = i([y]) \) iff \( h(x) = h(y) \) iff \( [x] = [y] \), so the kernel of \( i \) is the equality relation and it is injective.

For \( h \) strong the map \( i \) yields an isomorphism from \( S/\equiv \) to \( h[T] \), called the canonical isomorphism. Starting with a strong homomorphism, there is a canonical isomorphism from the quotient by its kernel to its image. This fact may also be viewed as follows; the epimorphism \( e \) and the monomorphism \( i \) make the “diagram” of Figure 1 “commutative”. By this we mean that if functions are composed along any two paths from one object to a second, the resulting function will be the same, regardless of the path. Note that the functions in the composition are written from right to left in order from the start node to the finish node.

![Figure 1](image.png)

There is an additional fact about this diagram. Replace the arrow from \( S/\equiv \) to \( T \) with a dashed arrow. We have shown that there is a map \( i \) which makes the diagram commutative. In algebra one often shows that there is a function making a diagram commutative. Often this function will be unique. That is the case here; to make the diagram commutative, \( i([x]) \) must equal \( h(x) \).

The restriction of a congruence relation \( \equiv \) to a substructure \( S' \subseteq S \) is \( \equiv \cap (S' \times S') \); this is readily seen to be a congruence relation. In fact, if \( \equiv \) is the kernel of a homomorphism \( h \) then the restriction of \( \equiv \) is the kernel of the restriction of \( h \).

General definitions involving structures are very useful in algebra. However usually some subcollection of all the structures is under consideration. We call these the distinguished structures. (In universal algebra the distinguished structures are often called the algebras. The term “algebra” is also used for a particular family of distinguished structures, and we will reserve it for this use.) Invariably, a homomorphism between distinguished structures is any homomorphism between the structures.
The distinguished structures are invariably specified as those satisfying a set of axioms. In elementary logic a formal definition can be given of what an axiom is, and what it means for it to be true in a structure. The axioms are strings of symbols in a formally defined language, which includes “logical symbols” (propositional connectives, universal and existential quantifier, equals sign), in addition to symbols for the constants, functions, and relations, these being called nonlogical symbols. The meaning of any legal string can be formally defined. We will give these definitions in chapter 11; for now we rely on the reader’s intuition.

A substructure of a distinguished structure need not be distinguished in general. Some simplifications result if this is the case. For example, the hull of $T$ is then the least distinguished structure containing $T$. Often the type can be enlarged to ensure that this is the case. However it is sometimes useful to be aware that a smaller type suffices to specify the structures under consideration.

A classic example of this situation occurs in the family of structures known as groups. A group may be given by a single binary function $\cdot$; the operator symbol is often omitted, so that we may write $xy$ for $x \cdot y$.

One set of axioms is as follows.
- For all $x, y, z$, $(xy)z = x(yz)$.
- There is an element $e$ such that for all $x$, $xe = ex = x$.
- For all $x$ there is an element $x^{-1}$ such that $xx^{-1} = x^{-1}x = e$.

With these axioms a substructure need not be a group, since the second and third axioms, which state the existence of certain elements, need not hold.

However one can show from the axioms that $e$ and $x^{-1}$ are unique. If the type is enlarged to include a constant $e$ and a unary function $x^{-1}$ (superscript notation is common for unary functions), the second and third axioms may be replaced by,
- for all $x$ $xe = ex = x$; and
- for all $x$ $xx^{-1} = x^{-1}x = e$.

The new axioms define exactly the same class of structures.

In the larger language, a substructure automatically satisfies the axioms. This is an instance of a general phenomenon. The axioms are all universally quantified; they state that for all values of the variables a certain relationship holds. In general one may allow propositional connectives, such as in the definition of transitivity,
- if $xRy$ and $yRz$ then $xRz$.

No existential quantifiers are allowed. A standard convention allows omitting all quantifiers, so that the formula is implicitly universally quantified on all its variables.

It is clear that any universally quantified statement holds in a substructure if it holds in the structure; we will prove this in chapter 11. Many basic classes of structures can be defined with universally quantified axioms; indeed, they can be defined with equations, as is the case for groups. Such classes of structures are studied in general, although for our purposes the important point is that they include many common classes. To repeat, when the distinguished structures are those satisfying a set of universally quantified axioms a substructure of a distinguished structure is distinguished. Note that, by theorem 2, an image or inverse image of a substructure under a homomorphism (between distinguished structures) is distinguished. For example, the image of a group under a group homomorphism is a subgroup of the codomain group.

Universally quantified axioms are clearly convenient. On the other hand, the multiplication operation of a group suffices to determine the identity element and the inverse function, so there is no loss in considering the structure as only specifying multiplication. A subgroup is a substructure which satisfies the axioms, so it must contain the identity element and be closed under the inverse function. One can choose which alternative one prefers at any point.

Given a structure $S_i$ for each $i$ in some index set $I$, all of of some given type, their product will be defined as a structure of the given type, whose domain is $\times_i S_i$, the Cartesian product of the domains of the
structure. Here, there are several domains of existence. Similar remarks apply to various other algebraic constructions.

Given a concrete category such as the category of structures one can, and indeed usually does to show in either case (indeed in any category). In category theory, one cannot single out "the" product; however in the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.

The definition of a product in terms of the diagram property is said to be arrow theoretic or category theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.

The definition of a product in terms of the diagram property is said to be arrow theoretic or category theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.

The definition of a product in terms of the diagram property is said to be arrow theoretic or category theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.

The definition of a product in terms of the diagram property is said to be arrow theoretic or category theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.

The definition of a product in terms of the diagram property is said to be arrow theoretic or category theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the language of category theory, theorem 7 shows that a product exists in the category of structures of a given certain maps between them, satisfying certain axioms; we know nothing further about the objects. In the theoretic. In category theory, which we will discuss in chapter 13, all that is given is certain objects, and further, the evaluation map at the theorem may be generalized to infinite sums by giving appropriate definitions. Similarly if one with disjoint domains. However given any multisorted structure there is an isomorphic one with disjoint domains.
“multiplication” operation · defined on it then \( \prod_{1 \leq i \leq n} a_i \) equals \( a_1 \cdot \ldots \cdot a_n \). The notation is also generalized to other operators (using the same symbol, as in \( \oplus_i x_i \)).

**Exercises.**

1. If \( T \subseteq S \), the inclusion map \( \subseteq: T \mapsto S \) (the “overloading” of the symbol \( \subseteq \) causes no confusion) maps \( x \in T \) to itself. Show that \( T \) is a substructure iff the inclusion map is a homomorphism, in which case it is strong.

2. Complete the proof of theorem 7.
3. Orders.

A binary relation $R$ is called
- a preorder if it is reflexive and transitive;
- antisymmetric if $x = y$ whenever $xRy$ and $yRx$;
- a partial order if $R$ is a preorder and antisymmetric;
- irreflexive if $xRx$ is false for all $x$.

A partially ordered set (or poset) is a structure $S$ with a single binary relation $\leq$, a partial order on $S$.

**Theorem 1.** If $\leq$ is a partial order, and $x < y$ is defined to hold if $x \leq y$ but $x \neq y$ then $<$ is irreflexive and transitive. If $<$ is irreflexive and transitive, and $x \leq y$ is defined to hold if $x < y$ or $x = y$, then $\leq$ is a partial order.

**Proof:** We leave the proof as an exercise, and also the proofs of the basic facts stated in the next five paragraphs.

The relation $<$ is called the strict or proper part of $\leq$; a partial order may be equally well given by giving its strict part. If $\leq$ is a preorder on a set $S$, two elements $x, y \in S$ are called comparable if $x \leq y$ or $y \leq x$; otherwise they are incomparable. The strict part of a total order may be defined equivalently as an irreflexive transitive binary relation satisfying the law of trichotomy, that $x < y$ or $x = y$ or $y < x$. The transpose of a preorder $\leq$ is denoted $\geq$. The transpose of a preorder is a preorder; that of a partial order a partial order, and its strict part is denoted $>$; and that of a total order a total order. The transpose order is also called the opposite order.

Suppose $\leq$ is a preorder on $S$ and $S' \subseteq S$. An element $x \in S$ is called
- a lower bound of $S'$ if $x \leq y$ for all $y \in S'$;
- an upper bound of $S'$ if $x \geq y$ for all $y \in S'$;
- an infimum of $S'$ if it is a lower bound, and for any other lower bound $w$, $x \leq w$;
- a supremum of $S'$ if it is an upper bound, and for any other upper bound $w$, $x \geq w$.

Note that an upper bound is the same as a lower bound in the opposite order; this duality holds for many definitions and theorems concerning preorders. An infimum is also called a greatest lower bound and a supremum a least upper bound.

If $\leq$ is a preorder on $S$ a subset $S' \subseteq S$ is called
- directed if every pair of its elements has an upper bound in $S'$;
- filtered if every pair of its elements has a lower bound in $S'$;
- a chain if every pair of its elements is comparable.

Clearly, $S'$ is directed (filtered) iff any nonempty finite subset has an upper (lower) bound in $S'$. Also, a chain is both directed and filtered.

If $\leq$ is a preorder on $S$ and $S' \subseteq S$, an element $x \in S'$ is called
- minimal if whenever $y \in S'$ and $y \leq x$ then $y = x$;
- maximal if whenever $y \in S'$ and $y \geq x$ then $y = x$.

If $S$ is a partial order $x$ is called
- least if $x \leq y$ for all $y \in S'$;
- greatest if $x \geq y$ for all $y \in S'$.

In a partially ordered set we have
- a least upper (greatest lower) bound for $S'$, or a least (greatest) element of $S'$, is unique if it exists;
- if an upper (lower) bound is a member of $S'$, it is a greatest (least) element, and also a least upper (greatest lower) bound;
a maximal (minimal) element of a directed (filtered) set is a greatest (least) element.

If $S$ is a preorder, the relation $x \equiv y$, defined to hold if $x \leq y$ and $y \leq x$, is a strong congruence relation. The preorder $S/\equiv$ is in fact a partial order, called the induced partial order.

From universal algebra a homomorphism between preorders is a function $f$ such that if $x \leq y$ then $f(x) \leq f(y)$. Such a map is also called order preserving, or monotone.

A partially ordered set $S$ is called inductive if every chain $C \subseteq S$ has an upper bound. Zorn’s lemma states that a nonempty inductive partial order contains maximal elements. As mentioned in chapter 1, this is equivalent to the axiom of choice, and the maximal principle is a special case. Transfinite induction may be used to prove Zorn’s lemma, although proofs have been devised which do not require it.

A well-order on a set $S$ is a total order such that, if $S' \subseteq S$ then $S'$ contains a least element. Another principle equivalent to the axiom of choice is the well-ordering principle. This states that given any set $S$, there is a well-order on $S$.

In a partially ordered set $S$, an ascending chain is a finite or countably infinite sequence $x_0 < x_1 < \cdots$. $S$ is said to satisfy the ascending chain condition if there is no infinite ascending chain. Equivalently, any nondecreasing infinite sequence $x_i$, i.e., sequence where $x_{i+1} \geq x_i$, must eventually become constant, i.e., for some $n$ and $x$, $x_i = x$ for $i \geq n$.

If there is an infinite ascending chain $C$ in a partially ordered set $S$ then $C$ has no maximal element. Conversely if a subset $S'$ of $S$ has no maximal elements we may successively choose (using the axiom of choice) elements $x_0 < x_1 \cdots$ to form an infinite ascending chain. Thus, $S$ satisfies the ascending chain condition iff every subset contains a maximal element.

The notion of a descending chain is dual to that of an ascending chain. $S$ satisfies the descending chain condition iff every subset contains a minimal element. We also say that $\leq$ is well founded, or that $S$ is partially well-ordered.

A finite chain is said to have length $n$ if there are $n+1$ elements in the chain. Thus, a chain $x_0 \leq \cdots \leq x_n$ has length $n$. A chain which is not finite is said to have infinite length.

In a partially ordered set, the least upper bound of a subset $S' \subseteq S$ is also called the join of $S'$. If this exists for every pair of elements, the partially ordered set is called a $\sqcup$-semilattice, and $x \sqcup y$ denotes the join. The greatest lower bound of a subset is also called the meet. If this exists for every pair of elements, the partially ordered set is called a $\sqcap$-semilattice, and $\sqcap$ denotes the meet. If both $x \sqcup y$ and $x \sqcap y$ always exist, the structure is called a lattice.

Define the $\sqcup$-semilattice axioms to be the axioms

\[
\begin{align*}
x \sqcup x &= x \\
x \sqcup y &= y \sqcup x \\
x \sqcup (y \sqcup z) &= (x \sqcup y) \sqcup z;
\end{align*}
\]

the $\sqcap$-semilattice axioms are defined dually, i.e., by replacing $\sqcup$ by $\sqcap$ (in fact these are the same axioms). The lattice axioms consist of both sets of semilattice axioms, and

\[
x \sqcup (x \sqcap y) = x, \quad x \sqcap (x \sqcup y) = x.
\]

Duality is especially useful in lattices. Informally, a term is an expression involving functions and variables; a formal definition is given in chapter 11. In lattices, the functions are $\sqcup$ and $\sqcap$; the dual $\hat{t}$ of a term $t$ is obtained by exchanging these. The dual of an equation $t = u$ is $\hat{t} = \hat{u}$; and of an inequality $t \leq u$ is $\hat{t} \geq \hat{u}$. The lattice axioms are closed under duality, so if an equation or inequality holds in a lattice so does its dual; this is intuitively clear and can be proved formally using the machinery presented in chapter 11.
LEMMA 2.

a. A $\sqcup$-semilattice satisfies the $\sqcup$-semilattice axioms, and $x \leq y$ iff $x \sqcup y = y$.

b. Given a structure with a binary operation $\sqcup$ satisfying the $\sqcup$-semilattice axioms, define $x \leq y$ if $x \sqcup y = y$; the structure is a $\sqcup$-semilattice, with $x \sqcup y$ the join.

c. In a $\sqcap$-semilattice, $\sqcap$ is monotone in each argument.

d. A map $f$ between $\sqcup$-semilattices is monotone iff $f(x \sqcup y) \leq f(x) \sqcup f(y)$.

PROOF: For part a, certainly $x \geq x$ and if $u \geq x$ then $u \geq x$, so $x = x \sqcup x$. The conditions $w \geq x, y$ and whenever $u \geq x, y u \geq w$ equally well specify $x \sqcup y$ and $y \sqcup x$. Now, the following are equivalent:

- $x \sqcup (y \sqcup z) = w$;
- $w \geq x, y \sqcup z$ and whenever $u \geq x, y \sqcup z$ then $u \geq w$;
- $w \geq x, y, z$ and whenever $u \geq x, y, z$ then $u \geq w$;
- $w \geq x \sqcup y, z$ and whenever $u \geq x \sqcup y, z$ then $u \geq w$;
- $(x \sqcup y) \sqcup z = w$.

To see that the second statement implies the third, since $w \geq x, y \sqcup z$ we have $w \geq x, y, z$; and if $u \geq x, y, z$ then $u \geq w$. To see that the third statement implies the second, since $w \geq x, y, z$ we have $w \geq x, y \sqcup z$; and if $u \geq x, y \sqcup z$ then $u \geq w$. The equivalence of the third and fourth statements follows similarly. If $x \leq y$ then $y \geq x$ and if $u \geq x$ then $u \geq y$; thus $x \sqcup y = y$. If $x \sqcup y = y$ then certainly $y \geq x$. For part b, suppose $\sqcap$ is given. We have $x \leq x$ since $x \sqcup x = x$. If $x \leq y, y \leq z$ and whenever $x \sqcup y \geq w$ then $x \sqcup y \geq x \sqcup z$ and whenever $x \sqcup y \geq d$ then $x \sqcup y \geq z$. The dual identity is readily verified to preserve any meets or joins which exist, and hence a partial order isomorphism between $\sqcup$-semilattices is a homomorphism iff $f(x \sqcup y) = f(x) \sqcup f(y)$. By part d such a map is monotone; but it is easy to see that the converse is not true. The dual statements to the lemma hold for a $\sqcap$-semilattice; in this case $x \geq y$ iff $x \sqcap y = y$, and $f$ is monotone iff $f(x \sqcap y) \geq f(x) \sqcap f(y)$. A partial order isomorphism is readily verified to preserve any meets or joins which exist, and hence a partial order isomorphism between $\sqcup$-semilattices is a $\sqcap$-semilattice isomorphism, etc.

THEOREM 3. A lattice satisfies the lattice axioms, and $x \leq y$ iff $x \sqcup y = y$ iff $x \sqcap y = x$. Conversely given a structure satisfying the lattice axioms, $x \sqcup y = y$ iff $x \sqcap y = x$; defining $x \leq y$ if these hold, the structure is a lattice.

PROOF: That $x \sqcup (x \sqcap y) = x$ follows because $x \geq x, x \sqcap y$ and if $u \geq x, x \sqcap y$ then $u \geq x$. The dual identity has a dual proof. The rest of the first claim follows by theorem 2. Given a structure satisfying the lattice axioms, if $x \sqcup y = y$ then $x \sqcap y = x \sqcup (x \sqcup y) = x$; and similarly $x \sqcap y = x$ implies $x \sqcup y = y$. It follows by theorem 2 that the structure is a lattice.

In a $\sqcup$-semilattice, there is a greatest element iff the join of the entire set exists; it will be denoted 1. There is a least element iff the join of the empty set exists; it will be denoted 0. If 1 exists all finite sets have joins; some authors require this in the definition. Dually in a $\sqcap$-semilattice, 1 is the meet of the empty set and 0 the meet of the entire set. A lattice is called complete if the meet and join exist for every subset $T$ of the domain $S$.

THEOREM 4. If $\leq$ is a partial order on $S$, and if the meet of every subset $T \subseteq S$ exists, the partially ordered set is a complete lattice.
Proof: Let $T \subseteq S$, and let $U$ be the set of upper bounds of $T$. Let $w$ be the meet of $U$. If $x \in T$ then $x \leq u$ for every $u \in U$, so $x \leq w$; hence $w$ is an upper bound of $T$, that is $w \in U$, and being a lower bound of $U$ it is the least element of $U$, that is, the least upper bound of $T$.

A collection $C$ of subsets of a set $S$ is called a closure system if it is closed under arbitrary intersection. In a closure system we use $\cap$ to denote the meet, and $\cup$ to denote the join. In a closure system, given any subset $T \subseteq S$ there is a least set $T^{cl} \in C$ containing $T$; this is called the closure of $T$. The join of a subcollection $C' \subseteq C$ clearly equals $(\bigcup C')^{cl}$. A closure system can be given by giving the map $T \mapsto T^{cl}$; see the exercises.

A closure system is called algebraic if the join of a directed subcollection is its union. Thus, an algebraic closure system is a family of subsets of $X$ which is closed under arbitrary intersection and unions of directed subcollections. Observe that an algebraic closure system is an inductive partial order. It is also true that it suffices to require that the join of a chain be its union; the proof requires transfinite induction and is given in appendix 1. The family of lattices which are that of an algebraic closure system can be characterized abstractly, and are called the algebraic lattices; see chapter 21.

The substructures of a structure $S$ yield a family $C$ of subsets of $S$. We have already seen that $C$ is closed under arbitrary intersections; that is, $C$ is a closure system. The closure of a subset $T \subseteq S$ is the hull of $S$, and the join of a subfamily $C'$ is the hull of $\bigcup C'$. $C$ is in fact an algebraic closure system. To see this, suppose $C'$ is directed and define a structure on $\bigcup C'$ as follows. If $x_1, \ldots, x_n \in \bigcup C'$ then $x_1, \ldots, x_n \in S$ for some $S \in C'$ (since $C'$ is directed). We may thus define $R(x_1, \ldots, x_n)$ or $f(x_1, \ldots, x_n)$ by considering the value in $S$; which $S$ is chosen is irrelevant, as long as it contains $x_1, \ldots, x_n$. Clearly this structure is the join of $C'$.

The above remarks hold if we consider only distinguished structures, provided a substructure of a distinguished structure is distinguished. Thus, for example, given any subset $S$ of a group $G$, there is a smallest subgroup containing $S$, namely the hull of $S$ when the language is taken as $1, \cdot, \cdot^{-1}$. This is also called the subgroup generated by $S$.

It is readily seen that in a lattice,

$$x \cap (y \cup z) \geq (x \cap y) \cup (x \cap z), \quad x \cup (y \cap z) \leq (x \cup y) \cap (x \cup z).$$

Indeed $x \cup (y \cap z) \leq x \cup y, x \cup z$, etc.

Lemma 5. In a lattice the following are equivalent.

1. $x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$,
2. $x \cup (y \cap z) = (x \cup y) \cap (x \cup z)$,
3. $x \cap (y \cup z) \leq x \cup (y \cap z)$.

Proof: Statement 3 is self-dual, so it suffices to show (1) iff (3). If (1) then $x \cap (y \cup z) = (x \cap y) \cup (x \cap z) \leq (x \cap y) \cup z$. If (3) then $x \cap (y \cup z) \leq y \cup (x \cap z)$, whence $x \cap (y \cup z) \leq x \cap (y \cup (x \cap z))$, whence $x \cap (y \cup z) \leq (x \cap y) \cup (x \cap z), and (1)$ follows.

A lattice satisfying the statements of the lemma is called distributive. A lattice is called modular if $x \cap (y \cup z) = (x \cap y) \cup z$ whenever $z \leq x$. Clearly a distributive lattice is modular; the converse is not true (see section 21.1).

If $z \leq x$ then $(x \cap y) \cup z \leq x, y \cup z$, so the modularity condition follows if $x \cap (y \cup z) \geq (x \cap y) \cup z$ whenever $z \leq x$. It is easy to see that a lattice is modular iff $x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$ whenever $y \leq x$ or $z \leq x$. The modular law is self-dual, so a lattice is modular iff $x \cup (y \cap z) = (x \cup y) \cap (x \cap z)$ whenever $y \geq x$ or $z \geq x$. Various other equivalent conditions can be given.
If a lattice $L$ contains a least (resp. greatest) element we use the symbol 0 (resp. 1) to denote it, and call $L$ a 0-lattice (resp. 1-lattice). A $\{0, 1\}$-lattice is a lattice with both 0 and 1. In a $\{0, 1\}$-lattice, an element $b$ is called a complement of $a$ if $b \cap a = 0$, $b \cup a = 1$. A distributive $\{0, 1\}$-lattice in which every element has a complement is called a Boolean algebra.

**Lemma 6.** In a distributive $\{0, 1\}$-lattice a complement is unique if it exists.

**Proof:** Suppose $b_1, b_2$ are both complements of $a$. Then $b_1 = b_1 \cap 1 = b_1 \cap (a \cup b_2) = (b_1 \cap a) \cup (b_1 \cap b_2) = 0 \cup (b_1 \cap b_2) = b_1 \cap b_2$. Similarly $b_2 = b_1 \cap b_2$.

The Boolean algebra axioms are the following statements, for a structure with binary operations $\cap$ and $\cup$, a unary operation $^c$, and constants 0 and 1.

- $x \cup y = y \cup x$, $x \cap y = y \cap x$
- $x \cup (y \cup z) = (x \cup y) \cup z$, $x \cap (y \cap z) = (x \cap y) \cap z$
- $x \cup (y \cap z) = (x \cup y) \cap (x \cup z)$, $x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$
- $x \cap 0 = x$, $x \cap 1 = 1$,
- $x \cap x^c = 0$

**Theorem 7.** A Boolean algebra satisfies the Boolean algebra axioms, where 0 (1) is the least (greatest) element, and $x^c$ is the complement. Conversely a structure satisfying the Boolean algebra axioms is a Boolean algebra, with least (greatest) element 0 (1), and $x^c$ the complement of $x$.

**Proof:** For the first claim, $x \cup 0 = x$ is just the claim that 0 is the least element; and similarly for $x \cap 1 = x$. The other axioms are all straightforward. For the second claim, the idempotent and absorption laws must be shown; this is left as an exercise.

In a Boolean algebra, $x - y$ is defined to be $x \cap y^c$, and $x \oplus y$ to be $(x - y) \cup (y - x)$. We leave it as an exercise to verify the following.

- $x \cap 1 = 1$ and $x \cap 0 = 0$
- $0^c = 1$ and $1^c = 0$
- $y = x^c$ if and only if $x \cup y = 1$ and $x \cap y = 0$
- $(x^c)^c = x$
- $(x \cup y)^c = x^c \cap y^c$ and $(x \cap y)^c = x^c \cup y^c$
- $x \leq y$ if and only if $y^c \leq x^c$
- $x \leq y$ if and only if $x \cap y^c = 0$
- $x \cap y = 0$ if and only if $x \leq y^c$
- $x \oplus y = y \oplus x$
- $(x \oplus y) \oplus z = x \oplus (y \oplus z)$
- $x \oplus 0 = x$
- $x \oplus 1 = x^c$
- $x \oplus x = 0$
- $x \oplus x^c = 1$
- $x \oplus y = (x \cup y) \cap (x \cap y)^c = (x \cup y) - (x \cap y)$

**Exercises.**

1. Prove theorem 1, and the facts stated in the following two paragraphs.
2. Complete the proof of theorem 2.
3. Show that the equivalence relations on a set $S$ form a closure system, as subsets of $S \times S$. Give an explicit characterization of the join. Show that they form an algebraic closure system. If $\equiv_1 \subseteq \equiv_2$, i.e., if
\[ x \equiv_2 y \text{ whenever } x \equiv_1 y, \equiv_2 \text{ is said to refine } \equiv_1. \] Show that this is so iff each equivalence class of \( \equiv_1 \) is a union of equivalence classes of \( \equiv_2 \).

4. Show that the congruence relations on a structure \( S \) form an algebraic closure system, when ordered by inclusion.

5. Let \( L \) be a complete lattice and \( K \subseteq L \). Define \( K \) to be a \( \sqcap^* \)-sublattice if it is closed under arbitrary meets. For example, if \( L = \text{Pow}(S) \) a \( \sqcap^* \)-sublattice is a closure system. Define a closure operator on \( L \) to be a map \( x \mapsto x^- \) such that \( x \leq y \Rightarrow x^- \leq y^- \), \( x^- = x^- \), and \( x \leq x^- \). Show the following.
   a. If \( K \) is a \( \sqcap^* \)-sublattice the map \( x \mapsto \sqcap\{y \in K : x \leq y\} \) is a closure operator.
   b. If \( x \mapsto x^- \) is a closure operator the set \( L^- \) (i.e., the image) is a \( \sqcap^* \)-sublattice. Hint: \( L^- = \{x \in L : x^- \leq x\} \), and if \( S \subseteq L^- \) and \( s = \sqcap S \) then \( s^- \leq \sqcap S \).
   c. The maps of parts a,b between \( \sqcap^* \)-sublattices and closure operators are inverse to each other.
   d. Write \( \sqcup^- \) for the join in \( L^- \); then for \( S \subseteq L \), \( \sqcup^- S^- = (\sqcup S)^- \). Hint: If \( s = \sqcup S \) then \( s^- \) is an upper bound for \( S^- \), and if \( t \in L^- \) is an upper bound then \( t \geq s^- \). In chapter 13 we will see that this follows because the corestriction of \( x \mapsto x^- \) to \( L^- \) is the left adjoint to the inclusion of \( L^- \) in \( L \), i.e., for \( y \in L^- \), \( x^- \leq y \) iff \( x \leq y \).
   e. \( L^- \) is closed under directed joins iff \( x \mapsto x^- \) preserves directed joins. Hint: If \( L^- \) is closed under directed joins and \( S \) is directed then
   \[
   (\sqcup S)^- = \sqcup S^- = \sqcup^- (\sqcup S)^- = (\sqcup S^-)^- = \sqcup S^-.
   \]
   If \( x \mapsto x^- \) preserves directed joins and \( S \subseteq L^- \) is directed then \( \sqcup S = \sqcup S^- = (\sqcup S)^- \).

6. Complete the proof of theorem 7, and prove the facts stated in the paragraph following it.

1. Basic definitions. A binary function \( \cdot \) on a set \( S \) is said to be associative if \( x \cdot (y \cdot z) = (x \cdot y) \cdot z \) for all \( x, y, z \in S \). A structure with a single associative binary function is called a semigroup. The function is called the multiplication of the semigroup. The \( \cdot \) sign may be omitted, so we may write \( xy \) rather than \( x \cdot y \). Any product of any number of elements may be simply written as \( x_1 \cdots x_n \), by the associative law and induction (the order matters, however). An element in a semigroup is called an identity element if \( xe = ex = x \) for all \( x \in S \). A semigroup with an identity element is called a monoid.

**Theorem 1.** In a semigroup, if there is an identity element it is unique.

**Proof:** If \( e, f \) are both identity elements then \( e = ef = f \).

There is a unique monoid consisting of the identity element alone; this is called the trivial monoid, and any other monoid is called nontrivial. Given a monoid \( S \), and an element \( x \in S \), the element \( x' \in S \) is called an inverse of \( x \) if \( x'x = xx' = e \), where \( e \) is the identity element.

**Theorem 2.** In a monoid, if an element has an inverse the inverse is unique.

**Proof:** If \( x', x^* \) are both inverses of \( x \) then \( x' = x' e = x' (x x^*) = (x' x) x^* = e x^* = x^* \).

The inverse of \( x \) is denoted \( x^{-1} \). A group is a monoid in which every element has an inverse. Note that the trivial monoid is a group, called the trivial group. Since the axioms are universally quantified a substructure of a semigroup is again a semigroup, called a subsemigroup. Universally quantified axioms may be given for monoids (add \( xe = ex = x \)) or groups (add also \( xx^{-1} = x^{-1} x = e \)); a substructure is then a monoid or group respectively, and called a submonoid or subgroup.

**Theorem 3.** The elements of a monoid which possess an inverse form a submonoid which is a group. Further, \( (xy)^{-1} = y^{-1} x^{-1} \), \( (x^{-1})^{-1} = x \), and \( e^{-1} = e \).

**Proof:** It suffices to observe that \( xyy^{-1} x^{-1} = e \), \( x^{-1} x = e \), and \( ee = e \).

The cancellation law states that if \( ax = ay \) or \( xa = ya \) then \( x = y \). Note that if \( a \) has an inverse then \( x = y \) follows; in particular, the cancellation law holds in a group.

By universal algebra, given any set \( S \subseteq G \) of elements of a semigroup there is a least subsemigroup containing \( S \), namely the least such substructure; this is called the subsemigroup generated by \( S \). The same is true of monoids or groups. A homomorphism between semigroups preserves multiplication; one between monoids preserves in addition the identity; and one between groups preserves in addition the inverse. The image or inverse image of a subsemigroup under a homomorphism between semigroups is again a subsemigroup; and similarly for submonoids or subgroups. Although it does not follow by universal algebra, it is easy to verify that a map between groups preserving multiplication is a homomorphism. For monoids, this is true if the codomain monoid satisfies the cancellation law or if the homomorphism is surjective.

An example of a monoid is provided by the functions from a set \( S \) to \( S \), equipped with composition as the multiplication operation; the identity function \( \iota_S \) is the identity element. The elements which possess an inverse are exactly the bijections; the bijections of a set thus form a group with composition as the multiplication operation. More generally given a structure \( S \), \( \text{End}(S) \) becomes a monoid when equipped with composition and the identity map. The elements possessing an inverse are the automorphisms, which form a group \( \text{Aut}(S) \), called the automorphism group of the structure.

A semigroup is said to be commutative if it satisfies the axiom \( xy = yx \), and similarly for monoids or groups. Commutative groups are also called Abelian. A commutative semigroup \( G \) is often “written additively”, meaning that \( + \) is used for the operation symbol; if \( G \) is a monoid \( 0 \) is used for the identity, and if \( G \) is a group \( -x \) is used for the inverse.
A semigroup $G$ is said to be finite or infinite according to whether its underlying set is. If $G$ is finite the cardinality $|G|$ of $G$ is called the order of $G$.

In a monoid define $x^0 = e$ and $x^{n+1} = x^n \cdot x$ for $n \in \mathbb{N}$. In a group define $x^{-n} = (x^{-1})^n$ for $x \in \mathbb{N}$, thus defining $x^n$ for all $n \in \mathbb{Z}$; note that the notation is consistent when $n = -1$.

**Theorem 4.** If $x \in G$, $G$ a monoid, and $m, n \in \mathbb{N}$ then $x^{m+n} = x^m x^n$ and $x^{mn} = (x^m)^n$.

**Proof:** We have $x^{m+0} = x^m = x^m e = x^m x^0$; and assuming $x^{m+n} = x^m x^n$, $x^{m+(n+1)} = x^{m+n} x = (x^m x^n) x = x^m (x^n x) = x^m x^{n+1}$. For the second identity, $x^0 = e = (x^m)^0$; and assuming $x^{mn} = (x^m)^n$, $x^{m(n+1)} = x^{mn+m} = x^m x^m = (x^m)^n x^m = (x^m)^{n+1}$.

**Theorem 5.** If $x \in G$, $G$ a group, and $m, n \in \mathbb{Z}$ then $x^{m+n} = x^m x^n$ and $x^{mn} = (x^m)^n$.

**Proof:** First, $x^{-n+1} = x^{-n} x$, $n \in \mathbb{N}$; this is clear for $n = 0$, and assume it for $n$,

$$x^{-(n+1)} = (x^{-1})^{n+1} x = x^{-(n+1)} x = x^{-n}$$

The proof of theorem 4 now goes through for $m \in \mathbb{Z}$, $n \in \mathbb{N}$. Next, $x^n x^{-n} = e$, $n \in \mathbb{N}$; this is clear for $n = 0$, and assume it for $n$,

$$x^{n+1} x^{-(n+1)} = x^{n+1} (x^{-1})^{n+1} = x^n x (x^{-1})^n x^{-1} = e$$

It follows that $(x^n)^{-1} = x^{-n}$, $n \in \mathbb{N}$. Thus,

$$x^{m-n} = x^{m-n} x^n (x^n)^{-1} = x^m x^{-n}, m \in \mathbb{Z}, n \in \mathbb{N},$$

which completes the proof of the first claim. For $m \in \mathbb{Z}$, $n \in \mathbb{N}$ the second claim holds by the proof of theorem 4; and

$$x^{m(-n)} = x^{-m(n)} = x^{-m} x^n = ((x^m)^{-1})^n = (x^m)^{-n}.$$

It follows that the subgroup generated by a single element $x$ of a group consists of the elements $x^n$, $n \in \mathbb{Z}$. If the group $G$ is infinite these may be all distinct. The subgroup in this case is called the infinite cyclic group. It is generated by either of the two elements $x$, $x^{-1}$ and no other. Any two infinite cyclic groups are isomorphic, by a function taking the generator of one to either generator of the other.

**Theorem 6.** In a group, if the elements $x^n$, $n \in \mathbb{N}$, are not distinct there is a unique positive integer $o$ such that $x^o = e$ and the elements $x^i$, $0 \leq i < o$, are distinct. For any $n \in \mathbb{Z}$, $x^n = x^r$ where $r = n \mod o$.

**Proof:** Suppose $x^m = x^n$, $m > n$; then $x^{m-n} = e$. Let $o$ be the least positive integer such that $x^o = e$. If $0 \leq i < j < o$ and $x^i = x^j$ then $x^{j-i} = e$, contradicting the choice of $o$; thus, the elements $x^i$, $0 \leq i < o$, are distinct. If $n = qo + r$ then $x^n = (x^o)^q x^r = e^q x^r = ex^r = x^r$, since $e^q = e$ for any $q \in \mathbb{Z}$, as is easily shown.

If $o$ exists, the element is said to have finite order (this can happen in an infinite group). In this case $o$ is called the order $o(x)$ of the element $x$. The elements $x^i$, $0 \leq i < o(x)$, clearly form a subgroup of order $o(x)$. Clearly if two elements have the same order $o$ the subgroups they generate are isomorphic; this group is the cyclic group of order $o$. It has several generators; later (chapter 6) we will say what they are. Note that a cyclic group, finite or infinite, is commutative.

Even in an infinite group $G$, there might be a $t$ such that $x^t = 1$ for all $x \in G$; such a $t$ is called an exponent for $G$. For finite $G$, $|G|$ is an exponent (see corollary 8); there might however be exponents which are proper divisors of $|G|$.

If $S, T \subseteq G$ let $ST$ denote $\{xy : x \in S, y \in T\}$; let $xS$ denote $\{x\}S$, and similarly for $Sx$. This operation is associative, that is, $S(TU) = (ST)U$. It follows by the cancellation law that the map $x \mapsto xa$ from $S$ to $Sa$, $S \subseteq G$, $a \in G$, is injective; by definition it is surjective. In particular if $S$ is finite then $S$ and $Sa$ have the same cardinality. If $H \subseteq G$ is a subgroup, then $Ha$ is called a right coset of $H$; a subset of the form $aH$, $a \in G$, is called a left coset. Finally, note that for a subgroup $H$, $HH = H$. 22
Theorem 7. Let $H \subseteq G$ be a subgroup; then the right (left) cosets of $H$ form a partition of $G$.

Proof: Since $e \in H$ a $\in Ha$. If $x \in Ha \cap Hb$ then $x = ha = h'b$ for some $h, h' \in H$, whence $a = h^{-1}h'b$ and $a \in Hb$. But then $Ha \subseteq HHb = Hb$; reversing the roles of $a, b$ $Hb \subseteq Ha$ also, and $Ha = Hb$. The proof for left cosets is the same.

Corollary 8 (Lagrange’s Theorem). Suppose $G$ is a finite group. If $H \subseteq G$ is a subgroup, then $|H| \mid |G|$, and for $x \in G$ $o(x) \mid |G|$.

Proof: These are immediate consequences of the theorem.

If $H \subseteq G$ is a subgroup and the number of right cosets of $H$ is finite, the number is called the index of $H$ in $G$. When $G$ is finite this equals $|G|/|H|$; the order of $H$ is the cardinality of a right coset, and the index the number of right cosets. The index is also the number of left cosets.

2. Normal Subgroups. Since the right cosets of $H$ form a partition of $G$, the relation of belonging to the same right coset is an equivalence relation. To answer the question of when this is a congruence relation, we define a subgroup $H \subseteq G$ to be normal if $aHa^{-1} = H$ for all $a \in G$.

Theorem 9. The relation $\equiv_H$ of belonging to the same right coset of a subgroup $H \subseteq G$ is a congruence relation iff $H$ is a normal subgroup. Given a congruence relation $\equiv$ in $G$, the equivalence class $H_\equiv$ of $e$ is a normal subgroup. The maps $H \mapsto H_\equiv$ and $\equiv \mapsto H_\equiv$ are inverse to each other.

Proof: First, $y \equiv_H x$ iff $x \in Hy$ iff $xy^{-1} \in H$; for $\equiv_H$ to be a congruence relation it is necessary that $y \equiv_H xy$ whenever $x \equiv_H e$, from which $yHy^{-1} \subseteq H$ for any $y$ follows, and $yHy^{-1} = H$ is immediate. Conversely suppose $yHy^{-1} = H$ for all $y \in G$; then $y_2Hy_1^{-1} = y_2y_1^{-1}y_1H = y_2y_1^{-1}H$, and so $y_2y_1^{-1} \in H$ iff $y_2Hy_1^{-1} = H$. Hence if $x_2x_1^{-1}, y_2y_1^{-1} \in H$ then $y_2x_2x_1^{-1}y_1^{-1} \in y_2Hy_1^{-1} = H$; also if $x_2x_1^{-1} \in H$ then $x_2x_1^{-1} = H$ so $x_2^{-1}x_1 \in H$. This shows that $\equiv_H$ is a congruence relation if $H$ is normal. Certainly if $x, y \equiv e$ where $\equiv$ is a congruence relation then $xy \equiv e$ and $x^{-1} \equiv e$ since $ee = e^{-1} = e$; hence by exercise 2 (or $e \equiv e$) $H_\equiv$ is a subgroup. Further if $x \equiv e$ then $xy^{-1} \equiv e$ and so $H_\equiv$ is normal. Next, $y \equiv_{H_\equiv} x$ iff $xy^{-1} \in H_\equiv$ iff $xy^{-1} \equiv e$ iff $y \equiv x$. Finally $x \in H_\equiv$ iff $x \equiv_H e$ iff $x \in eH = H$.

Thus, there is a one to one correspondence between congruence relations and normal subgroups of a group $G$. Given a homomorphism $h : G \rightarrow H$, the normal subgroup corresponding to its kernel congruence relation is $\{x : h(x) = e\}$. This is also called the kernel of the homomorphism. Note that for $a \in G$, $[a]$ under the congruence $\equiv$ equals $K_a a$. For the remainder of this text by the kernel of a group homomorphism we mean the normal subgroup. We write $\text{Ker}(h)$ and $\text{Im}(h)$ for the kernel and image of a homomorphism $h$. We write $G/K$ for $G/\equiv_K$ when $K$ is a normal subgroup.

The setwise product of two right cosets of a normal subgroup is again a right coset; indeed $KaKb = Kaa^{-1}Kab = KKab = Kab$. Thus, if $G/K$ is given as the right cosets the multiplication operation is setwise product. For a normal subgroup $K$, the left and right cosets are the same; indeed, $Ka = aKa^{-1}a = aK$.

Finally note that any subgroup of a commutative group is commutative; it is also normal, and the quotient group exists.

An alternative (and somewhat older) term for the quotient group $G/K$ is the factor group. The notation $K \trianglelefteq G$ is used to denote that $K$ is a normal subgroup of $G$. We will not introduce a notation for $H$ being a subgroup of $G$, using instead “if $H \subseteq G$ is a subgroup” or some such.

3. An example of a group. Many important finite groups occur as “symmetry groups” of geometric figures; these are the rigid transformations of the plane (or space) which map the figure to itself. For example, the reader may verify that for $n \geq 3$ the symmetry group of a regular $n$-gon has order $2n$, with the $n$ rotations (counting the identity map as a rotation) comprising a subgroup of order $n$. This group is called the dihedral group.
When \( n = 3 \) the dihedral group is the group of all permutations of 3 points. Let \( S_3 \) be the group of bijections of the set \( \{1, 2, 3\} \). Letting \( a, b \) be given by

\[
\begin{array}{c|ccc}
   x & 1 & 2 & 3 \\
\hline
   a(x) & 2 & 3 & 1 \\
   b(x) & 2 & 1 & 3 \\
\end{array}
\]

the elements of \( S_3 \) may be written as \( e, a, a^2, b, ba, ba^2 \). The equations \( a^3 = e, b^2 = e, bab^{-1} = a^2 \) hold, and completely determine the “multiplication table” of the group; this is

\[
\begin{array}{cccccc}
e & e & a & a^2 & b & ba & ba^2 \\
e & e & a & a^2 & b & ba & ba^2 \\
a & a & a^2 & e & ba^2 & b & ba \\
a^2 & a^2 & e & a & ba & ba^2 & b \\
b & b & ba & ba^2 & e & a & a^2 \\
ba & ba & ba^2 & b & a^2 & e & a \\
ba^2 & ba^2 & b & ba & a & a^2 & e
\end{array}
\]

For any \( a \in G \) the function \( \sigma_a : G \to G \) defined by \( x \mapsto axa^{-1} \) is readily verified to be an automorphism of \( G \). It is injective (if \( axa^{-1} = aya^{-1} \) then \( x = y \)), and surjective (\( (a^{-1}xa)a^{-1} = x \)). Also \( a(xy)a^{-1} = axa^{-1}aya^{-1} \) so it is a homomorphism (using exercise 2). This operation on \( G \) is called conjugation by \( a \). An automorphism determined by conjugation by some element is called an inner automorphism; a group may possess additional automorphisms, which are called outer.

Let \( \sigma_a \) denote conjugation by \( a \), and \( \text{Aut}(G) \) the group of automorphisms of \( G \). While not of great importance at this point, it is easily verified that the function \( h : G \to \text{Aut}(G) \) defined by \( h(a) = \sigma_a \), is a homomorphism (exercise 3). Its image is the inner automorphisms; its kernel is those elements \( a \in G \) such that \( \sigma_a(x) = x \) for all \( x \in G \), that is, such that \( xa = ax \) for all \( x \in G \). This normal subgroup of \( G \) is called the center of \( G \); clearly it is commutative. Say that two elements \( x, y \in G \) commute if \( xy = yx \); the center of \( G \) can also be described as those elements which commute with every element of \( G \). A group is commutative iff it equals its center. In this case the inner automorphisms consist of the identity automorphism alone, so if a commutative group possesses automorphisms they are outer automorphisms.

The additive group of the integers possesses an automorphism, namely \( x \mapsto -x \). Indeed, \( x \mapsto x^{-1} \) is an automorphism of any commutative group; it may be the identity, indeed this is so iff every element has order 2. In \( S_3 \), note that an automorphism is completely determined by where it maps the elements \( a \) and \( b \) defined above, since these generate the group. For the inner automorphisms, this is as follows.

\[
\begin{array}{cccccc}
   \sigma_e & \sigma_a & \sigma_{a^2} & \sigma_b & \sigma_{ba} & \sigma_{ba^2} \\
\hline
   a & a & a & a & a^2 & a^2 & a^2 \\
   b & ba & ba^2 & b & ba^2 & ba
\end{array}
\]

Only \( \sigma_e \) fixes \( G \), so the center of \( S_3 \) is trivial; equivalently, distinct elements yield distinct conjugations. Now, an automorphism must map an element to an element of the same order (exercise 4), so in this case must map \( a \) to one of \( a, a^2 \), and \( b \) to one of \( b, ba, ba^2 \). Thus, for \( S_3 \), all automorphisms are inner.

4. Homomorphism theorems. The collection of subgroups of a group \( G \) form an algebraic closure system when ordered by inclusion. This is also true of the normal subgroups. If \( C \) is a collection of normal subgroups, \( K = \bigcap C \), and \( y \in K, x \in G \) then \( xyx^{-1} \in H \) for all \( H \in C \), so \( xyx^{-1} \in K \); thus \( K \) is normal. If \( D \) is a directed set and \( K = \bigcup C \), \( y \in K, x \in G \) then \( y \in H \) for some \( H \in C \), whence \( xyx^{-1} \in H \) and \( xyx^{-1} \in K \), so \( K \) is again normal. We will use \( \sqcup \) for the join in the lattice of subgroups.

An alternative proof that the normal subgroups form an algebraic closure system is as follows. The bijection \( K \mapsto \equiv K \) from normal subgroups to congruence relations is readily seen to preserve all meets and joins. The claim then follows by exercise 3.7.
Lemma 10. Suppose \( h : G \rightarrow G' \) is a homomorphism, and \( S, T \subseteq G \); then \( h[ST] = h[S]h[T] \).

Proof: We have \( x \in h[ST] \) iff \( x = h(st) \) for some \( s \in S, t \in T \), iff \( x = h(s)h(t) \) for some \( s \in S, t \in T \), iff \( x \in h[S]h[T] \).

Suppose that \( K \triangleleft G \) and \( H \subseteq G \) is a subgroup with \( K \subseteq H \). It is immediate that \( K \triangleleft H \); the map \( Kx \rightarrow Kx \) is an embedding of \( H/K \) in \( G/K \), by which \( H/K \) may be considered a subgroup of \( G/K \). It will follow by theorem 11 that if \( H \triangleleft G \) then \( H/K \triangleleft G/K \).

Theorem 11. Suppose \( h : G \rightarrow G' \) is a homomorphism with kernel \( K \) and image \( h[G] \); let \( C \) be the collection of subgroups \( H \subseteq G \) with \( K \subseteq H \), and \( C' \) the collection of subgroups \( H' \subseteq G' \) with \( H' \subseteq h[G] \). The map \( H \mapsto h[H] \) is a one to one correspondence between \( C \) and \( C' \); the inverse map is \( H' \mapsto h^{-1}[H'] \).

For \( H \in C \), \( H/K \) is mapped to \( h[H] \) by the canonical isomorphism. For \( H \in C \), \( H \triangleleft G \) iff \( h[H] \triangleleft h[G] \), and if \( H \triangleleft G \) then \( G/H \) is isomorphic to \( h[G]/h[H] \), whence to \( (G/K)/(H/K) \).

Proof: The first claim follows from the claim that if \( H \subseteq C \) then \( h^{-1}[h[H]] = H \). The latter is equivalent to the claim that whenever \( h(x) \in h[H] \), already \( x \in H \); but if \( h(x) \in h[H] \) then \( h(x) = h(w) \) for some \( w \in H \), whence \( x \equiv_K w \), or \( x \in K \subseteq H \). For the second claim, \( Ku \) maps to \( h(u) \), and if \( u \in H \) then \( h(u) \in h[H] \); it follows that the image of \( H/K \) is exactly \( h[H] \). For the third claim, \( h[H] \) is normal in \( h[G] \) iff \( h(x)h[H]h(x^{-1}) = h[H] \) for any \( x \in G \). By lemma 10 the left side is \( h[xHx^{-1}] \). Since \( h \) is bijective on \( C \) and \( xHx^{-1} \subseteq C \) this equals \( h[H] \) iff \( xHx^{-1} = H \). If \( H \subseteq C \) is normal we claim that the map \( g : G/H \rightarrow h[G]/h[H] \) defined by \( g(aH) = h(a)h[H] \) is an isomorphism. It is a homomorphism, since

\[
g(aHbH) = g(abH) = h(ab)h[H] = h(a)h[bH]h[H] = g(aH)g(bH).
\]

It is surjective, since every element of \( h[G]/h[H] \) is \( h(a)h[H] \) for some \( a \in G \). If \( g(aH) = h[H] \) then \( h(a) \in h[H] \); but we saw above that then \( a \in H \). This implies that the kernel of \( g \) is trivial, so \( g \) is injective.

Theorem 12. Suppose \( K \triangleleft G \) and \( H \subseteq G \) is a subgroup; let \( \eta : G \rightarrow G/K \) be the canonical epimorphism.

a. \( H \cup K = KH \).

b. \( KH \) is the unique subgroup \( H' \subseteq G \) with \( K \subseteq H' \subseteq G \) and \( \eta[H'] = \eta[H] \).

c. \( H \cap K \triangleleft H \).

d. \( KH/K \) is isomorphic to \( H/K \cap K \), via \( Kx \mapsto H \cap Kx \), with inverse map \( (H \cap K)x \mapsto Kx \).

Proof: That \( H \cup K = KH \) follows because \( aK = Ka \) for any \( a \in G \). Now, \( \eta[HK] = \eta[H]\eta[K] = \eta[H]\{e\} = \eta[H] \), and certainly \( K \subseteq KH \) (uniqueness follows by theorem 11). If \( h \in H, k \in H \cap K \) then \( hkh^{-1} \in H, K \). For the last claim, note that any \( Kx, x \in KH \), equals \( Kh \) for some \( h \in H \), since \( x = kh \) for some \( h \in H, k \in K \); and that if \( h \in H \) then \( (H \cap K)h = H \cap Kh \). Thus,

\[
Kx = Kh \rightarrow H \cap Kh = (H \cap K)h \rightarrow Kh, \quad \text{and} \quad (H \cap K)h \rightarrow Kh \rightarrow H \cap Kh = (H \cap K)h.
\]

5. Product of Groups. The product \( \times_i G_i \) of semigroups \( G_i \) is defined to be the product of the structures. This construction is also called the direct product. Since the axioms are equations, the product structure is a semigroup. The analogous facts hold for monoids or groups. In the case of monoids we consider the empty product to be the trivial monoid. The proofs of the claims of this section will be left as an exercise.

If the \( G_i \) are monoids there is an embedding \( \mu_i : G_i \rightarrow \times_i G_i ; \mu_i(x) \) is the sequence \( \langle x_i \rangle \) where \( x_i = x \) and \( x_j = 1 \) for \( j \neq i \). This map is a homomorphism. It is a “canonical” embedding of \( G_i \) into the product, where the term is applied loosely to mean that it is natural and has many properties which are a result of
general facts. Canonical embeddings are common in algebra, and it is common practice to identify the range of the embedding with the domain, so that the domain may be considered as a substructure of the codomain.

Via the canonical embedding, we may consider $G_i$ as a submonoid of the product. Note that $\pi_i \mu_j$ is the identity map on $G_i$ if $j = i$, and if $n \neq i$ it maps every element to 1. When the $G_i$ are written additively, this latter map is called the 0 map.

If the $G_i$ are groups, $G_i$ considered as a subgroup of the product is a normal subgroup. The quotient $\times_i G_i / G_i$ is isomorphic to $\times_{j \neq i} G_j$. More generally if $I \subseteq I$, then $\times_i \in J G_i$ is a normal subgroup, and the quotient is isomorphic to $\times_{i \not\in J} G_i$.

Note that the canonical embedding relies on the presence of the identity element, and does not apply to products of structures in general. In general, in a product of structure $\times_i S_i$ the relation $x_i = y_i$ is a congruence relation, and the quotient is isomorphic to $S_i$.

If the $G_i$ are monoids, the subset of the product consisting of the elements which are 1 in all but finitely many components is clearly a submonoid; if the $G_i$ are groups it is a subgroup. This monoid (or group) is called the direct sum of the $G_i$; the notation $\oplus_i G_i$ is used to denote it. If the index set is finite, the product and the direct sum coincide.

If the $G_i$ are commutative the direct sum, together with the injections, satisfies a diagram property. Namely, if $S'$ is a commutative group (or monoid) and $m'_i : G_i \mapsto S'$ are homomorphisms then there is a unique homomorphism $f : \oplus_i G_i \mapsto S'$ such that $f \circ \mu_i = m'_i$. Writing the groups additively, clearly $f(\langle x_i \rangle) = \sum_{x_i \neq 0} m'_i(x_i)$. We leave it as an exercise to show that this diagram property characterizes the direct sum up to isomorphism.

**Exercises.**

1. Show that if a structure $\langle S, \cdot, e, \cdot^{-1} \rangle$ satisfies the axioms $x(yz) = (xy)z, ex = x, x^{-1}x = e$ then it is a group. Hint: First show that $x^2 = e$ implies $x = e$.

2. Show the following, for a group $G$.
   a. If $H \subseteq G$ is closed under $\cdot$ and $^{-1}$ is a subgroup.
   b. If $G$ is finite and $H \subseteq G$ is closed under $\cdot$ then $H$ is a subgroup.
   c. If $H$ is a group and $f : G \mapsto H$ preserves $\cdot$ then $f$ is a homomorphism.

3. Show that the function $h$ from $G$ to $\text{Inn}(G)$ defined by $h(a) = \sigma_a$ is an epimorphism. Hint: First show that $\sigma_{ab} = \sigma_a \circ \sigma_b$.

4. Suppose $\phi : G \mapsto G$ is an automorphism of the group $G$ and $x \in G$. Show that $o(\phi(x)) = o(x)$.

5. Prove the claims made about the product or direct sum of monoids or groups.
5. Permutation groups.

Suppose $M$ is a monoid and $S$ a set. A map $\cdot : M \times S \to S$ is called an action of $M$ on $S$ provided $(ab)x = a(bx)$ and $1x = x$, for all $x \in S$ and $a, b \in M$, where as usual we have omitted the operation symbol, and written $1$ for the identity of $M$. Each $a \in M$ induces a function $\psi_a : S \to S$ defined by $\psi_a(x) = ax$. We can consider an action as a two sorted structure. Alternatively, we can consider the $\psi_a$ as (not necessarily distinct) functions of a structure with domain $S$; this invariably leads to correct definitions of substructures of $S$, etc.

The requirement $(ab)x = a(bx)$ is equivalent to $\psi_{ab} = \psi_a\psi_b$, and the requirement $1x = x$ to $\psi_e = \iota_S$. The map $a \mapsto \psi_a$ is thus a homomorphism from $M$ to the monoid of functions on $S$; indeed the definition of an action is exactly the requirement that this be so. The action is called faithful when this homomorphism is injective, that is, when the $\psi_a$ are distinct for distinct $a$.

In the case that $M$ is a group $G$, the homomorphism $a \mapsto \psi_a$ is a group homomorphism, so the $\psi_a$ are bijections. Bijections are also called permutations, especially when $\psi$ is injective.

We have already seen some examples of actions. The monoid of functions on a set is clearly a faithful action of the monoid on the set. A group acts on itself by conjugation; this action may or may not be faithful. A group acts on itself in another important way, by left translation; here $\psi_a(x) = ax$. A group acts on itself in another important way, by left translation; here $\psi_a(x) = ax$. The order of $\text{Sym}(n)$ is the set of the $\psi_a$ isomorphisms on an $n$-element set appears as a subgroup of $\text{Sym}(n)$.

Two permutation groups $(G_1, S_1)$ and $(G_2, S_2)$ are called isomorphic if there is an isomorphism $j : G_1 \to G_2$, and a bijection $p : S_1 \to S_2$, with $j(a)p(x) = p(ax)$ for all $a \in G, x \in S$. Equivalently, define $j \times p : G_1 \times S_1 \to G_2 \times S_2$ by $j \times p(\langle a, x \rangle) = \langle j(a), p(x) \rangle$; this map must preserve the action, in the sense that $p \circ j^{-1} = j \circ (j \times p)$.

Let $\text{Sym}_n$ denote the group of permutations of $S$. If $S, T$ are finite sets of the same cardinality $n$ then $\text{Sym}_S, \text{Sym}_T$ are isomorphic permutation groups (exercise). In particular the groups are isomorphic, and the group is called the symmetric group of degree $n$. The map $a \mapsto \psi_a$, together with the identity on $S$, yields an isomorphism from any permutation group on $S$ to a subgroup of $\text{Sym}_S$. Thus up to isomorphism every permutation group on an $n$ element set appears as a subgroup of $\text{Sym}_n$ for a fixed $n$ element set $S$. The cardinality of $S$ is called the degree of the permutation group. Isomorphic groups may have representations as permutation groups of different degrees, or nonisomorphic representations of the same degree; for example a representation of degree $n$ yields a representation of degree $m$ for any $m \geq n$, where the additional elements are fixed by every permutation. The following theorem shows that any group has at least one representation as a permutation group.

**Theorem 1 (Cayley’s theorem).** A group is isomorphic to the group of its left translations.

**Proof:** The map $a \mapsto \psi_a$ has already been shown to be an injective homomorphism; it is surjective since the group of left translations is defined to be the set of the $\psi_a$.

The order of $\text{Sym}_S$ where $S$ is an $n$ element set may be computed as follows. We may assume $S = \{1, \ldots, n\}$. For $\pi \in \text{Sym}_S$ there are $n$ choices for $\pi(1)$; having fixed $\pi(1)$ there are then $n - 1$ choices for $\pi(2)$; in general having selected $\pi(i), 1 \leq i < k$, there are $n - (k - 1)$ choices for $\pi(k)$. Altogether there are $n(n - 1) \cdots 1$, permutations $\pi$ of an $n$ element set. That is, $o(\text{Sym}_n) = n!$.

Given $x \in S$ and $\pi \in \text{Sym}_S$, let $k$ be the least positive integer such that $\pi^k(x) = x$, if such exists. The elements $\pi^i(x), 0 \leq i < k$, are distinct, for if $\pi^i(x) = \pi^j(x), 0 \leq i \leq j < k$, then $\pi^{j-i}(x) = x$ and $0 \leq j - i < k$, so $j = i$. These elements are called the orbit of $x$ under $\pi$. Clearly $\pi^k(x) = x$ for any $q \in \mathbb{Z}$, and for any $i \in \mathbb{Z}$, $\pi^i(x) = \pi^r(x)$ where $r = i \mod k$. If $y$ is in the orbit of $x$ then the orbit of $y$ and the orbit of $x$ are equal; for $y = \pi^i(x)$ for some $x$, so $\pi^i(y) = \pi^{i+l}(x)$, and as $i$ runs from $0$ to $k - 1$ so does $i + l \mod k$. The orbits thus form a partition of $S$. 27
Fixing $S = \{1, \ldots, n\}$ denote $\text{Sym}_S$ by $\text{Sym}_n$. An element $\pi$ of $\text{Sym}_n$ may be denoted by giving the list $\pi(1) \ldots \pi(n)$; the members of $\text{Sym}_3$ represented this way are $123, 132, 213, 231, 312, 321$. The cycle decomposition is another representation. Each orbit is written in order $x\pi(x)\pi^2(x) \ldots \pi^{k-1}(x)$ and enclosed in parentheses; this is called a cycle of the permutation. The cycles are written in some order. The members of $\text{Sym}_3$ represented this way, in the same order as above, are $(1)(2)(3), (1)(23), (12)(3), (123), (132), (13)(2)$. Note that a cycle of length $k$ may be written in $k$ different ways, and the cycles may be written in any order. Cycles of the form $(x)$ are often omitted.

A cycle represents a permutation, namely that where the other elements are fixed. If the cycle has length $k$ the permutation has order $k$, since successive powers map the first element to the second, third, and so forth. The permutation represented by a cycle decomposition is the product of the permutations successively determining the image under the product of the suffixes of the sequence of cycles.

The notation may be generalized, so that the cycles may have common members. This denotes the permutation obtained by multiplying the permutations denoted by the cycles. The result of multiplying two cycles $C_1$ and $C_2$ may determined as follows. In the product $C_1C_2$, $C_2$ is applied first; if $C_2(i) = j$ and $C_1(j) = k$ then $C_1C_2(i) = k$. For example, $(123)(134) = (234)$, since $1 \mapsto 3 \mapsto 1, 2 \mapsto 2 \mapsto 3$, etc. One can determine the effect of several cycles on an element $i$ by tracing through the cycles from right to left, successively determining the image under the product of the suffixes of the sequence of cycles.

Given $\pi, \rho \in \text{Sym}_n$, the cycle representation of $\rho\pi\rho^{-1}$ is determined from that of $\pi$ as follows. The element $\rho(i)$ is mapped to $\rho(\pi(i))$ by $\rho\pi\rho^{-1}$. Hence if each element $i$ in the cycle representation of $\pi$ is replaced by $\rho(i)$, the result is the cycle representation of $\rho\pi\rho^{-1}$.

Define the sign $\text{sg}(\pi)$ of a permutation $\pi$ to be

$$\prod_{1 \leq i < j \leq n} \text{sg}_\pi(i, j)$$

where $\text{sg}_\pi(i, j) = +1$ if $\pi(i) < \pi(j)$ ($\pi$ preserves the order of $i$ and $j$), and $\text{sg}_\pi(i, j) = -1$ if $\pi(i) > \pi(j)$ ($\pi$ reverses the order of $i$ and $j$). Permutations $\pi$ with $\text{sg}(\pi) = +1$ reverse an even number of pairs and are called even; those with $\text{sg}(\pi) = -1$ reverse an odd number of pairs and are called odd.

**Theorem 2.** $\text{sg}(\pi\rho) = \text{sg}(\pi)\text{sg}(\rho)$.

**Proof:** Recall that $|S|$ denotes the cardinality of the finite set $S$. Let

$$S_\alpha = \{(i, j) : \text{sg}_\pi(i, j) = \alpha\}$$

and let

$$S_{\alpha\beta} = \{(i, j) : \text{sg}_\pi(i, j) = \alpha \text{ and } \text{sg}_\pi(\tau(i), \tau(j)) = \beta\},$$

where $\alpha$ or $\beta$ is $+$ (for $+1$) or $-$ (for $-1$). Then the $S_\alpha$ are disjoint and their union is $\{1, \ldots, n\}$; this is also true of the $S_{\alpha\beta}$. Further $\text{sg}(\tau) = (-1)^{|S_+|}$, $\text{sg}(\sigma) = (-1)^{|S_+ \cup S_-|}$, and $\text{sg}(\sigma\tau) = (-1)^{|S_+ \cup S_-|}$. The lemma is equivalent to

$$(-1)^{|S_+| + |S_+ \cup S_-|} = (-1)^{|S_+ \cup S_-|},$$

which follows since the exponent of the left side is $|S_+| + 2|S_-| + |S_+|$. A transposition is defined to be a permutation which exchanges two elements, i.e., for some $i, j \pi(i) = j, \pi(j) = i$, and $\pi(k) = k$ for $k \neq i, j$. Clearly $\text{sg}(\pi) = -1$. We leave it as an exercise to show that the transpositions generate the group $S_n$, i.e., that every permutation can be written as a product of transpositions. The number of transpositions in any such product is even if the permutation is even, and odd if the permutation is odd.
Suppose the group $G$ acts on the set $S$. The orbit under $G$ of an element $x \in S$ is defined to be \{ax : a \in G\}. What we called the orbit above is the orbit under the cyclic group generated by the given permutation. We continue to consider finite $S$, although many facts are true of infinite $S$ as well.

**Theorem 3.** The orbits of the elements of $S$ form a partition of $S$.

**Proof:** Let $x \equiv y$ if $x, y$ belong to the same orbit; this is trivially reflexive and symmetric. It holds iff $y = ax$ for some $a \in G$; if it holds then $x = au, y = bu$ for some $a, b \in G, u \in S$, whence $y = ba^{-1}x$, and the other direction is trivial. If $x \equiv y, y \equiv z$ then $y = ax, z = by$ for some $a, b \in G$, whence $z = bax$ and $z \equiv x$.

The parts of the partition corresponding to $\equiv$ are clearly the orbits.

$G$ acts on each orbit separately. The action is called transitive if there is only a single orbit, that is, if for any $x, y \in S$ there is an $a \in G$ with $ax = y$. This follows, provided for some $x \in S$ and all $y \in S$ there is an $a \in G$ with $ax = y$. The stabilizer $\text{Stab}_x$ of an element $x \in S$ is defined to be those $a \in G$ such that $ax = x$. Two subsets $S_1, S_2$ of a group $G$ are said to be conjugate if $S_2 = aS_1a^{-1}$ for some $a \in G$.

**Theorem 4.** Suppose $G$ is a group of permutations of $S$, and let $T$ be the orbit containing $x, y \in S$. The map $ax \mapsto a\text{Stab}_x$ is well defined, and is a 1-1 correspondence between $T$ and the left cosets of $\text{Stab}_x$; the induced action of $G$ on the left cosets is left multiplication. In particular, for finite $S$ the size of the orbit of $x$ (which is the degree if $G$ is transitive) equals the index of $\text{Stab}_x$, and so divides $|G|$. If $y \in T$ then $\text{Stab}_x$ and $\text{Stab}_y$ are conjugate.

**Proof:** We have that $ax = bx$ iff $b^{-1}a \in \text{Stab}_x$, so the map is well defined and injective; it is clearly onto. By definition of the induced action $b(a\text{Stab}_x)$ is the coset corresponding to $b(ax) = (ba)x$, which is $(ba)\text{Stab}_x$, so the induced action is left multiplication. If $y = bx$ then $a \in \text{Stab}_y$ iff $ay = y$ iff $abx = bx$ iff $b^{-1}abx = x$ iff $b^{-1}ab \in \text{Stab}_x$ iff $a \in b\text{Stab}_xb^{-1}$, proving the last claim.

A group $G$ acting on a set $S$ induces an action on the $t$-tuples of elements of $S$, and on the $t$ element subsets. An element $g \in G$ is said to stabilize a subset $T \subseteq S$ if $g[T] = T$, that is, if $g$ stabilizes the set in the action of $G$ on the subsets; $g$ is said to fix, or pointwise fix, $T$ if $g(x) = x$ for all $x \in T$ (equivalently if $g$ stabilizes $T$ ordered in any way, in the action of $G$ on the tuples). The terminology of permutation groups is generalized to these actions; thus, we may speak of the orbit or stabilizer of a subset or tuple. Theorem 4 applies immediately; for example letting $\text{Stab}_T$ denote the group of elements of $G$ which (setwise) stabilize $T$, the subsets in the orbit of $T$ are in 1-1 correspondence with the cosets of $\text{Stab}_T$; and the stabilizer of any subset in the orbit is conjugate to $\text{Stab}_T$.

A permutation group $G$ on a set $S$ is called
- semiregular if the stabilizers of a point all have size 1;
- regular if it is transitive and semiregular;
- $t$-transitive if it is transitive on the $t$-tuples;
- sharply $t$-transitive it is regular on the $t$-tuples;
- $t$-homogeneous if it is transitive on the $t$ element subsets.

If $S$ is finite, $G$ is semiregular iff the blocks of transitivity all have the same size as the group; $G$ is regular if $|G| = |S|$ and it is either transitive or semiregular. For finite $S$, the term half-transitive is used to denote that the blocks of transitivity have the same size.

If $G$ is (sharply) $m$-transitive on $S$ then clearly any stabilizer $\text{Stab}_x$ is (sharply) $m$-transitive. The converse follows if $G$ is transitive, and $|S| > m$. That $G$ is $m$-transitive also follows if the stabilizer of every $m$ element subset is transitive, and $|S| > m + 1$. To see this, note that inductively $G$ is $i$-transitive for all $i \leq m + 1$.

**Theorem 5.** If $G$ acts on $S$ regularly then this action is permutation group isomorphic to the action of $G$ on itself by left translation.
Proof: The cosets of the stabilizer of some element $x$ are just the group elements, and these are in 1-1 correspondence with the elements of $S$, since $G$ is transitive. The correspondence maps $y \in S$ to the unique $g \in G$ such that $g(x) = y$, which we denote $g_y$. Together with the identity map on $G$, this yields the required permutation group isomorphism, since $g(u) = v$ iff $gg_u(x) = g_v(x)$ iff $gg_u = g_v$.

Thus, there is only one regular representation of a group $G$; we may call it the regular representation. It is also called the left regular representation, since the action is left multiplication. One may define actions on the right, by requiring $(xa)b = x(ab)$. Note that this is not the same as an action on the left; we now have $\psi_a \psi_b = \psi_{ab}$. Right multiplication is a right action, and yields a permutation group which is in bijective correspondence with $G$, reversing multiplication rather than preserving it. This action is called the right regular representation of $G$. A map between semigroups which reverses multiplication is called an ; thus, the right regular representation is anti-isomorphic to $G$. Many authors write functions on the right, and define composition so that $x(f \circ g) = (xf)g$; in this case it is the right regular representation which is isomorphic.

Theorem 6. If $G$ is commutative and acts transitively on $S$ then it acts regularly on $S$.

Proof: Since the stabilizers are all conjugate they are all equal; any stabilizer thus fixes every element of $S$, and so is the trivial group.

Exercises.

1. Show that if $S, T$ are finite sets of the same cardinality then $\text{Sym}_S, \text{Sym}_T$ are isomorphic permutation groups.
2. Show that $e, (12), (13), (23)$ and $e, (12)(34), (13)(24), (14)(23)$ are isomorphic groups.
3. Show that the transpositions $(1i), 2 \leq i \leq n$, generate $\text{Sym}_n$. 
6. Rings.

1. Basic definitions. A ring is a structure \( \langle R, +, \times, 0, 1, - \rangle \) (+, \times binary, − unary, 0, 1 constants) such that

- \( \langle R, +, 0, - \rangle \) is a commutative group;
- \( \langle R^\times, \times, 1 \rangle \) is a monoid; and
- \( x(y + z) = xy + xz \) and \( (y + z)x = yx + zx \).

In the last axiom, as usual we have omitted the \( \times \) sign; it is called the distributive law. \( \langle R, +, 0, - \rangle \) is called the additive group of the ring; the additive inverse of \( x \) is written \( -x \), and \( x + (-y) \) is written \( x - y \).

\( \langle R^\times, \times, 1 \rangle \) is called the multiplicative monoid.

**Theorem 1.** The following hold in any ring:

- \( x0 = 0x = 0 \);
- \( x(-z) = (-x)z = -(xz) \);
- \( -(x)(-y) = xy \);
- \( x(y - z) = xy - xz \) and \( (y - z)x = yx - zx \);
- \( -1)x = -x. \)

**Proof:** Exercise.

If \( 0 = 1 \) the ring equals \( \{0\} \), and is called trivial. The definition given here is the most common, but there are variations. The requirement that a multiplicative identity exist is sometimes omitted. Structures where multiplication need not be associative are sometimes called rings.

Since the axioms are universally quantified a substructure of a ring is itself a ring, and is called a subring. The subrings of a ring form an algebraic closure system. A map between rings which preserves \(+, \times, 1\) also preserves \(0, -\) and is a homomorphism. Images or inverse images under a homomorphism of subrings are subrings. The elements of a ring \( R \) which possess a multiplicative inverse form a group under multiplication, called the group of units; we denote this as \( \text{Units}(R) \). These facts all follow by universal algebra or group theory.

A familiar example of a ring is the integers \( \mathbb{Z} \). For a more abstract example, let \( A \) be any commutative group, written additively. The "pointwise" addition operation can be defined on the monoid \( \text{End}(A) \) of endomorphisms of \( A \), by defining \( (f + g)(x) = f(x) + g(x) \). The endomorphisms with this operation form a commutative group; the additive identity is the function whose value at any argument is 0, and \(-f\) is defined by \( (-f)(x) = -f(x) \). Finally,

\[
 f \circ (g + h)(x) = f((g + h)(x)) = f(g(x) + h(x)) = f(g(x)) + f(h(x)) \\
 = f \circ g(x) + f \circ h(x) = (f \circ g + f \circ h)(x),
\]

verifying one distributive law; the other is similar.

In fact, any ring \( R \) is isomorphic to a subring of \( \text{End}(A) \), where \( A \) is the additive group of \( R \), since the map \( a \mapsto \psi_a \) where \( \psi_a(x) = ax \) is a monomorphism of rings. Indeed, it preserves \(+\) (\( \psi_{a+b} = \psi_a + \psi_b \)), \( \times \) (\( \psi_{ab} = \psi_a \circ \psi_b \)), and \( 1 \) (\( \psi_1(x) = x \), all \( x \)); and if \( \psi_a(x) = x \) for all \( x \in R \) then \( a = 1 \).

2. Ideals. An ideal \( I \) in a ring \( R \) is a subgroup of the additive group which is closed under left and right multiplication by \( R \), that is, such that \( ax, xa \in I \) whenever \( a \in R, x \in I \). For subsets \( S, T \subseteq R \) define \( S + T = \{ s + t : s \in S, t \in T \} \) and \( ST = \{ st : s \in S, t \in T \} \), with as usual \( x + T \) for \( \{x\} + T \) and \( xT \) for \( \{x\}T \). These operations are associative; \(+\) is commutative; and \( \times \) distributes over \(+\) from the right and left. A coset of an ideal \( I \) is defined to be \( x + I \) where \( x \in R \).
Theorem 2. Given an ideal \( I \) in a ring \( R \) the relation \( \equiv_I \) of belonging to the same coset of \( I \) is a congruence relation. Given a congruence relation \( \equiv \) in \( R \), the equivalence class \( I_\equiv \) of 0 is an ideal. The maps \( I \mapsto \equiv_I \) and \( \equiv \mapsto I_\equiv \) are inverse to each other.

Proof: If \( I \), it follows by group theory that \( \equiv_I \) is a congruence relation in the additive group, and that \( x \equiv_I y \) iff \( x - y \in I \). If \( x_2 - x_1, y_2 - y_1 \in I \) then \( x_2y_2 - x_1y_1 = (x_2 - x_1)y_2 + x_1(y_2 - y_1) \in I \), so \( \equiv_I \) respects \( \times \) and is a congruence relation in the ring. Given \( \equiv \), it follows by group theory that \( I_\equiv \) is a subgroup of the additive group. If \( x \equiv 0 \) and \( a \in R \) then \( ax \equiv 0 \) and \( xa \equiv 0 \); \( I_\equiv \) is thus an ideal. The last claim follows by group theory.

Similarly to groups, by the kernel of a ring homomorphism we mean the ideal rather than the congruence relation. The ideals form an algebraic closure system (by exercise 3.7 as for groups). The ideal generated by a subset \( S \subseteq R \) is all elements which can be written in the form \( a_1x_1b_1 + \cdots + a_nb_n \) for some \( a_1, \ldots, a_n, b_1, \ldots, b_n \in R \) and \( x_1, \ldots, x_n \in S \) (these must be contained in any ideal containing \( S \), and comprise an ideal). The quotient ring consists of the cosets of \( I \) in the additive group; the canonical homomorphism maps \( x \) to \( x + I \). In the quotient ring addition corresponds to setwise addition, i.e., \( (x + y) + I = (x + I) + (y + I) \) (it is not true that \( xy + I = (x + I) \times (y + I) \)).

An ideal \( I \subseteq R \) is called proper if \( I \subset R \); and trivial if \( I = \{0\} \), else nontrivial. Note that \( I = R \) iff \( I \) contains 1 iff \( I \) contains a unit. Also \( Ix = I \) if \( x \) is a unit.

Some homomorphism properties obeyed by rings are as follows. Suppose \( h : R \rightarrow R' \) is a homomorphism with kernel \( I \) and image \( h[R] \); \( C \) is the collection of ideals \( J \subseteq R \) with \( I \subseteq J \); and \( C' \) is the collection of ideals \( J' \subseteq R' \) with \( h^{-1}[J] \subseteq C \); this follows using theorem 4.11 and \( h(\gamma)h(x) = h(\gamma x) \). For \( J \in C \), \( h[I] \) (i.e., the cosets \( \{x + I : x \in J\} \)) is mapped to \( h[J] \) by the canonical isomorphism. \( R/J \) is isomorphic to \( h[R]/h[I] \). If \( I \) and \( J \) are ideals, then \( I \cap J = I + J \). The ideals form a monoid under +, with the trivial ideal being the identity; the map \( I \mapsto h[I] \) is a monoid homomorphism.

Let \( \mathbb{Z}_n \) denote the quotient of \( \mathbb{Z} \) by the ideal of integers divisible by \( n \); \( \mathbb{Z}_n \) is called the ring of integers mod \( n \) (the set of integers divisible by \( n \) is commonly written as \( n\mathbb{Z} \), and \( \mathbb{Z}/n\mathbb{Z} \) is often used instead of \( \mathbb{Z}_n \)). Two integers in the same coset iff their difference is divisible by \( n \); the notation \( x \equiv y \mod n \) (congruence mod \( n \)) is commonly used for this. \( \mathbb{Z}_n \) contains \( n \) elements, the cosets of the integers from 0 to \( n - 1 \); these are called the congruence classes, mod \( n \).

There is a unique homomorphism \( \phi \) from \( \mathbb{Z} \) into any ring \( R \), where the image of \( n \) is \( 1 + \cdots + 1 \). We can write \( n \) for this element of \( R \), but the \( n \) need not not all be distinct. The image of \( \phi \) is called the prime subring. If \( \phi \) is injective the prime subring is a copy of \( \mathbb{Z} \); \( R \) is said to have characteristic 0. If \( \phi(n) = 0 \) for some nonzero \( n \), it is readily verified that there is a least positive \( n \) such that \( \phi(n) = 0 \), and that the kernel of \( \phi \) is the set of integers divisible by \( n \). The prime subring is a copy of \( \mathbb{Z}_n \) for some \( n \); \( R \) is said to have finite characteristic, and \( n \) is called the characteristic. The prime subring is clearly the smallest subring of a ring \( R \).

The product \( \times_i R_i \) of rings \( R_i \) is their product as structures; it is readily verified to be a ring (the axioms are equations). The canonical embedding of \( R_i \) as an additive group is not a ring homomorphism, since it does not preserve 1, and its image is not in general a subring. It is an ideal, and the quotient ring is isomorphic to the product of the remaining factors. The direct sum of the \( R_i \) is the subring of \( \times_i R_i \) where the sequences are 0 at all but a finite number of places; the notation \( \oplus R_i \) is used for this.

3. Commutative rings. A ring is called commutative if the axiom \( xy = yx \) holds. The center of a ring \( R \) is defined to be \( \{x \in R : xy = yx \text{ for all } y \in R\} \); this is readily seen to be a commutative subring. A ring is called a division ring or skew field if \( 0 \neq 1 \) and every nonzero element possesses a multiplicative inverse. A commutative division ring is called a field.
For the remainder of this chapter only commutative rings are considered. An additive subgroup which is closed under left multiplication is an ideal; and \( Rx_1 + \cdots + Rx_k \) is the smallest ideal containing \( x_1, \ldots, x_k \).

In a commutative ring \( R \), an ideal \( I \subseteq R \) is

- maximal if it is proper and whenever \( I \subseteq J \subseteq R \) for an ideal \( J \) then \( J = I \) or \( J = R \);
- prime if it is proper and whenever \( xy \in I \), \( x, y \in R \), then either \( x \in I \) or \( y \in I \);
- finitely generated if \( I = Rx_1 + \cdots + Rx_k \) for some \( x_1, \ldots, x_k \in I \).
- principal if \( I = Rx \) for some \( x \in I \).

Suppose \( I \) is a proper ideal. Let \( S \) be the set of proper ideals containing \( I \). \( S \) is partially ordered by inclusion. Clearly it is nonempty, and it is readily seen to be inductive (the union of a chain of ideals containing \( I \) is an ideal containing \( I \)). By Zorn’s lemma \( S \) contains a maximal element, which is clearly a maximal ideal of \( R \) containing \( I \). In particular, if \( R \) is nontrivial it has maximal ideals.

The only ideals in a field \( F \) are \( \{0\} \) and \( F \), since if the ideal contains a nonzero element it contains 1. (It follows from theorem 3.b below that a commutative ring with this property is a field.) Given a ring homomorphism \( \sigma : F \to R \) where the domain \( F \) is a field, the kernel is either \( \{0\} \), whence \( \sigma \) is injective; or \( \sigma \) is the zero map, defined as \( \sigma(x) = 0 \) for all \( x \in F \).

The following terms apply in a commutative ring \( R \); let \( w, x, y, u \) denote elements of \( R \).

- Say that \( x \mid y \) (\( x \) divides, or is a divisor of, \( y \)) if \( y = wx \) for some \( w \). Exercise 3 gives the basic properties of this relation; in particular it is a preorder.
- An element \( w \) is a greatest common divisor (abbreviated gcd) of elements \( x, y \) if \( w \mid x, w \mid y \) and whenever \( u \mid x, u \mid y \) then \( u \mid w \). Elements \( x, y \) are said to be relatively prime if 1 is a gcd.
- An element \( w \) is a least common multiple (abbreviated lcm) of elements \( x, y \) if \( x \mid w, y \mid w \) and whenever \( x \mid u, y \mid u \) then \( w \mid u \).
- Elements \( x, y \) are called associates if \( y = ux \) where \( u \) is a unit. We use \( \sim \) in this chapter to denote the relation of being associates; it is a strong congruence relation with respect to \( | \), and if \( x \sim y \) then \( Rx = Ry \) (exercise 3).
- An element \( x \) is called a zero divisor if it is nonzero, and \( xy = 0 \) for some nonzero \( y \).

Zero divisors are in fact defined in noncommutative rings; \( x \neq 0 \) is a zero divisor if \( xy = 0 \) or \( yx = 0 \) for some \( y \neq 0 \). Note that gcd\((0,0) = 0 \); some authors do not define this.

A commutative ring is called an integral domain if it is nontrivial and contains no zero divisors; equivalently, if the multiplicative monoid satisfies the cancellation law, or if the trivial ideal is prime. In an integral domain \( R \),

- a nonzero element \( p \in R \) is called irreducible if it is not a unit and whenever \( p = xy \) then one of \( x, y \) is a unit;
- a factorization of a nonzero element \( x \in R \) is a product \( p_1 \cdots p_n \) equaling \( x \), where each \( p_i \) is irreducible;
- an element \( p \in R \) is called prime if \( p \neq 0 \) and whenever \( p \mid xy \) then \( p \mid x \) or \( p \mid y \).

Suppose \( R \) is an integral domain.

- If \( Rx = Ry \) then \( x \sim y \).
- \( \sim \) is the canonical congruence relation of the preorder \( | \). The map \( [x] \mapsto Rx \) is an isomorphism from \( | \sim \) to the collection of principal ideals ordered by superset. The gcd’s (lcm’s) of two elements are a class of associates, indeed the meet (join) of the classes of the elements.
- If \( R \) has finite characteristic \( n \) then \( n \) must be a prime.

These facts are left as exercises.

**Theorem 3.** Suppose \( R \) is a nontrivial commutative ring and \( I \subseteq R \) is an ideal.

a. \( R/I \) is an integral domain iff \( I \) is prime.

b. \( R/I \) is a field iff \( I \) is maximal.

33
c. If $I$ is maximal it is prime.

**Proof:** Suppose $I$ is prime; if $xy \equiv_I 0$ then $xy \in I$, so $x \in I$ or $y \in I$, that is, $x \equiv_I 0$ or $y \equiv_I 0$. This shows that $R/I$ is an integral domain. Suppose $R/I$ is an integral domain; if $xy \in I$ then $xy \equiv_I 0$, so $x \equiv_I 0$ or $y \equiv_I 0$, that is, $x \in I$ or $y \in I$. This shows that $I$ is prime. Suppose $I$ is maximal; if $x \not\equiv_I 0$ then $x \notin I$, so the ideal generated by $I$ and $x$ is $R$, whence $1 = x'x + u$ where $x' \in R$, $u \in I$, whence $x'x + I = 1 + I$. This shows that $R/I$ is a field. Suppose $R/I$ is a field; if $x \notin I$ then $x \not\equiv_I 0$, so for some $x' \in R$, $x'x + I = 1 + I$, whence $1 = x'x + u$ where $u \in I$, whence the ideal generated by $I$ and $x$ is $R$. This shows that $I$ is maximal.

Part c follows by parts a and b, since a field is an integral domain. For a direct argument, if $I$ is maximal, $xy \in I$, and $x \notin I$ then for some $x' \in R$, $u \in I$, $x'x = 1 + u$; also $x'xy \in I$, so $y = x'xy - uy \in I$.

A commutative ring $R$ is called Noetherian if its lattice of ideals satisfies the ascending chain condition.

**Theorem 4.**

a. A commutative ring $R$ is Noetherian iff every ideal of $R$ is finitely generated.

b. In a Noetherian integral domain every nonzero element has a factorization.

**Proof:** If $I_0 \subset I_1 \subset \cdots$ is an infinite ascending chain its union $I$ is an ideal. If $x_1, \ldots, x_k \in I$ then for some $j$, $x_1, \ldots, x_k \in I_j$, whence $Rx_1 + \cdots + Rx_k \subseteq I_j$. Thus, $I$ is not finitely generated. If $I \subseteq R$ is not finitely generated define a sequence $x_0, x_1, \ldots$ so that the ideals $Rx_0, Rx_0 + Rx_1, \ldots$ are an ascending chain as follows. Let $x_0$ be arbitrary; having chosen $x_j$ for $j < i$ let $x_i$ be any element of $I$ not in $Rx_0 + \cdots + Rx_{i-1}$. For part b, suppose $R$ is an integral domain, and suppose $x_0 \neq 0$ has no factorization. It cannot be irreducible, so $x_0 = yz$ where neither $x_0 \sim y$ nor $x_0 \sim z$, and so $Rx_0 \subset Ry, Rz$. If both $y, z$ had factorizations then $x_0$ would, so one of them does not; call this $x_1$. Continuing, we obtain an ascending chain $Rx_0 \subset Rx_1 \subset \cdots$.

An integral domain is called
- a factorial domain if every nonzero $x \in R$ has a unique factorization, that is, has a factorization and given two such, say $x = u_1 \cdots u_k = v_1 \cdots v_l$, then $l = k$ and the $v_i$ can be reordered so that $v_i \sim u_i$;
- a principal ideal domain if every ideal is principal;
- a Euclidean domain if a Euclidean norm can be defined on it, where a Euclidean norm is a function $\nu: R - \{0\} \rightarrow \mathbb{N}$ such that given $x, y \in R - \{0\}$, there are $q, r \in R$ with $x = qy + r$ and either $r = 0$ or $\nu(r) < \nu(y)$; further $\nu(xy) \geq \nu(x)$.

If the reals are extended with $-\infty$, given a Euclidean norm $\nu$, $\nu(0)$ may be considered to be $-\infty$.

**Theorem 5.** Let $R$ be an integral domain.

a. If $R$ is a Euclidean domain it is a principal ideal domain.

b. If $R$ is a principal ideal domain it is Noetherian.

c. If $R$ is a principal ideal domain it is a factorial domain.

d. If $p \in R$ is nonzero then $p$ is prime iff $pR$ is a prime ideal.

e. If $p \in R$ is prime it is irreducible.

f. If $R$ is a factorial domain,

1. a gcd of $x, y$ exists, and $\lvert / \rvert \sim$ is a lattice;
2. if $w, x$ are relatively prime and $w \mid xy$ then $w \mid y$;
3. if $p$ is irreducible it is prime;
4. if $d$ is a gcd of $x, y$ then $xy/d$ is an lcm.

g. If $R$ is a principal ideal domain

1. a gcd of $x, y$ equals $ax + by$ for some $a, b \in R$;
2. a prime ideal is maximal.

h. If $R$ is finite it is a field.
 Proof: Suppose $R$ is a Euclidean domain, say with norm $\nu$, and let $I$ be an ideal. Let $d \in I, d \neq 0$, be such that $\nu(d)$ is least, among the members of $I$. Given $x \in I$ we have $x = qd + r$ where either $\nu(r) < \nu(d)$ or $r = 0$. By the assumption on $d$ and the fact that $r \in I$ we cannot have that $\nu(r) < \nu(d)$, so $r = 0$ and $x \in dR$. This shows that $I \subseteq dR$, and certainly $dR \subseteq I$; hence $I = dR$ and every ideal is principal. Part b is a trivial consequence of theorem 4. For part c, every $x \in R$ has some factorization by theorem 4. In a principal ideal domain, the map $[x] \mapsto Rx$ from $|/\sim$ to the complete lattice of ideals ordered by superset is an isomorphism. Hence if $Rx + Ry = Rz$ then $z$ is a gcd of $x, y$; that $z = ax + by$ for some $a, b \in R$ follows because $Rx + Ry$ is exactly such elements of $R$. If $w, x$ are relatively prime then $1 = aw + bx$ for some $a, b \in R$, so $y = awy + bxy$; if in addition $w|xy$ then $w|y$. If $p$ is irreducible then $p$ and $x$ are relatively prime unless $p|x$; hence if $p|xy$ then $p|x$ or $p|y$. Now, given factorizations $p_1 \cdots p_s, q_1 \cdots q_t$ of $x, q_1|p_1 \cdots p_s$, so $q_1$ must be an associate of some $p_i$. Continuing, the factorizations must be essentially the same. This completes the proof of part c. For part d, $p|x$ iff $x \in pR$, and similarly for $y$ and $xy$; the claim is thus immediate from the definition. For part e, if $p = xy$ where $x, y$ are nonunits then $p$ is certainly not prime; $xy|xy$, but if $xy|x$ then $y$ is a unit, and similarly for $y$. For part f, if $x = p_1^{e_1} \cdots p_s^{e_s}$ and $y = p_1^{f_1} \cdots p_t^{f_t}$ where the $p_i$ are distinct and $e_i, f_i$ may be zero, then $p_1^{e_1} \cdots p_s^{e_s} p_1^{f_1} \cdots p_t^{f_t}$ iff $e_i \leq f_i$ for all $i$. The claims follow easily from this. Part g.1 was proved above. For part g.2 it suffices to show that if $p$ is prime then $pR$ is maximal. This follows because the ideal lattice is isomorphic to $/$, and prime elements are minimal nonunits. For part h, if $x \neq 0$ the map $x \mapsto xy$ is injective, hence bijective, so $xy = 1$ for some $y$.

It is easily verified that $\mathbb{Z}_n$ is an integral domain iff $n$ is a prime $p$; by theorem 5.h, in this case it is a field.

There is a converse to theorem 5.f.3. Say that a commutative ring satisfies the divisor chain condition if there is no infinite sequence $a_i$ where $a_{i+1}$ is a proper divisor of $a_i$ for all $i \in \mathbb{N}$. Then an integral domain is factorial iff it satisfies the divisor chain condition, and every irreducible element is prime. The proof is left to the reader, noting that the divisor chain condition guarantees some factorization, as in theorem 4.b. Note also that the divisor chain condition holds in a Noetherian integral domain.

Theorem 5.g.1 readily generalizes to several elements. That is, in a principal ideal domain a gcd of $x_1, \ldots, x_k$ can be written as $a_1x_1 + \cdots + a_kx_k$. In a Euclidean ring Euclid’s algorithm provides a means for computing $d = \gcd(x_1, \ldots, x_k)$; this is as follows.

1. If $x_i \neq 0$, $x_j = 0$ for $j \neq i$, set $d = x_i$ and stop.
2. Find a nonzero $x_i$ of least norm.
3. Reduce the norm of some nonzero $x_j$, $j \neq i$, by writing $x_j$ as $qx_i + r$ and replacing $x_j$ by $r$.
4. Go back to step 1.

Step 3 does not change the set of common divisors, and the algorithm terminates because the least norm decreases. This is an algorithm only in a generalized sense, unless we can actually compute $r$ in step 3.

Values $a_i$ such that $\sum a_i x_i = d$ can be determined by writing the $x_i$ as a column vector $x$, and keeping track of a matrix $A$ so that $Ax^0 = x$ where $x^0$ is the initial value; this is sometimes called the extended Euclidean algorithm. We leave the details as an exercise, and also the fact that in a factorial domain,

$$\gcd(\gcd(x_1, \ldots, x_{k-1}), x_k) = \gcd(x_1, \ldots, x_k),$$

and an algorithm for finding the $a_i$ which makes use of this fact.

4. Fractions. There are two important ways of obtaining new rings from a commutative ring $R$, namely the ring of fractions and the ring of polynomials. The ring of polynomials is considered in the next chapter.

The ring of fractions is a generalization of the construction of the rational numbers from the integers, where each rational number is considered as an equivalence class of fractions $m/n$, $n \neq 0$, where $m_1/n_1 \equiv m_2/n_2$ if $m_1n_2 = m_2n_1$. This construction goes through virtually unchanged in any integral domain. It may be
generalized to any commutative ring; also, the denominators may be required to be in an appropriate subset of the ring.

Define a multiplicative subset of $R$ to be a submonoid of $R^\times$, i.e., a subset $S \subseteq R$ with $0 \notin S$, $1 \in S$, and $xy \in S$ whenever $x,y \in S$. For example if $P$ is a prime ideal then $R - P$ is a multiplicative subset, immediately from the definitions. $R^\times$ itself is a multiplicative subset iff $R$ is an integral domain, iff $\{0\}$ is a prime ideal. $S$ need not be the complement of an ideal, but if it is the ideal is prime.

For the remainder of the section $x/y$ denotes an ordered pair $(x,y)$ from $R$, with $y \neq 0$; $x$ is called the numerator and $y$ the denominator. Given a multiplicative subset $S \subseteq R$, we define a relation $\equiv_S$ (or just $\equiv$) on the pairs with denominators in $S$, namely $x/s \equiv x'/s'$ iff $t(xs' - x's) = 0$ for some $t \in S$. If $R$ is an integral domain the requirement becomes $xs' = x's$.

**Lemma 6.** The relation $\equiv_S$ on the pairs is an equivalence relation. If $x_1/s_1 \equiv x'_1/s'_1$, $x_2/s_2 \equiv x'_2/s'_2$ then $(x_1s_2 + x_2s_1)/(s_1s_2) \equiv (x'_1s'_2 + x'_2s'_1)/(s'_1s'_2)$, and $(x_1x_2)/(s_1s_2) \equiv (x'_1x'_2)/(s'_1s'_2)$.

**Proof:** Reflexivity and symmetry of $\equiv$ are immediate. For transitivity, if $t_1(xs' - x's) = 0$, $t_2(x's'' - x''s') = 0$ then $t_1t_2s'(xs'' - x''s') = 0$. Suppose $t_1(x_1s'_1 - x'_1s_1) = 0$, $t_2(x_2s'_2 - x'_2s_2) = 0$. $t_1t_2((x_1s_2 + x_2s_1)s'_2 - (x'_1s'_2 + x'_2s'_1)s_2) = 0$ and $t_1t_2(x_1x_2s'_2 - x'_1x'_2s_2) = 0$.

If $S$ is a multiplicative subset of $R$ consider the equivalence classes under $\equiv$ of pairs with denominators in $S$. By the lemma the functions $[x_1/y_1] + [x_2/y_2] = [(x_1y_2 + x_2y_1)/(y_1y_2)]$ and $[x_1/y_1][x_2/y_2] = [(x_1x_2)/(y_1y_2)]$ are well defined. Let $R_S$ denote the equivalence classes with these operations. A more common notation is $S^{-1}R$; the notation $R_x$ is commonly used in the case that $S$ is the powers of $x$ (provided no $x^i = 0$), and $R_P$ when $P$ is a prime ideal and $S = R - P$. Other notations are also used, including $R[\{S^{-1}\}]$ and $R\{S\}$.

**Theorem 7.** Suppose $S$ is a multiplicative subset of a commutative ring $R$.

a. $R_S$ is a commutative ring, with $- [x/y] = [(-x)/y]$, additive identity $[0/1]$, and multiplicative identity $[1/1]$.

b. The map $\mu : x \mapsto [x/1]$ is a homomorphism, and $\mu(S) \subseteq \text{Units}(R_S)$.

c. If $\phi : R \to R'$ is a ring homomorphism such that $\phi(S) \subseteq \text{Units}(R')$ then there is a unique $\phi' : R_S \to R'$ such that $\phi' = \mu \phi$ (namely $\phi'([x/s]) = \phi(x)\phi(s)^{-1}$).

d. The diagram property of part c characterizes $R_S$ up to isomorphism.

e. If $R$ is an integral domain then $R_S$ is an integral domain and $\mu$ is injective.

**Proof:** Part a is left as an exercise. We have $x/1 + y/1 \equiv (x+y)/1$ and $(x/1)(y/1) \equiv (xy)/1$, so $\mu$ is a homomorphism. Since $(1/s)(s/1) = 1/1 \mu(s)$ is a unit for $s \in S$. If $\phi'$ in part c exists it must clearly be as specified, so we need only show that this map is well-defined and a homomorphism. If $x/s \equiv x'/s'$ then $t(xs' - x's) = 0$ for some $t \in S$. Applying $\phi$ and multiplying by $\phi(t)^{-1}\phi(s)^{-1}\phi(s')^{-1}$ yields $\phi(x)\phi(s)^{-1} = \phi(x')\phi(s')^{-1}$. That $\phi'$ is a homomorphism is straightforward and left as an exercise. Part d follows from part c in the usual manner and is left as an exercise. If $R$ is an integral domain and $[x/s][x'/s'] = 0$ then $xx' = 0$, so $x = 0$ or $x' = 0$; thus, $R_S$ is an integral domain. For any $R$, $x \in R$ is in the kernel of $\mu$ iff $tx = 0$ for some $t \in S$; if $R$ is an integral domain then $x$ must equal 0, so $\mu$ is injective. This proves part e.

If $S \subseteq T$ where $S,T$ are multiplicative subsets of $R$, there is a canonical homomorphism from $R_S$ to $R_T$; namely, $[x/y]$ in $R_S$ maps to $[x/y]$ in $R_T$. This is readily verified to be well defined. When $R$ is an integral domain the map is injective.

When $R$ is an integral domain $R_{R^\times}$ is a field, called the field of fractions of $R$. The canonical embedding $x \mapsto [x/1]$ can be ignored, and the field of fractions considered the smallest field containing $R$. For any multiplicative subset $S$, $R_S$ may be viewed as a subring of the field of fractions; it is those classes which contain a fraction with a member of $S$ in the denominator.
For a concrete example, \( \mathbb{Q} \) is the field of fractions of \( \mathbb{Z} \). If \( p \) is a prime number \( p\mathbb{Z} \), the set of integers divisible by \( p \), is a prime ideal. The \( S \) set of integers not divisible by \( p \) is a multiplicative subset of \( \mathbb{Z} \): \( \mathbb{Z}_S \) consists of all those fractions in lowest terms, whose denominator is not divisible by \( p \). The units are those where the numerator is not divisible by \( p \) either. Given any fraction \( m/n \), define its order at \( p \) to be that integer \( r \) such that \( m/n = p^r(u/v) \) where \( u, v \) are not divisible by \( p \); \( r \) may be positive, negative, or 0. The elements of the ring are those fractions whose order is nonnegative, and the units those whose order is 0. These remarks apply in any principal ideal domain.

5. **Ordered commutative rings.** An ordered commutative group is a commutative group, together with a one place relation “positive”, which satisfies the following axioms.

- 0 is not positive.
- If \( x \neq 0 \) then either \( x \) or \( -x \) is positive, but not both.
- If \( x \) and \( y \) are positive so is \( x + y \).

The two argument predicate \( x < y \) is defined to hold if \( y - x \) is positive; and the two argument predicate \( x \leq y \) to hold if \( x < y \) or \( x = y \). It is readily verified that the following hold in an ordered commutative group.

- \( \leq \) is a linear order;
- if \( x < y \) then \( x + z < y + z \), and if \( x \leq y \) then \( x + z \leq y + z \);
- if \( x < y \) then \( -y < -x \), and if \( x \leq y \) then \( -y \leq -x \);

The absolute value \( |x| \) is defined to be \( x \) if \( x \) is positive or 0, else \(-x\). This may be verified to satisfy the triangle inequality \( |x + y| \leq |x| + |y| \).

Ordered commutative groups illustrate a fine point of chapter 2. In this case the kernel relation of a homomorphism is not necessarily a congruence relation. Taking the quotient disregarding the order, the resulting group may not be ordered, as the example of \( \mathbb{Z}_n \) shows. Note also that the product of ordered commutative groups need not be ordered.

An ordered commutative ring is a commutative ring, together with a one place relation “positive”, which makes + an ordered commutative group, and in addition satisfies

- if \( x \) and \( y \) are positive so is \( xy \).

One readily verifies that

- 1 is positive (provided \( 1 \neq 0 \));
- if \( x < y \) and \( 0 < z \) then \( xz < yz \), and if \( x \leq y \) and \( 0 \leq z \) then \( xz \leq yz \);
- if \( xy = 0 \) then \( x = 0 \) or \( y = 0 \); and
- \( |xy| = |x||y| \).

In particular a nontrivial ordered commutative ring is an integral domain; also it clearly has characteristic 0.

If \( R \) is an ordered integral domain then an order may be defined on \( R_S \), for any multiplicative subset \( S \). Namely, a class is positive if it contains a fraction \( a/b \) where \( a \) and \( b \) are positive. It is readily verified that this satisfies the required axioms.

6. **Chinese remainder theorem.** In a commutative ring \( R \), for \( S \subseteq R \) let \([S]\) denote the ideal generated by \( S \). For ideals \( I, J \subseteq R \) \([IJ]\) is called the product of the ideals \( I, J \). The product is often denoted \( IJ \), but to avoid confusion we generally reserve this for the setwise product. Clearly \([IJ]\) is those elements of the form \( x_1y_1 + \cdots + x_ky_k, x_i \in I, y_i \in J \). The following facts are readily verified.

- Ideal multiplication is associative, and \([IK] = [[IJ]K], \text{etc.}\n- In fact the ideals form a commutative monoid under ideal multiplication; the identity is \( R \).
- \([I_1 \cdots I_k] \subseteq I_1 \cap \cdots \cap I_k \).
- The map \( a \mapsto aR \) is a monoid monomorphism from the multiplicative monoid of \( R \) to that of the ideals.
- An ideal $P \subseteq R$ is prime iff whenever $[IJ] \subseteq P$, either $I \subseteq P$ or $I \subseteq P$.

The term “ideal” arose due to the monomorphism $a \mapsto aR$; the ideals behave enough like elements in some contexts to consider them as “idealized” elements.

**Theorem 8.** Suppose $R$ is a commutative ring: $I_j$, $1 \leq j \leq k$ are ideals; and $h_j : R \rightarrow R/I_j$ is the canonical homomorphism. Let $h : R \rightarrow R/I_1 \times \cdots \times R/I_k$ be defined by $h(x) = \langle h_1(x), \ldots, h_k(x) \rangle$. Then $h$ is a homomorphism with kernel $I_1 \cap \cdots \cap I_k$. If $I_1 + I_2 = R$ for $i \neq j$ then $h$ is surjective, and $[I_1 \cdots I_k] = I_1 \cap \cdots \cap I_k$.

**Proof:** The first claim is straightforward. For the second, let $J_i$ be $[S_i]$ where $S_i$ is the setwise product of the $I_j$, $j \neq i$. By hypothesis, for $i \neq j$ there are $u_{ij} \in I_i$, $v_{ij} \in I_j$ with $1 = u_{ij} + v_{ij}$. It follows that for a particular $i$ there are $u_i \in I_i$ and $v_i \in I_i$, with $1 = u_i + v_i$; this may be seen by rewriting $\prod_{j \neq i}(u_{ij} + v_{ij})$.

It follows that $h(v_i)$ is 1 in the $i$th component and 0 in the remaining components; it now follows that $h$ is surjective. For the last claim, suppose $1 = u + v$ where $u \in I$ and $v \in J$, and suppose $x \in I \cap J$; then $x = xu + xv \in [IJ]$. This proves the claim for $k = 2$, and may be used in an inductive argument completing the proof.

The term comaximal is used for ideals $I, J$ such that $I + J = R$. If $R$ is a principal ideal domain the restriction $aR + bR = R$ is equivalent to $\gcd(a, b) = 1$; in this case $[aRbR] = abR$. In the case $R = \mathbb{Z}$ the theorem implies that if $\gcd(m_i, m_j) = 1$, $i \neq j$, then for any $a_1, \ldots, a_k$ the system of congruences $x \equiv a_i \bmod m_i$ has exactly one solution with $0 \leq x < m_1 \cdots m_k$. This fact is called the Chinese remainder theorem, and was known to the first century Chinese mathematician Sun-Tsu. Note that the elements $v_i$ of the proof satisfy $v_i \equiv 1 \bmod m_i$, and $v_i \equiv 0 \bmod m_j$ for $j \neq i$.

Suppose $m = p_1^{e_1} \cdots p_k^{e_k}$ is the prime factorization of $m$. The theorem implies that $\mathbb{Z}_m$ is isomorphic to $\mathbb{Z}_{p_1^{e_1}} \times \cdots \times \mathbb{Z}_{p_k^{e_k}}$, via the map which sends $n$ to $\langle n_1, \ldots, n_r \rangle$ where $n_i = n \bmod p_i^{e_i}$; this may be easily proved directly. The following lemma is a useful fact about direct products of rings; it implies, for example, that $\text{Units}(\mathbb{Z}_m)$ is isomorphic to $\text{Units}(\mathbb{Z}_{p_1^{e_1}}) \times \cdots \times \text{Units}(\mathbb{Z}_{p_k^{e_k}})$.

**Lemma 9.** $\text{Units}(R_1 \times \cdots \times R_k)$ is isomorphic to $\text{Units}(R_1) \times \cdots \times \text{Units}(R_k)$.

**Proof:** Clearly, $\langle a_1, \ldots, a_n \rangle$ is a unit iff each $a_i$ is. The isomorphism simply views $\text{Units}(R_1 \times \cdots \times R_k)$ as a substructure of $R_1 \times \cdots \times R_k$.

**7. M"{o}bius inversion.** M"{o}bius inversion is a useful tool, which is most generally given in the context of a locally finite partially ordered set. A partially ordered set $(S, \leq)$ is said to be locally finite if given $u, v \in S$, $u \leq v$, the set $\{x : u \leq x \leq v\}$ is finite. Let $F$ be a field, and let $A$ be the family of functions $f : S \times S \rightarrow F$ such that $f(x, y) = 0$ unless $x \leq y$.

A multiplication operation may be introduced on $A$, namely,

$$(f * g)(x, y) = \sum_{x \leq z \leq y} f(x, z)g(z, y), \quad x \leq y.$$ 

It is readily verified that this operation, together with pointwise addition, make $A$ a ring (in fact an $F$-algebra; see chapter 8). The multiplicative identity is the function $\delta(x, y)$ which is 1 if $x = y$ else 0. The function $f$ has a multiplicative inverse iff $f(x, x) \neq 0$ for all $x$; this may be defined by induction on the length of the longest chain from $x$ to $y$. We leave the verification of these claims as an exercise.

Let $\chi(x, y)$ be 1 if $x \leq y$ else 0; the inverse $\mu$ of $\chi$ is called the M"{o}bius function of the partial order. It is the unique function satisfying $\mu(x, y) = 0$ unless $x \leq y$; $\mu(x, x) = 1$ for any $x \in S$; and $\sum_{x \leq z \leq y} \mu(x, z) = 0$ for $x < y$. 

38
Lemma 10 (Möbius inversion). For $f, g \in A, x, y \in S, x \leq y,$

$$g(x, y) = \sum_{w : x \leq w \leq y} f(w, y) \text{ iff } f(x, y) = \sum_{w : x \leq w \leq y} \mu(x, w)g(w, y); \text{ and}$$

$$g(x, y) = \sum_{w : x \leq w \leq y} f(x, w) \text{ iff } f(x, y) = \sum_{w : x \leq w \leq y} \mu(w, y)g(x, w).$$

Proof: The first equation is equivalent to $g = \chi * f$, the second to $f = \mu * g$, the third to $g = f * \chi$, and the fourth to $f = g * \mu$. The lemma follows by $\mu = \chi^{-1}$ and algebra.

Outside of partial order theory, Möbius inversion is mostly of interest for the partially ordered set $(\mathcal{N}^>, |)$, with $F = \mathcal{C}$. The functions $f : S \mapsto F$ may be viewed as members of $A$ by defining $\tilde{f}(x, y) = f(x/y)$ when $x|y$; indeed, they may be identified with the subalgebra of $A$ consisting of the $\tilde{f}$ which satisfy $\tilde{f}(x, y) = \tilde{f}(1, y/x)$; call the algebra of single argument functions $B$.

Multiplication in $B$ becomes $(f * g)(x) = \sum_{d|e} f(d)g(x/d)$, or $\sum_{de=x} f(d)g(e)$. This is called the Dirichlet product, and is clearly commutative. The multiplicative identity $\delta(x)$ is 1 if $x = 1$, else 0. The functions with a multiplicative inverse are those with $f(1) \neq 0$. The $\mu$ function is the inverse of the function $\chi(x) = 1$ for all $x$. The Möbius inversion theorem states that $\tilde{f}(x) = \sum_{d|x} f(d)$ for all $x$ iff $f(y) = \sum_{d|y} \mu(d)g(y/d)$ for all $y$. To give $\mu$ explicitly, define an integer $n \in \mathcal{N}^>$ to be square-free if in its factorization each prime occurs to the first power.

Lemma 11. If $x \in \mathcal{N}^>$ is square-free, $\mu(x) = (-1)^k$ where $k$ is the number of prime divisors; otherwise $\mu(x) = 0$.

Proof: If $x$ is not square-free we have by induction

$$\mu(p_1^{e_1} \cdots p_k^{e_k}) = -\sum_{S \subseteq \{1, \ldots, k\}} (-1)^{|S|} = -\sum_{0 \leq j \leq k} \binom{k}{j}(-1)^j = -(1 - 1)^k = 0.$$

If it is, the term $(-1)^k$ is omitted from the sum, and the claim follows in this case also.

The Euler function is defined by $\phi(n) = |\{m : \gcd(m, n) = 1, 1 \leq m \leq n\}|$; equivalently, $\phi(n)$ is the cardinality of the group of units in the ring $\mathbb{Z}_n$. A function $f : \mathcal{N}^> \mapsto \mathcal{C}$ is called multiplicative if it is not identically 0 and $f(mn) = f(m)f(n)$ whenever $\gcd(m, n) = 1$. This is clearly so iff $f(n) = f(p_1^{e_1}) \cdots f(p_k^{e_k})$ where $n = p_1^{e_1} \cdots p_k^{e_k}$ is the factorization of $n$.

Lemma 12.

a. $\phi$ is multiplicative;

b. $\phi(n) = n \prod_{p|n, p \text{prime}} \left(1 - \frac{1}{p}\right)$;

c. $\phi(n) = \sum_{d|n} \mu(d)(n/d)$;

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

Proof: Part a follows by lemma 9 (see the comments preceding it). Now, there are $p^{e-1}$ integers $i$ satisfying $0 \leq i < p^{e-1}$ of these are divisible by $p$, namely those of the form $pj, 0 \leq j < p^{e-1}$; part b follows. Part c follows because the product in part b is equal to

$$\sum_{S \subseteq \{1, \ldots, k\}} (-1)^{|S|} \frac{1}{d_S}$$

where $d_S$ is the product of the primes $p_i, i \in S$ ($d_{\emptyset} = 1$). Part d follows from part c by Möbius inversion.
Lemma 12.c can be proved more directly as follows. Let \( N \) be the function where \( N(n) = n \) for all \( n \); the claim is that \( \mu * N = \phi \). Indeed,

\[
(\mu * N)(n) = \sum_{d|n} \mu(d) \sum_{e=1}^{n/d} 1 = \sum_{e=1}^{n} \sum_{d|e} \mu(d);
\]

the inner sum in the last term equals \( \sum_{d|\gcd(e,n)} \mu(d) \), which equals 1 if \( \gcd(e, n) = 1 \) and 0 otherwise, proving the claim. That \( \phi \) is multiplicative now follows from the claim that the Dirichlet product of multiplicative functions is multiplicative, since \( \mu \) and \( N \) are clearly multiplicative. To see this, note that

\[
\sum_{d|m} f(d)g\left(\frac{mn}{d}\right) = \sum_{d_1|d_2|\frac{m}{d}} f(d_1)f(d_2)g\left(\frac{m}{d_1}\right)g\left(\frac{d_1}{d_2}\right) = \left(\sum_{d_1|m} f(d_1)g\left(\frac{m}{d_1}\right)\right)\left(\sum_{d_2|n} f(d_2)g\left(\frac{n}{d_2}\right)\right).
\]

Yet another proof of lemma 12.b makes use of the principle of inclusion and exclusion, which states that if \( C \) is a collection of \( k \) events then

\[
\Pr\left[\bigcup_{i=1}^{k} C_i\right] = \sum_{D \subseteq C} (-1)^{|D|-1} \Pr[\bigcap_{D \subseteq C}] ;
\]

a proof is given below. The events are \( p_i|m, 1 \leq m \leq n \), where \( p_i \) is a prime divisor of \( n \), and we want 1 minus the probability of the union; this yields (using the notation \( d_S \) from above)

\[
\phi(n) = \sum_{S \subseteq \{1, \ldots, k\}} (-1)^{|S|} \frac{n}{d_S}.
\]

Lemma 12.b and 12.c follow; note that lemma 12.a follows from lemma 12.b.

Suppose \( G \) is a cyclic group of order \( n \), with generator \( x \). The order \( o(x^i) \) of \( x^i \) equals \( n/\gcd(i,n) \); for \( o(x^i) \) is the smallest \( d \) such that \( n|di \), and (writing \( g \) for \( \gcd(i,n) \)) \( n|di \) iff \( n|g|\gcd(i,n) \) iff \( n/g|d \). Thus, when \( d|n \), \( o(x^i) = d \) iff \( \gcd(i,n) = n/d \), which is so iff \( \gcd(i/(n/d),d) = 1 \). There are thus \( \phi(d) \) elements of order \( d \) in \( G \), namely \( x^{(n/d)j} \) where \( j \) runs over a reduced system of residues mod \( d \). For example if \( n = 12 \),

- \( o(x^i) = 12 \) if \( i = 1, 5, 7, 11 \); \( o(x^i) = 6 \) if \( i = 2, 10 \); \( o(x^i) = 4 \) if \( i = 3, 9 \);
- \( o(x^i) = 3 \) if \( i = 4, 8 \); \( o(x^i) = 2 \) if \( i = 6 \); \( o(x^i) = 1 \) if \( i = 0 \).

This provides yet another proof of lemma 12.d.

Two elements of order \( d \) in \( G \) generate the same cyclic subgroup. For if \( i_1 = (n/d)j_1 \) and \( i_2 = (n/d)j_2 \), where \( \gcd(j_1, d) = \gcd(j_2, d) = 1 \), then there is a \( k \) with \( j_1k \equiv j_2 \mod d \), whence \( i_1k \equiv i_2 \mod n \). We claim that this is all of the subgroups, and so the subgroups of \( G \) consist of exactly one subgroup of order \( d \) for each \( d|n \). Indeed, let \( H \) be a subgroup, and let \( i \) be least such that \( x^i \in H \). Then \( x^i \) generates \( H \), for if \( x^{ri+s} \in H \) where \( 0 < s < i \) then \( x^s \in i \), a contradiction.

A proof of the principle of inclusion and exclusion may be given by induction on \( n \), the basis \( n = 1 \) being immediate. Assume the theorem for \( n \); and let \( C = \{S_1, \ldots, S_n, S_{n+1}\} \). Now,

\[
P(S_1 \cup \cdots \cup S_n \cup S_{n+1}) = P(S_1 \cup \cdots \cup S_n) + P(S_{n+1}) - P((S_1 \cup \cdots \cup S_n) \cap S_{n+1});
\]

write \( P = P_1 + P_2 - P_3 \) for this. For each subcollection \( D \) of \( C \) not containing \( S_{n+1} \), \( P(\bigcap D) \) appears in the inductively given sum for \( P_1 \), with the same sign as it must have in the sum for \( P \). \( P_3 \) equals \( P(S_1 \cap S_{n+1} \cup \cdots \cup S_n \cap S_{n+1}) \); for each subcollection \( D \) of \( C \) containing \( S_{n+1} \) and with \( \text{card}(D) \geq 2 \), \( P(\bigcap D) \) appears in the inductively given sum for \( P_3 \), but with the opposite sign, because \( S_{n+1} \) is not counted in the size of the subcollection. Finally \( P(S_{n+1}) \) appears with the correct sign as \( P_2 \).
Exercises.

1. Prove theorem 1.

2. If \( h : S \mapsto T \) is a function between sets, define its kernel to be the equivalence relation \( h(x) = h(y) \). If \( S \) and \( T \) are structures of the same type, show that if this is a congruence relation then \( h \) is a homomorphism.

3. Show that the divisibility relation in a commutative ring \( R \) possesses the following properties.
   - \( x|y \) if \( y \) and \( z \) then \( x|z \);
   - \( x|0 \);
   - \( 1|x \);
   - \( x|y \) and \( x|z \) then \( x|y + z \);
   - \( x|y \) and \( w \) is any element then \( x|wy \);
   - \( x|y \) iff \( y \in Rx \) iff \( Ry \subseteq Rx \).
   
   Show also that
   - \( \sim \) is a strong congruence relation with respect to \( | \);
   - if \( x \sim y \) then \( Rx = Ry \).

4. Show that if \( R \) is an integral domain and \( x|y \) and \( y|x \) then \( x \sim y \).

5. Show that the requirement \( \nu(xy) \geq \nu(x) \) in the definition of a Euclidean norm is inessential, in that if a ring possesses a norm with only the first property it possesses one with the second also. Hint: Define \( \nu'(x) \) to be the least \( \nu(xy) \), \( y \neq 0 \).

6. Show that in a factorial domain the lattice \( / \sim \) is distributive. Hint: First show that \( (N, \leq) \) is.

7. Fill in the remaining details of the proof of theorem 7.

8. Prove the following version of the Möbius inversion theorem. Let \( A \) be a commutative group, written additively; let \( nx \) denote \( -x, 0, +x \) for \( n = -1, 0, +1 \) respectively. Suppose \( f, g : N \mapsto A \); then

   \[ g(n) = \sum_{d|n} f(d) \iff f(n) = \sum_{d|n} \mu(d) g(n/d). \]

9. Prove that in a factorial domain, \( \gcd(\gcd(x_1, \ldots, x_{k-1}), x_k) = \gcd(x_1, \ldots, x_k) \).

10. Let \( x^s \) denote the column vector of the values \( x_i \) after \( s \) steps of Euclid’s algorithm. Show that \( x^{s+1} = E_{s+1} x^s \) where \( E_{s+1} \) is the matrix of an elementary row operation (see appendix 2). Show that the product \( E_s \cdots E_1 \) can be kept track of by performing the same row operation on \( E_s \) as on \( x_s \), where \( E_0 = I \). This gives a method for computing the \( a_i \) in the extended Euclid’s algorithm.

11. In the extended Euclid’s algorithm for 2 elements \( x_1, x_2 \), keep track of only the first column of \( E_n \cdots E_1 \). Show how to determine \( a_1 \) and \( a_2 \).

12. Let \( d_1 = x_1, d_i = \gcd(d_{i-1}, x_i) \) for \( 1 < i \leq k \), and \( b_{i-1}d_{i-1} + c_ia_i = d_i \); say how to find the \( a_i \) from the \( b_i, c_i \).
7. Polynomials.

1. Single variable polynomials. Throughout this chapter, a ring \( R \) is commutative. A polynomial \( p \) over \( R \) is a formal object, namely an expression involving a single variable \( x \), of the form

\[
\sum_{i=0}^{n} p_i x^i, \quad \text{or} \quad p_0 + p_1 x + \cdots + p_n x^n
\]

where \( p_i \in R; p_i \) is called the \( i \)th coefficient. The coefficient \( p_n \) is required to be nonzero, unless \( n = 0 \) (so that the “zero polynomial” is the expression “0”); this value is called the “leading” coefficient of \( p \). If \( p \) is not the zero polynomial \( n \) is called its degree. In some contexts, the degree of the zero polynomial is considered to be \(-\infty\). We write \( \deg(p) \) for the degree of the polynomial \( p \).

An alternative way of viewing the polynomials which is sometimes convenient is to first consider the infinite expressions of the form \( \sum_{i=0}^{\infty} p_i x^i \); these are called the formal power series over \( R \). A polynomial may be seen to be exactly a formal power series where only finitely many coefficients are nonzero. Given formal power series \( p \) and \( q \), define \( p + q \) to be the formal power series \( r \) where \( r_i = p_i + q_i \), and \( pq \) to be the formal power series \( r \) where \( r_i = p_0 q_i + \cdots + p_i q_0 \). Addition and multiplication of polynomials are thus easily defined as the restriction of these operations to the polynomials, without the need of “extending” the polynomials with 0 coefficients.

Let \( R[x] \) denote the collection of polynomials over \( R \). A polynomial \( p \) is called a constant polynomial if \( p_i = 0 \) for \( i > 0 \). There is a map from \( R \) to \( R[x] \), which takes \( r \in R \) to the constant polynomial where \( p_0 = r \). This is a canonical embedding of \( R \) in \( R[x] \), by which \( R \) may be considered a subset.

**Theorem 1.** The formal power series, with the operations of \( +, \times \) given above, form a commutative ring. \( R[x] \) is a subring, and \( R \) is in turn a subring of \( R[x] \). The additive identity is 0; the additive inverse \( -p \) has \((-p)_i = -p_i \); and the multiplicative identity is 1. If \( r \in R \) then \((rp)_i = rp_i \).

**Proof:** The method of proof is to show that the \( i \)th coefficient is identical, on either side of the equation, using properties of \( R \). Thus, commutativity of \( + \) follows since \( a_i + b_i = b_i + a_i \) by commutativity of \( + \) in \( R \). Associativity of \( + \), that 0 is the additive identity, and that \( \sum_{i=0}^{n} (-a_i)x^i \) is the additive inverse of \( \sum_{i=0}^{n} a_i x^i \) follow similarly. Commutativity of \( \times \) follows by \( \sum_{j=0}^{i} a_0 b_{i-j} = \sum_{j=0}^{i} a_j b_{i-j} \); associativity of \( \times \) follows by

\[
\sum_{k=0}^{i} a_k \left( \sum_{j=0}^{i-k} b_j c_{k-j} \right) = \sum_{k=0}^{i} \left( \sum_{j=0}^{k} a_j b_{k-j} \right) c_{i-k};
\]

and the distributive law follows by

\[
\sum_{j=0}^{i} a_j (b_{i-j} + c_{i-j}) = \sum_{j=0}^{i} a_j b_{i-j} + \sum_{j=0}^{i} a_j c_{i-j}.
\]

These are all identities in \( R \). It is easily seen that multiplication by \( r \) multiplies the coefficients by \( r \), in particular multiplication by 1 leaves them unchanged; and addition or multiplication of two elements of \( R \), considered as formal power series, is the same as the operation in \( R \). Clearly the sum or product of polynomials is a polynomial.

If \( p \) is a polynomial of degree \( n \) we associate with it the function \( p(x) = p_0 + p_1 x + \cdots + p_n x^n \). \( R^R \), the collection of functions from \( R \) to \( R \), is (as noted earlier) a commutative ring when equipped with the “pointwise” operations \((f + g)(x) = f(x) + g(x)\) and \((fg)(x) = f(x)g(x)\). The map taking a polynomial \( p \) to the function \( p(x) \) is a ring homomorphism, since \((p + q)(x) = p(x) + q(x)\) and \((pq)(x) = p(x)q(x)\) follow from the definitions by identities in \( R \). The kernel of this homomorphism may not be trivial, and it is necessary to consider the ring \( R[x] \) of formal polynomials in some contexts. Often, however, it is safe to blur the distinction between the formal polynomial and the function; in particular we write \( p(x) \) for either, and if \( a \in R \) then \( p(a) \) denotes the value of the function at \( a \).
Theorem 3. If $R$ is a field then $R[x]$ is an integral domain; further for $p, q \in R[x]$, \( \deg(pq) = \deg(p) + \deg(q) \), with the usual understanding for $-\infty$.

**Proof:** The coefficient of $x^{n+m}$ in $pq$ is $p_nq_m$, where $p_n, q_m$ are the leading coefficients of $p, q$, when $p, q \neq 0$. If $R$ is an integral domain this is nonzero, proving the theorem.

Over an arbitrary ring $R$ a polynomial $p$ is said to be monic if its leading coefficient is 1. If $p$ and $q$ are both monic then $pq$ is monic, and $\deg(pq) = \deg(p) + \deg(q)$.

**Theorem 2.** If $F$ is a field then $F[x]$ is a Euclidean domain; in fact, the degree is a Euclidean norm.

**Proof:** Let $p(x), d(x)$ be given with $d_n$ the leading coefficient of $d(x)$; we must write $p(x) = q(x)d(x) + r(x)$ where $\deg(r(x)) < n$. If $\deg(p(x)) < n$ we are done. Otherwise we proceed by induction on $\deg(p)$. Let $p_m$ be the leading coefficient of $p(x)$; then $p(x) - (a_m/d_n)x^{m-n}d(x)$ has lower degree than $p(x)$.

The existence of $q$ and $r$ as in the proof is called the division law for $F[x]$. If $R$ is a factorial domain $p \in R[x]$ is called primitive if the gcd of its coefficients is 1. Note that monic polynomials are primitive.

**Theorem 4.** Suppose $R$ is a factorial domain, and $F$ its field of fractions.

a. The product of primitive polynomials is primitive.

b. Any $p(x) \in R[x]$ equals $cp'(x)$, for $c \in R$, $p'(x) \in R[x]$ primitive, uniquely up to a unit.

c. If $p(x)$ is primitive and $p(x) = q(x)r(x)$ in $F[x]$ then $p(x) = q'(x)r'(x)$ in $R[x]$ where $q'(x), r'(x) \in R[x]$ are primitive.

d. $R[x]$ is a factorial domain.

**Proof:** For part a, suppose

$$p(x) = p_0 + \cdots + p_nx^n = (q_0 + \cdots + q_sx^s)(r_0 + \cdots + r_tx^t) = q(x)r(x),$$

where $q(x), r(x)$ are primitive, and suppose there is a prime element $a \in R$ dividing all the $p_i$. Let $j$ be least such that $a$ does not divide $q_j$, and let $k$ be least such that $a$ does not divide $r_k$. Now,

$$p_{i+j} = q_0r_{i+j} + \cdots + q_ir_j + \cdots + q_{i+j}r_0;$$

$a$ divides all terms in the sum other than $q_ir_j$, and $a$ divides the sum, so $a$ divides $q_ir_j$. But this is a contradiction, since then $a$ must divide $q_i$ or $r_j$. For part b, if $p(x) = cp'(x)$ then $c$ is a gcd of the coefficients of $p(x)$ and the claim follows. For part c, $p(x) = (a/b)q'(x)(c/d)r'(x)$ where $a, b, c, d \in R$ and $q'(x), r'(x) \in R[x]$ are primitive. Hence $bdp(x) = acq'(x)r'(x)$; by parts 1 and 2, $bd \sim ac$ and the claim follows. For part d, suppose $p(x) \in R[x]$; then $p(x) = cp'(x)$ where $c \in R$ and $p'(x)$ is primitive in an essentially unique way. Any factorization of $p(x)$ must be a product of a factorization of $c$ and a factorization of $p'(x)$. Since $R$ is factorial $c$ has an essentially unique factorization. By part 3 $p'(x)$ has a factorization into primitive irreducible polynomials, which is essentially unique in $F[x]$ by theorem 3; it is therefore essentially unique in $R[x]$.

The element $c$ in theorem 4.b is called the content of $p$; it is the gcd of the coefficients. The polynomial $p' = p/c$ is called the primitive part. It follows that the content of $pq$ is the product of the contents of $p$ and $q$. It also follows that in $R[x]$ a divisor of a primitive polynomial is primitive. Part c is known as Gauss’ lemma.

A root of a polynomial $p \in R[x]$ is an $a \in R$ such that $p(a) = 0$. Suppose the ring of coefficients is a field $F$ and $p \in F[x]$ is nonzero. We claim that $a \in F$ is a root of $p$ iff $x - a$ is a divisor of $p$. Indeed, write $p(x) = q(x)(x - a) + r$ where $r$ is a constant, which is possible since the degree is a Euclidean norm; then
$p(a) = 0$ iff $r = 0$ iff $(x - a)|p(x)$. By theorem 4 the claim also follows if the ring of coefficients is a factorial domain.

If the ring of coefficients is a field and $a$ is a root there will be some maximum integer $m > 0$ such that $(x - a)^m$ divides $p$; $m$ is called the multiplicity of $a$. If $m > 1$ $a$ is said to be a multiple root. It follows that if $\text{deg}(p) = n$ then $p$ has at most $n$ roots, where each root is counted according to its multiplicity. This also follows if the ring of coefficients is an integral domain, by considering the coefficients as elements of the field of fractions. It also follows that if the ring of coefficients is an infinite integral domain (which it is if the characteristic is 0 for example) then the only polynomial representing the identically 0 function is the 0 polynomial.

If $p$ is a polynomial in $R[x]$ its derivative $p'$ is defined to be the polynomial $\sum_{i>0} ip_ix^{i-1}$. The map $p \mapsto p'$ is a derivation on $R[x]$, that is, it satisfies the algebraic laws

\[(p + q)' = p' + q', \quad (cp)' = cp' \text{ for } c \in R, \quad (pq)' = pq' + p'q;\]

the first two state that the map is $R$-linear, and the third is called Liebnitz’ rule or the product rule. We leave it as an exercise to show that $p \mapsto p'$ is a derivation, $(q')' = iq^{-1}q'$ (for any derivation), and the chain rule $(p \circ q)' = (p' \circ q)q'$. The notation $p^{(i)}$ is used for the result of iterating the derivative operation $i \geq 0$ times, with $p^{(0)}$ equaling $p$.

We suppose for the remainder of the section that the ring of coefficients is a field $F$. If $a$ is a root of $p$ then $p = (x - a)q$ for some nonzero $q$, so $p' = (x - a)q' + q$ and $p'(a) = q(a)$. Now, $a$ is a multiple root of $p$ iff it is a root of $q$, and this is so iff $a$ is a root of $p'$. In particular $p$ has a multiple root iff $\gcd(p, p') \neq 1$; this is called the derivative test for multiple roots.

If $p = (x - a)^m q$ where $a$ is not a root of $q$ then $p' = (x - a)^{m-1}(mq + (x - a)q')$. If $F$ has characteristic 0, or $m$ is not divisible by the characteristic, then $mq + (x - a)q'$ does not have $a$ as a root (and is nonzero), so $a$ has multiplicity $m - 1$ in $p'$. Inductively, if $F$ has characteristic 0, or $m$ is less than the characteristic, then $m$ is the least value of $i$ such that $a$ is not a root of $p^{(i)}$.

A nonzero $p \in F[x]$ is said to split in $F$ if all its roots lie in $F$. This is readily seen to be so iff $p$ factors into linear factors, or if $p$ has $n$ roots in $F$, counting multiplicities. A field $F$ is said to be algebraically closed if every $p \in F[x]$ splits in $F$. Clearly this is so iff every polynomial has some root in $F$. The complex numbers $\mathbb{C}$ are an example of an algebraically closed field.

The fact that the complex numbers are algebraically closed is called the fundamental theorem of algebra (the algebra of polynomials). There is a short proof, which assumes some topology, namely the fact that if $D$ is a closed disc in the complex plane and $f : D \rightarrow \mathbb{C}$ is continuous then $|f|$ assumes the inf of its values in $D$ (where $|z|$ denotes the norm of the complex number $z$). The reader should have no trouble proving this after reading chapter 17.

By induction it suffices to show that if $p(x)$ is a nonconstant polynomial with complex coefficients, then $p$ has at least one root in the complex numbers. Let $p(x) = \sum_{i=0}^{n} a_i x^i$ where $n \geq 1$. We may assume $a_n = 1$; then $|p(x)| \geq |x|^n (1 - \sum_{i=1}^{n-1} |a_i||x|^{n-i})$. This becomes arbitrarily large as $|x| \rightarrow \infty$, and so if $m = \inf\{|p(x)| : x \in C\}$ there is an $R$ such that $|x| > R$ implies that $|p(x)|$ is bounded away from $m$. Thus, $m = \inf\{|p(x)| : |x| \leq R\}$, and as noted above there is an $x_0$ with $|p(x_0)| = m$ (and $|x_0| \leq R$). To complete the proof, it suffices to show that if $|p(x_0)| > 0$ then there is a direction one can go from $x_0$ in which $|p(x)|$ decreases. Letting $q(y) = p(x_0 + y)/p(x_0)$, $q(y) = 1 + \sum_{i=1}^{n} b_i y^i$ where not all $b_i$ are 0. Let $k$ be the least $i$ such that $b_i \neq 0$, and write $b_k$ as $|b_k| e^{i\theta}$. For real $r > 0$

$$|q(re^{(\pi - \theta)/k})| \leq 1 - r^k \left(|b_k| - \sum_{i=k+1}^{n} |b_i|^r i^{-k}\right).$$

For sufficiently small $r$ the second term is positive, and $q(y) < 1$. 44
Theorem 5. If $F$ is algebraically closed then $F$ is infinite.

Proof: Suppose $F$ is finite, say $F = \{a_i : 1 \leq i \leq n\}$. Consider the polynomial $p(x) = 1 + \prod_i (x - a_i)$. Clearly this does not have any $a_i$ as a root.

2. Interpolation. Let $F$ be a field, and suppose $f : F \mapsto F$; the “divided difference” functions are defined by the following recursion.

$$f^{[0]}(x_0) = f(x_0),$$
$$f^{[i+1]}(x_0, \ldots, x_{i+1}) = \frac{f^{[i]}(x_0, \ldots, x_i) - f^{[i]}(x_1, \ldots, x_{i+1})}{x_0 - x_{i+1}}.$$ 

These functions are defined when the values of the arguments are distinct; below it is shown that the domain can be extended when the field is $\mathcal{R}$.

Lemma 6.

$$f^{[i]}(x_0, \ldots, x_i) = \sum_{j=0}^{i} \frac{f(x_j)}{\prod_{k \neq j} (x_j - x_k)}.$$ 

Proof: This may be proved by induction on $i$ by a straightforward calculation.

An $n$-ary function $f : S^n \mapsto S$ on a set $S$ is said to be symmetric if $f(x_{\pi(1)}, \ldots, x_{\pi(n)}) = f(x_1, \ldots, x_n)$ for any permutation $\pi$ of $\{1, \ldots, n\}$ and any values $x_1, \ldots, x_n$. It is obvious from the lemma that the functions $f^{[i]}$ are symmetric.

Theorem 7. Let $x_0, \ldots, x_n$ be distinct elements of $F$, and $x$ distinct from these. Then there is a unique polynomial $p$ of degree at most $n$ such that $p(x_i) = f(x_i)$ for $0 \leq i \leq n$. Further,

a. $p = \sum_{i=0}^{n} f(x_i)/\prod_{j \neq i} (x_i - x_j)$; 

b. $p = \sum_{i=0}^{n} f^{[i]}(x_0, \ldots, x_i) \prod_{j=0}^{i-1} (x_i - x_j)$; 

c. $f(x) = p(x) + f^{[n+1]}(x_0, \ldots, x_n, x) \prod_{i=0}^{n} (x - x_i)$.

Proof: Write $g_i$ for $\prod_{j \neq i} (x_i - x_j)/\prod_{j \neq i} (x_j - x_i)$; clearly $g_i(x_j) = \delta(i, j)$, and so the polynomial $p$ of part a satisfies $p(x_i) = f(x_i)$ for $0 \leq i \leq n$. If $p_1$ and $p_2$ both satisfy the requirement then $p_1 - p_2$ has degree at most $n$ and at least $n + 1$ roots, so equals 0. That the unique $p$ of part a satisfies part b is proved by induction on $n$, the basis $n = 0$ being trivial. Letting $p_k$ denote the interpolating polynomial on $x_0, \ldots, x_k$, by induction $p_k$ for $k < n$ is the sum up to $k$ of the terms of the expression. Now, $p_n - p_{n-1}$ has the $n$ zeros $x_0, \ldots, x_{n-1}$, so is of the form $c \prod_{j=0}^{n-1} (x - x_j)$ for some constant $c$. Clearly $c$ is the leading coefficient of $p = p_n$; this may be seen to be as stated by the first claim and lemma 6. Part c now follows (again, $p_n$ is the first $n + 1$ terms of $p_{n+1}$).

The expression in part a of theorem 7 is called the Lagrange form of the interpolating polynomial; that in part b is called the Newton form. Lemma 6 gives a method for computing the $p^{[i]}$, called the divided difference table. This involves computing the differences $f^{[i]}(x_t, \ldots, x_{t+i})$ for $t = 0, \ldots, n - i$, for $i = 0, \ldots, n$ successively (for those familiar with the term, this is an example of dynamic programming). The table can be arranged in a triangle, as in figure 1 (this is exactly the same arrangement as Pascal’s triangle for computing the binomial coefficients).

The divided difference table can be used to compute $p(x)$. Since $p$ and $f$ have the same values at the $x_i$, $p^{[i]}(x_t, \ldots, x_{t+i}) = f^{[i]}(x_t, \ldots, x_{t+i})$ for $0 \leq t \leq n - i$, $0 \leq i \leq n$. Also, $p^{[n+1]}(x_0, \ldots, x_n, x) = 0$; this can be seen by theorem 7.c. The divided difference table for $p$ on $x_1, \ldots, x_n, x$ can thus be computed by adding the value 0 as the new apex of the divided difference, and computing backwards. The new entry of the base row is $p(x)$.
An important special case of Newton interpolation occurs when the $x_i$ are equally spaced, say $x_i = x_0 + id$. Usually $F$ is $\mathbb{R}$, but we only need characteristic 0. Let $y(i) = f(x_i)$. For a function $y : \mathbb{Z} \mapsto F$ let \( \Delta y(i) = y(i+1) - y(i) \). \( \Delta \) is called the forward difference operator; it is readily seen to be linear. Also, \[
abla^ny(i) = \sum_{j=0}^{n} (-1)^j \binom{n}{j} y(i+j).
\]

It is readily verified by induction that \[
\n[n](x_0, \ldots, x_{t+i}) = \frac{\Delta^ny(t)}{t!d^n}.
\]

Using this it is straightforward to show that for the interpolating polynomial $p$,
\[
p(x_0 + td) = \sum_{i=0}^{n} \Delta^iy(0) \binom{t}{i} \text{ where } \binom{t}{i} = \frac{t(t-1) \cdots (t-i+1)}{i!}
\]

for all $t \in F$. The computation of the coefficients no longer involves division, and the table is called the difference table rather than the divided difference table.

It is readily verified that for $k > 0$
\[
\Delta \binom{x}{k} = \binom{x}{k-1}.
\]

For $f : \mathbb{Z} \mapsto F$ if
\[
\Delta f = \sum_{k=0}^{n} a_k \binom{x}{k}
\]

let
\[
g = \sum_{k=1}^{n+1} a_k \binom{x}{k}, \quad x \in \mathbb{Z}.
\]

Then \( \Delta(g - f) = 0 \), from which it follows that $g - f$ is constant. This shows that if $\Delta f$ is a polynomial of degree $n$ then $f$ is a polynomial of degree $n + 1$. It follows that $f$ is a polynomial of degree at most $n$ iff $\Delta^{n+1}f = 0$.

Those polynomials $f$ such that $f(n) \in \mathbb{Z}$ for $n \in \mathbb{Z}$ are of interest; they are sometimes called numerical polynomials. Again usually $F = \mathbb{R}$, but characteristic 0 suffices.

**Theorem 8.** Suppose $f$ is a polynomial; the following are equivalent.
1. $f(x) \in \mathbb{Z}$ for all $x \in \mathbb{Z}$.
2. $f(x) \in \mathbb{Z}$ for all sufficiently large $x \in \mathbb{Z}$.
3. $f = \sum a_i \binom{x}{i}$ where the $a_i$ are integers.

**Proof:** 1$\Rightarrow$2 is trivial. 2$\Rightarrow$3 follows by induction on the degree $n$, the basis $n = 0$ being straightforward. Now, $\Delta f$ has the required property, so the $a_i$, $i > 0$, are integers. Choosing $x \in \mathbb{Z}$ sufficiently large, $a_0$ must be an integer also. 3$\Rightarrow$1 follows since binomial coefficients are integers.
Any polynomial can be written as a linear combination of functions of the form \( x \cdot (x - i + 1) \), and conversely. The translation between these two forms may be expressed using Stirling numbers; the conventions for these vary, and we use those of [Knuth]. We define \( S_1(k, j) \), \( S_2(k, j) \), \( 0 \leq j \leq k \), to be the unique values such that

\[
k! \binom{x}{k} = \sum_{j=0}^{k} (-1)^{k-j} S_1(k, j) x^j, \quad x^k = \sum_{j=0}^{k} S_2(k, j) j! \binom{x}{j}.
\]

We leave it as an exercise to show

\[
S_1(k, 0) = \delta(k, 0), \quad S_1(k, k) = 1, \quad S_1(k, j) = (k-1)S_1(k-1, j) + S_1(k-1, j-1)
\]

\[
S_2(k, 0) = \delta(k, 0), \quad S_2(k, k) = 1, \quad S_2(k, j) = jS_2(k-1, j) + S_2(k-1, j-1)
\]

where \( 0 < j < k \). From this it is immediate that the values of \( S_1 \), \( S_2 \) are nonnegative integers.

In the case \( F = \mathbb{R} \) some additional facts about the functions \( f^{[k]} \) can be proved; for the rest of the section we assume \( F = \mathbb{R} \) and some facts from basic calculus. Suppose \( f \) is sufficiently differentiable; then

\[
f^{[n+1]}(x_0, \ldots, x_n, x) = \frac{f^{(n+1)}(\xi)}{k!}
\]

for some \( \xi \in (a, b) \). If \( n = 0 \) this is just the mean value theorem. For \( n > 0 \) let \( p_1 \) interpolate \( x_0, \ldots, x_n, x \) and let \( e_1 = f - p_1 \); then \( e_1 \) has \( n + 2 \) zeros. Applying Rolle’s theorem \( n + 1 \) times, \( e_1^{n+1} \) has a zero \( \xi \) in \( (a, b) \). But \( p_1^{(n+1)}(\xi) \) equals \( (n + 1)! \) times the leading coefficient \( f^{[n+1]}(x_0, \ldots, x_n, x) \) of \( p_1 \). Note that this gives an alternative form for the “error term” in theorem 7.c.

Provided \( f \) is sufficiently differentiable, the functions \( f^{[n]} \) may be extended continuously to all values in a unique manner. The Newton form gives the unique \( n \)th degree polynomial matching the derivatives \( f^{(0)}, \ldots, f^{(r)} \) at \( x_i \), where \( r \) is the number of times \( x_i \) occurs in the list of interpolating points. Theorem 7.c holds for all values of the variables. The proofs of these facts are left to the exercises.

### 3. Multi-variable polynomials.

Recalling that \( R \) is a commutative ring, the ring \( R[x_1, \ldots, x_k] \) of polynomials in the variables \( x_1, \ldots, x_k \) can be described as follows. By a monomial is meant an expression of the form \( x_1^{e_1} \cdot x_k^{e_k} \). A polynomial is an expression of the form \( \sum a_{\nu} \nu \) over finitely many distinct monomials \( \nu \), where \( a_{\nu} \in R^{\neq} \); or 0. (Alternatively it as a function from \( N^k \) to \( R \) which is nonzero only finitely often.) We also write \( R[x] \) for \( R[x_1, \ldots, x_k] \).

We leave it to the reader to fill in the details of the following.

- Each polynomial determines a function in \( R^{(R^k)} \), which is manifest from the expression.
- Operations +, \( \times \) can be defined uniquely on \( R[x] \), so that the map from a polynomial \( p \) to the function \( p(x) \) is a ring homomorphism.
- There is an obvious isomorphism between \( R[x_1, \ldots, x_k] \) and \( R[x_1, \ldots, x_{k-1}][x_k] \).
- If \( R \) is an integral domain \( R[x] \) is; if also \( R \) is infinite then \( R[x] \) is infinite and the map from \( R[x] \) to \( R^{(R^k)} \) is injective.

The degree (or total degree) of the monomial \( x_1^{e_1} \cdot x_k^{e_k} \) is defined to be \( e_1 + \cdots + e_k \). A polynomial in \( R[x] \) is called homogeneous if every monomial has the same degree. If the degree is \( m \) the polynomial is said to be homogeneous of degree \( m \). If \( p, q \) are homogeneous of degree \( m \) then \( p + q \) is, unless it equals 0. If \( p \) is homogeneous of degree \( m \) and \( q \) is homogeneous of degree \( n \) then \( pq \) is homogeneous of degree \( m + n \), unless it is 0 (which is impossible if \( R \) is an integral domain).

**Theorem 9.** If \( R \) is Noetherian then \( R[x] \) is.
Clearly symmetric polynomials define symmetric functions, and are homogeneous. For an abelian monoid \(G\) is considered a polynomial in \(n\) variables, and leading coefficient of \(a\) is the subring of \(R\) generated by \(a\). A derivation on \(R\), let \(p\) be the ideal generated by \(a\) and degree \(m\). Then \(p\) is homogeneous of degree \(m\) such that for every \(p \in I\) there is \(p' \in I'\) with the same degree and leading coefficient (for then \(p - p' \in I\) has lower degree). If \(p\) has degree \(d\) and leading coefficient \(a\), let \(e\) be least such that \(a \in J_e\). Then \(e \leq d\), and there is a polynomial in \(I'\) with leading coefficient \(a\) and degree \(e\).

If \(F\) is a field, by results of this chapter and the last \(F[x]\) is a Noetherian factorial domain. It is not, however, a principal ideal domain if there is more than one variable (exercise).

The notation \(R[x]\) has the following generalization. Suppose \(R, S\) are commutative rings with \(R \subseteq S\), and \(a_1, \ldots, a_k \in S\). Then \(R[a_1, \ldots, a_k]\), or \(R[a]\), denotes \(\{p(a_1, \ldots, a_k) : p \in R[x]\}\). This is readily seen to be the subring of \(S\) generated by \(R \cup \{a_1, \ldots, a_k\}\). It is the image of \(R[x]\) under the homomorphism which maps the polynomial \(p\) to the element \(p(a_1, \ldots, a_k)\) of \(S\). This homomorphism is called the evaluation map. Its kernel is those \(p \in R[x]\) for which \(p(a_1, \ldots, a_k) = 0\). \(R[a]\) is isomorphic to \(R[x]\) exactly if this ideal is trivial.

Write \(H_m\) for the collection of polynomials homogeneous of degree \(m\) together with 0. Then for the additive groups, \(R[x] = \bigoplus_m H_m\). Also, \(H_mH_n \subseteq H_{m+n}\). A ring with these properties is called graded, and the elements of \(H_m\) are called the elements homogeneous of degree \(m\). More generally the index set can be a commutative monoid \(G\), and the ring is called \(G\)-graded, so that by graded we mean \(N\)-graded.

The partial derivative \(\frac{\partial p}{\partial x_i}\) of a polynomial \(p \in R[x]\) with respect to \(x_i\) is just the derivative, when \(p\) is considered a polynomial in \(S[x]\) where \(S = R[x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]\). The operator \(p \mapsto \frac{\partial p}{\partial x_i}\) is a derivation on \(R[x]\), satisfying the chain rule

\[
\frac{\partial p(q_1, \ldots, q_t)}{\partial x_i} = \sum_{j=1}^{t} \frac{\partial p}{\partial y_j} \left( \frac{\partial y_j}{\partial x_i} \right)
\]

where \(q_i \in R[x]\) and \(p \in R[y_1, \ldots, y_t]\); we leave the proof as an exercise, and also the proof of Euler’s theorem that for \(p\) homogeneous of degree \(m\)

\[
\sum_{i=1}^{k} x_i \frac{\partial p}{\partial x_i} = mp.
\]

The multinomial theorem states that

\[
(x_1 + \cdots + x_k)^n = \sum_{i_1 + \cdots + i_k = n} \frac{n!}{i_1! \cdots i_k!} x_1^{i_1} \cdots x_k^{i_k}.
\]

This may be proved using the observation that the coefficient of \(x_1^{i_1} \cdots x_k^{i_k}\) is the number of ways of selecting \(x_j\) \(i_j\) times for each \(j\), when selecting one term from each copy of \((x_1 + \cdots + x_n)\) in the product; this is exactly the number of partitions of an \(n\) element set into parts of size \(i_1, \ldots, i_k\). The exercises show that this is \(n!/(i_1! \cdots i_k!)\). These numbers are called the multinomial coefficients; when \(k = 2\) they are the binomial coefficients.

A polynomial in \(R[x_1, \ldots, x_n]\) is called symmetric if it is left unchanged by a permutation of the variables. Clearly symmetric polynomials define symmetric functions, and are homogeneous. For \(0 \leq k \leq n\) the polynomial

\[
\sigma^k_n(x) = \sum_{i_1 < \cdots < i_k} x_{i_1} \cdots x_{i_k}
\]

48
is symmetric, homogeneous of degree \( k \). These polynomials are called the elementary symmetric polynomials. It is readily verified that

\[
(y - x_1) \cdots (y - x_n) = \sum_{i=0}^{n} (-1)^{n-i} \sigma_{n-i}^n(x) y^i. \tag{1}
\]

**Theorem 10.** If \( p \in R[x_1, \ldots, x_n] \) is a symmetric polynomial then there is a \( P \in R[y_0, \ldots, y_n] \) such that \( p(x_1, \ldots, x_n) = P(\sigma_0^n, \ldots, \sigma_n^n) \).

**Proof:** Define the type of a term to be the sequence \( I = i_1 \leq \cdots \leq i_n \) of degrees of its variables. The terms of given type are permuted among themselves, so it suffices to show that a polynomial consisting of the sum of all terms of a given type (call such a polynomial special) can be expressed as a polynomial in the \( \sigma_i^n \) (in which case say the type is expressible). Order the types first by dictionary order, that is, \( i_1 \ldots i_n \) precedes \( j_1 \ldots j_n \) if in the least index \( t \) where they differ \( i_t > j_t \). Clearly, the types with \( i_n = 1 \) are expressible. Given a type \( I = i_1 \ldots i_n \), let \( t \) be least such that \( i_t > 0 \), and consider \( J = j_1 \ldots j_n, K = k_1 \ldots k_n \), where \( j_s = 1 \) for \( s \geq t \) and 0 otherwise; and \( k_s = i_s - j_s \). It is not difficult to verify that the types in the product of the special polynomials of types \( J, K \) are \( I \) and types preceding \( I \) in the order (consider how the exponent vectors “line up”).

**Theorem 11 (Vandermonde determinant).** Let \( M \) be the \( n \times n \) matrix where \( M_{ij} = x_i^{n-j}, 1 \leq i, j \leq n \). Then \( \det(M) = \prod_{i < j} (x_i - x_j) \).

**Proof:** The proof is by induction on \( n \); the theorem is immediate for \( n = 2 \) and vacuously true for \( n = 1 \). Let \( \sigma_i^{n-1} \) denote \( \sigma_i^n(x_2, \ldots, x_n) \). By (1), if for \( 2 \leq i \leq n \) column \( i \) is multiplied by \(( -1 )^{n-1} \sigma_i^{n-1} \) and added to the first column, \( M_{i1} \) becomes \( \prod_{i=2}^n (x_1 - x_i) \) and \( M_{i1} \) becomes 0 for \( i > 1 \). The claim now follows by induction.

This theorem may also be proved as follows. If \( x_i \) is replaced by \( x_j \) the determinant \( p \) vanishes. Hence \( p \), considered as an element of \( S[x_i] \) where \( S = \mathbb{Z}[x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n] \), is divisible by \( x_i - x_j \). It follows that \( p \) is divisible by the product of these where \( i < j \), and since this has the same degree \( n(n-1)/2 \) as \( p \), it can differ by at most a constant. The constant is readily seen to be 1 by considering the term of \( p \) corresponding to the diagonal.

Suppose \( a = \sum_{i=0}^{n} a_i x^i \) and \( b = \sum_{i=0}^{m} b_i x^i \) are polynomials in \( R[x] \), where \( m, n > 0 \). Let \( M \) be the \((m+n) \times (m+n)\) matrix where

\[
\begin{array}{lc}
\text{for } 0 \leq i < m, & M_{ij} = \begin{cases} a_{m-(j-i)} & \text{if } i \leq j \leq i + n \\ 0 & \text{otherwise} \end{cases} \\
\text{for } m \leq i < m+n, & M_{ij} = \begin{cases} b_{n-(j-i+m)} & \text{if } i - m \leq j \leq m \\ 0 & \text{otherwise} \end{cases}
\end{array}
\]

For example, with \( n = 3, m = 2 \) \( M \) is

\[
\begin{bmatrix}
a_3 & a_2 & a_1 & a_0 \\
& a_3 & a_2 & a_1 & a_0 \\
b_2 & b_1 & b_0 \\
& b_2 & b_1 & b_0 \\
& b_2 & b_1 & b_0
\end{bmatrix}
\]

The above matrix is known as the Sylvester matrix. Row \( i \) is the coefficients of the polynomial \( x^i a \) for \( 0 \leq i < m \), and row \( m+i \) those of \( x^i b \) for \( 0 \leq i < n \). When the ring \( R \) is a field the determinant of the matrix is 0 iff these polynomials are linearly dependent, which is so iff there are polynomials \( \alpha \) of degree less than \( m \) and \( \beta \) of degree less than \( n \), not both 0, such that \( \alpha a + \beta b = 0 \). It follows that (when \( R \) is a field) the determinant is nonzero iff \( a \) and \( b \) are relatively prime (in \( R[x] \)). If they are relatively prime then such
an $\alpha, \beta$ cannot exist, since $a$ would have to divide $\beta$, a polynomial of smaller degree. Conversely if $d$ is a nontrivial common divisor of $A$ and $B$ then let $\alpha = b/d$ and $\beta = -a/d$.

The determinant of the Sylvester matrix is called the resultant of the polynomials $a$ and $b$. Write $a$ ($b$) for the vector of coefficients of $a$ ($b$). Considering the $a_i, b_i$ as indeterminates, det$(M)$ is a polynomial in $\mathcal{Z}[a, b]$, which we denote as $\rho(a, b)$ or simply $\rho$. The resultant is the value of $\rho$ when the coefficients of the polynomials are substituted for the indeterminates. It is easily seen that $\rho$ is homogeneous of degree $m + n$, $\rho$ contains the monomial $a_0^m b_0^n$ with coefficient 1, and $\rho(b, a)$ is up to sign the same polynomial.

**Lemma 12.** There are polynomials $\alpha, \beta \in \mathcal{Z}(a, b, x)$ (of degree in $x$ less than $m, n$ respectively) such that $\rho = a\alpha + b\beta$.

**Proof:** For each $j > 0$ add $x^j$ times column $j$ to column 0. In the resulting matrix $M'$, $M_{00}$ equals $x^j a$ for $0 \leq i < m$ and $x^{i-m}b$ for $m \leq i < m + n$. Over $\mathcal{Z}(a, b, x)$ this does not change the determinant. Expanding by minors on the first column yields the lemma.

Introduce new variables $r_1, \ldots, r_n$ and $s_1, \ldots, s_m$. Let $a = a_n(x - r_1) \cdots (x - r_n), b = b_n(x - s_1) \cdots (x - s_n)$; then $a_i = a_n(-1)^{n-i}\sigma_{n-i}(r), b_i = b_m(-1)^{m-i}\sigma_{m-i}(s)$.

**Lemma 13.** With $a, b$ as above,

$$\rho(a, b) = a_n^m b_m^n \prod_{i,j}(r_i - s_j).$$

**Proof:** Substituting $r_i$ for $x$ in the expression of lemma 12, it follows that $\rho$ is divisible by $r_i - s_j$ for any $i, j$ (the term involving $a$ drops out and the term involving $b$ has a factor $(r_i - s_j)$ for each $j$). Clearly $\rho$ is of the form $a_n^m b_m^q$ where $q$ is a polynomial in the $r_i, s_j$; it follows that $\rho = p a_n^m b_m^n \prod_{i,j}(r_i - s_j)$ for some polynomial $p$ in the $r_i, s_j$. The terms in $\rho$ of largest total degree in the $r_i$ consist of the single monomial $a_n^m b_m^n(r_1 \cdots r_n)^m$. Comparing coefficients, $p = 1$.

It follows that $\rho = a_n^m \prod_i b(r_i)$. Letting $b$ be the derivative $a'$, it is easily seen that $a'(r_i) = a_n \prod_{j, j \neq i}(r_i - r_j)$, so that

$$\rho(a, a') = a_n^{2n-1} \prod_{j \neq i}(r_i - r_j).$$

The quantity

$$\delta(a) = a_n^{2n-2} \prod_{j \neq i}(r_i - r_j)$$

is called the discriminant of $a$. The equation $\rho = a_n \delta$ gives a means of computing it. For example, for $n = 2$

$$\rho = a_2(4a_0 a_2 - a_1^2), \quad \delta = 4a_0 a_2 - a_1^2.$$ The discriminant is frequently defined with an additional factor $(-1)^{n(n-1)/2}$, in which case for $n = 2$ it becomes $a_1^2 - 4a_0 a_2$.

Suppose the ring is a field $F$. It will be shown in section 9.3 that given polynomials $a, b \in F[x]$ there is an extension $E \supseteq F$ in which $a, b$ both split. Using lemma 13 the following are easily shown.

- Letting the $r_i$ be the roots of $a$ and the $s_i$ the roots of $b$, the resultant is 0 iff the two polynomials have a common root in (any such) $E$, provided $a_n b_m \neq 0$.
- $a$ has a multiple root iff its discriminant equals 0, provided $a_n \neq 0$.
- The coefficients of a monic polynomial are, up to sign, the elementary symmetric polynomials of the roots.
- The discriminant of a monic polynomial is, up to sign, the square of the Vandermonde determinant in the roots (a symmetric polynomial in the roots).

**Exercises.**

1. Prove that the derivative operator has the stated algebraic properties. Do the same for the partial derivative operators.
2. Prove the facts stated near the beginning of section 3.
3. Show that, for $F$ a field and $k > 1$, $F[x_1, \ldots, x_n]$ is not a principal ideal domain.
4. Show that $\binom{n}{k}$ is the number of ways of partitioning an $n$ element set into parts of size $i$ and $n - i$.
   Hint: use induction. Then show that $n!/(i_1! \cdots i_k!)$ is the number of ways of partitioning an $n$ element set into parts of size $i_1, \ldots, i_k$.
5. Prove the recursion equations for the Stirling numbers. Hint: For $S_2$, write $x^{k+1} = xx^k$ and use $xx(i) = x(i+1) + ix(i)$ where $x(i) = x(x - 1) \cdots (x - i + 1)$.
6. Fill in the details of the following. The points in $\mathbb{R}^{n+1}$ with unequal components are a dense subset. It follows that $f^{[n]}$ can have at most one extension continuous on all of $\mathbb{R}^{n+1}$. By continuity such an $f$ must be symmetric. To show that $f^{[i]}$ exists, use the recursion, with the extra case

$$f^{[i+1]}(x_0, \ldots, x_i, x_0) = \frac{\partial f^{[i]}}{\partial x_0}(x_0, \ldots, x_n).$$

Inductively, if the $n$th derivative of $f$ exists and is continuous then the $(n - i)$th partials of $f^{[i]}$ exists and are continuous; in particular $f^{[n]}$ exists and is continuous. Using theorem 7.c, with the alternative expression for the error term, and taking a limit, $f^{[n]}(x, \ldots, x) = f^{(n)}(x)/n!$. By continuity, theorem 7.c holds everywhere; in particular $p$ does not depend on the order of the $x_i$. If, for example, the first $r$ values are $x_0$, a direct computation shows $p^{(i)}(x_0) = f^{(i)}(x_0)$ for $0 \leq i < r$.

7. Show that $R[x]$ is not a principal ideal domain if there is more than one variable.
8. Modules.

1. Basic definitions. A module is a commutative group $M$, with a ring $R$ acting on it so that certain additional axioms are satisfied. $M$ is written additively, and the action as $rx$, $r \in R, x \in M$. The axioms which must be satisfied are as follows; the first two state that the multiplicative monoid of $R$ acts on $M$.

$$r(sx) = (rs)x, \quad 1x = x, \quad (r + s)x = rx + sx, \quad r(x + y) = rx + ry,$$

for $r, s \in R, x, y \in M$.

Often, $R$ is fixed, and $M$ is referred to as a module over $R$ or an $R$-module. The map $\psi_r$ defined by $\psi_r(x) = rx$ is a homomorphism of the additive group of $M$; such a homomorphism is called a principal homomorphism. The map $r \mapsto \psi_r$ is a ring homomorphism from $R$ to the ring of endomorphisms of the additive group; this need not be injective. If $R$ is commutative $\psi_r$ is an $R$-module homomorphism.

**Theorem 1.** If $M$ is an $R$-module, $r, s \in R, x, y \in M$, then

$$0x = 0, \quad r0 = 0, \quad r(-x) = -(rx), \quad (-r)(-x) = rx,$$

$$r(x + y) = rx + ry, \quad (r - s)x = rx - sx, \quad (-1)x = -x$$

**Proof:** Exercise.

The notation $SU$ for $S \subseteq R, U \subseteq M$ may be used, to denote as usual $\{rx : r \in S, x \in U\}$, with $Rx$ denoting $R\{x\}$, and $rM$ denoting $\{r\}M$. Note that $(ST)U = S(TU), (S + T)U = ST + SU,$ and $S(U + V) = SU + SV$, where $T \subseteq R, V \subseteq M$. Also $0M = \{0\}$ and $RM = M$; in particular if $R$ is trivial so is $M$.

Fixing $R$ and regarding the elements of $R$ as unary functions on an $R$-module $M$, $M$ becomes a structure to which the definitions of universal algebra may be applied. $N$ is a submodule of $M$ iff $N$ is a subgroup of $M$, and is closed under the action of $R$, that is, whenever $r \in R$ and $x \in N$ then $rx \in N$. A map $h : M \mapsto N$ between $R$-modules is a homomorphism of $R$-modules if it is a group homomorphism ($h(x + y) = h(x) + h(y)$), and preserves the action of $R$, that is, $h(rx) = rh(x)$ for $r \in R, x \in M$. Such a map is also called linear, or $R$-linear when it is necessary to specify the ring of coefficients. Images or inverse images of submodules under homomorphisms are submodules. A congruence relation in an $R$-module $M$ must respect the action of $R$ in addition to the addition of $M$, that is, if $x \equiv y$ then $rx \equiv ry$, for $r \in R$.

**Theorem 2.** Let $M$ be an $R$-module. Given a submodule $N$ the relation $\equiv_N$ of belonging to the same coset of $N$ is a congruence relation. Given a congruence relation $\equiv$ in $M$, the equivalence class $N_\equiv$ of 0 is a submodule. The maps $N \mapsto \equiv_N$ and $\equiv \mapsto N_\equiv$ are inverse to each other.

**Proof:** We need only show that $\equiv_N$ respects the action of $R$, that is, if $r \in R$ and $x - y \in N$ then $rx - ry \in N$; and that $N_\equiv$ is closed under the action of $R$, that is, if $x \equiv 0$ then $rx \equiv 0$. Both claims are immediate by theorem 1.

As usual by the kernel of an $R$-module homomorphism we mean the submodule. If $N$ is a submodule of $M$ then in the quotient $M/N$, $(x + y) + N = (x + N) + (y + N)$. The submodules of an $R$-module form an algebraic closure system. The submodule generated by a subset $S \subseteq M$ is all elements which can be written in the form $r_1x_1 + \cdots + r_nx_n$ for some $r_1, \ldots, r_n \in R$ and $x_1, \ldots, x_n \in S$. Such an expression is called a linear combination of members of $S$; the $r_i$ are called the coefficients. We write Span($S$) for the submodule generated by $S$.

$R$-modules enjoy basic homomorphism properties obeyed by ideals. Suppose $h : M \mapsto M'$ is a homomorphism with kernel $I$ and image $h[M]$; $C$ is the collection of submodules $J \subseteq G$ with $I \subseteq J$; and $C'$ is
the collection of submodules \( J' \subseteq M' \) with \( J' \subseteq h[M] \). The map \( J \mapsto h[J] \) is a one to one correspondence between \( C \) and \( C' \). For \( J \in C \), \( J/I \) is mapped to \( h[J] \) by the canonical isomorphism. \( M/J \) is isomorphic to \( h[M]/h[J] \). If \( I \) and \( J \) are two submodules, then \( I \cup J = I + J \). The submodules form a monoid under +, with the trivial module being the identity; the map \( I \mapsto h[I] \) is a monoid homomorphism.

If \( I \) and \( J \) are submodules of an \( R \)-module \( N \), one verifies as in theorem 4.12.d that \( (I + J)/I \) is isomorphic to \( I/(I \cap J) \). This isomorphism is often called the second Noether isomorphism.

The product \( \times_i M_i \) of \( R \)-modules \( M_i \) is their product as structures; since the axioms are equations this is an \( R \)-module. The canonical embedding of \( M_i \) as an additive group is an embedding as an \( R \)-module. The elements of \( \times_i M_i \) which are nonzero in only finitely many components comprise a submodule, called the direct sum of the \( M_i \) and denoted \( \oplus_i M_i \). Together with the embeddings of the \( M_i \), this satisfies the usual diagram property for a direct sum.

If \( R \) is \( \mathcal{Z} \), then \( nx = x + \cdots + x \), \( n \) times. In this case submodules are simply subgroups, homomorphisms group homomorphisms, etc.

If \( M \) is generated by a finite subset it is said to be finitely generated; if it is generated by a single element it is called principal. A module is said to satisfy the ascending chain condition if its lattice of submodules does; such a module is called Noetherian. An \( R \)-module is Noetherian iff every submodule is finitely generated; the proof is identical to that of theorem 6.4.a.

A module as defined above is also called a left module. One may also define a right module, as one where \( (rs)x \) equals \( s(rx) \) rather than \( r(sx) \). The ring action is usually written on the right, so that \( x(rs) = (xr)s \). Modules are also defined where there is more than one action. In this case the actions must commute, or respect each other. The rings need not be the same, and some actions might be left and some right. A module where there is one ring acting both on the left and on the right is called two-sided. The requirement that the actions commute is \( r(xs) = (rx)s \). For \( R \) commutative there is no need to distinguish between left and right; writing the action on the left is a common convention.

We consider a ring \( R \) to be a (left) module over itself, by the action of left multiplication. The submodules are those additive subgroups closed under left multiplication, which are called left ideals. Right ideals are also defined; they must be closed under right multiplication, and are right modules. What we have called ideals in chapter 6 are also called two-sided ideals; they are two-sided modules. A ring may satisfy the ascending or descending chain condition for left, right, or two-sided ideals. By a Noetherian ring we mean one whose left ideals satisfy the ascending chain condition, since we consider a ring a module over itself by left multiplication. In a commutative ring there is no distinction between left, right, or two-sided ideals, so this definition agrees with the previous.

Next some additional basic definitions and facts are gathered together.

- The product of a ring \( R \) with itself any number of times is an \( R \)-module with pointwise left “scalar multiplication”\). The direct sum is the usual submodule. The canonical embeddings of \( R \) are \( R \)-module monomorphisms.

- If \( \phi : R \rightarrow S \) is a ring homomorphism and \( M \) is an \( S \)-module, the map \( (r, x) \mapsto \phi(r)x \) is an action of \( R \) on \( M \), which makes \( M \) an \( R \)-module. If \( N \subseteq M \) is an \( S \)-submodule and \( \phi \) is surjective, \( M \) is readily seen to be an \( R \)-submodule.

- A submodule isomorphic to \( R/I \) for some ideal \( I \) equals \( Rx \) where \( x \) is the element corresponding to \( 1 + I \).

- For \( r \in R \) and \( x \in M \), \( r \) is said to annihilate \( x \) if \( rx = 0 \).

- If \( x \in M \) the map \( r \mapsto rx \) is readily seen to be an \( R \)-module homomorphism from \( R \) to \( M \). Its image is the principal submodule \( Rx \). Its kernel, the ring elements which annihilate \( x \), is a left ideal, which may be denoted \( \text{Ann}(x) \), and \( Rx \) is isomorphic to the \( R \)-module \( R/\text{Ann}(x) \).

- For \( R \) commutative, \( \text{Ann}(x) \) is an ideal, and equals \( \text{Ann}(Rx) \); \( R/\text{Ann}(x) \) is a ring.
- More generally $r \in R$ is said to annihilate $M$ if $rM = 0$. The collection of annihilators of $M$ is called the annihilator of $M$; we denote it as $\text{Ann}_R(M)$ or simply $\text{Ann}(M)$. This is a two-sided ideal. For $R$ commutative it is the kernel of the homomorphism $r \mapsto \psi_r$ mentioned in section 1. The annihilator of a left ideal is its annihilator as an $R$-module.

- An $R$-module is called faithful if $\text{Ann}(M) = 0$; equivalently if the map $r \mapsto \psi_r$ is injective.

- $x \in M$ is called a torsion element if $\text{Ann}(x)$ is nontrivial, that is, if $rx = 0$ for some nonzero $r \in R$. $M$ is called torsion-free (torsion) if none (all) of its elements are torsion. If $R$ is an integral domain the torsion elements are readily verified to form a submodule, which is denoted $\text{Tor}(M)$.

- The subgroup $\psi_r^{-1}(0) = \{x \in M : rx = 0\}$ of $M$ is sometimes of interest; the notation $M_r$ will occasionally be used to denote it. It consists of torsion elements, and is annihilated by $r$. If $R$ is commutative it is a submodule of $M$, and for $R$ an integral domain a submodule of $\text{Tor}(M)$. Also if $r, s$ are associates then $M_r = M_s$.

2. Multilinear maps. Given an $R$-module $M$ and a set $S$, the collection $M^S$ of maps from $S$ to $M$ is as usual an $R$-module with the “pointwise” operations $(f + g)(x) = f(x) + g(x)$ and $(rf)(x) = rf(x)$. If $S$ is an $R$-module, the collection $\text{Hom}(S, M)$ of $R$-module homomorphisms (linear maps) from $S$ to $M$ is a subgroup of the additive group of $M^S$; if $R$ is commutative it is a submodule (exercise; commutativity of $R$ is required to show that a scalar multiple of an $R$-linear map is $R$-linear).

More generally for $R$ commutative and $N_i, M$ $R$-modules, a map $f : N_1 \times \cdots \times N_k \rightarrow M$ is said to be $R$-multilinear if it is $R$-linear in each variable, for any values of the remaining variables. That is, for each $i$,

$$f(x_1, \ldots, x_i + x_i', \ldots, x_k) = f(x_1, \ldots, x_i, \ldots, x_k) + f(x_1, \ldots, x_i', \ldots, x_k),$$

$$f(x_1, \ldots, rx_i, \ldots, x_k) = rf(x_1, \ldots, x_i, \ldots, x_k).$$

If addition and scalar multiplication are defined pointwise, the result is an $R$-module, as is readily verified. We denote this module as $L_R(N_1, \ldots, N_k; M)$. As usual, the subscript $R$ may be omitted when no confusion arises from doing so. A map in $L_R(N_1, \ldots, N_k; R)$, which we also denote as $L_R(N^k; R)$, is called a multilinear form.

Note that $L(N; M)$ is $\text{Hom}(N, M)$, enriched with addition and scalar multiplication. With composition as multiplication $\text{Hom}(M; M)$ becomes a ring, even for noncommutative $R$; the identity element is the identity map. Assuming $R$ is commutative, it may also be considered an $R$-module; this situation is considered further in the next section.

A multilinear form $f(x_1, \ldots, x_n)$ from $N^n$ to $M$ is said to be symmetric if it is a symmetric function, which recall from chapter 7 means that $f(x_{\pi(1)}, \ldots, x_{\pi(n)}) = f(x_1, \ldots, x_n)$ for any permutation $\pi$ of $\{1, \ldots, n\}$ and any values $x_1, \ldots, x_n$. It is said to be alternating if $f(x_1, \ldots, x_n) = 0$ whenever $x_i = x_j$ for some $i \neq j$. From

$$f(\ldots, x + y, \ldots, x + y, \ldots) =$$

$$f(\ldots, x, \ldots, x, \ldots) + f(\ldots, x, \ldots, y, \ldots) + f(\ldots, y, \ldots, x, \ldots) + f(\ldots, y, \ldots, y, \ldots),$$

it follows that if $f$ is alternating then

$$f(\ldots, x, \ldots, y, \ldots) = -f(\ldots, y, \ldots, x, \ldots),$$

and hence that

$$f(x_{\pi(1)}, \ldots, x_{\pi(n)}) = (-1)^{\text{sg}(\pi)} f(x_1, \ldots, x_n),$$

where $\text{sg}(\pi)$ is the sign of the permutation, as defined in chapter 5. The form is also said to be antisymmetric. If 2 has a multiplicative inverse then the second (antisymmetry) law implies the first (alternating) law, as can be seen by substituting $x$ for $y$.
3. Algebras. Let $R$ be a commutative ring. An $R$-algebra $A$ is an $R$-module, with a multiplication operation $\times$ defined on it which satisfies the following axioms.

- $A$ is a (not necessarily commutative) ring with $+$ (module addition) and $\times$
- $(rx) \times y = r(x \times y)$, and $x \times (ry) = r(x \times y)$

Given that $A$ is an $R$-module, the second requirement is that $\times$ be $R$-bilinear. Confusion can be avoided when using $+$ and $\times$ to denote the operations of both $R$ and $A$ by using different letters for the arguments.

The notions of an $R$-algebra homomorphism, $R$-subalgebra, and the product of $R$-algebras are left to the reader, along with their basic properties.

If $\phi : R \to A$ is a ring homomorphism then the action $\langle r, a \rangle \mapsto \phi(r)a$ makes $A$ an $R$-module. If further $\phi$ maps $R$ to the center of $A$ then $A$ an $R$-algebra. In particular $R$ is an $R$-algebra. On the other hand given an $R$-algebra $A$, the map $\phi(r) = r1$ from $R$ to $A$ is a ring (in fact $R$-algebra) homomorphism such that $\phi[R]$ is contained in the center of $A$.

An important example of an $R$-algebra is provided by $L_R(M; M)$, with composition for the multiplication operation. If $M$ is $R^n$ this algebra is easily seen to be identical to that of the $n \times n$ matrices with entries from $R$, exactly as in the case of fields (see appendix 2). The image of $r \mapsto r1$ is the matrices which take some constant value on the diagonal.

$R[x_1, \ldots, x_n]$ is another example of an $R$-algebra. This is an $R$-module with polynomial addition and scalar multiplication (by members of $R$), and with polynomial multiplication it becomes an $R$-algebra.

4. Generation of modules. Call a linear combination trivial if all its coefficients are 0 (or it is empty). A set $S$ of elements of an $R$-module $M$ is called linearly independent if the only linear combination of elements of $S$ which equals 0 is the trivial one. Call $S$ linearly dependent otherwise, i.e., if there is a nontrivial linear combination equaling 0. A linearly independent generating set is called a basis, and an $R$-module $M$ is called free if it has a basis.

For any ring $R$ $R^n$ may be considered an $R$-module with componentwise addition and left multiplication. If $R$ is trivial so is $R^n$; otherwise the “standard unit vectors”, i.e., those vectors which are 1 in some component and 0 elsewhere, form a basis, so $R^n$ is a free $R$-module. If a free $R$-module $M$ has a finite basis, of size $n$, it is isomorphic to $R^n$. We will see below that in this case all its bases have size $n$.

Given a ring $R$ and a set $S$ one can define a free module $M$ which is generated by (an injective image of) $S$. $M$ is the submodule of $R^S$ consisting of the functions which are nonzero at only finitely many places (if $|S| = n$ then then $M = R^S = R^n$). Members of $M$ may be written $a_1 x_1 + \cdots + a_n x_n$, $a_i \in R, x_i \in S$. By definition nontrivial linear combinations of members of $S$ are nonzero, and certainly $S$ generates $M$. We call $M$ the free $R$-module generated by $S$. It is specified up to isomorphism by a diagram property; $i : S \to M$ specifies a free module over $S$ if, given a map $\phi$ from $S$ to another module $N$, there is a unique $R$-module homomorphism $\psi : M \to N$ such that $\psi \circ i = \phi$.

As another example of an algebra, suppose $R$ is commutative and $S$ is a monoid. A multiplication may be defined on the free $R$-module $M$ with basis $S$, by setting $(ax)(by) = (ab)(xy)$ and using the distributive law to multiply arbitrary elements. It is easily checked that this multiplication makes $M$ an $R$-algebra, called the free $R$-algebra on $S$. It satisfies the same diagram property as the free module, where now $\phi$ is a monoid homomorphism and $\psi$ an $R$-algebra homomorphism.

The ring of polynomials $R[x]$ is the special case where $S$ is $N$. When $S$ is a group this algebra is called the group algebra, or group ring, of $S$, with coefficients in $R$; a common case is $R = Z$. A convenient notation for group rings is to introduce “variables” $x^g$ for the group elements, with the “exponential” multiplication law $x^g x^h = x^{gh}$.

If an algebra $A$ over a commutative ring $R$ has a basis $S$, the multiplication operation is determined by its restriction to $S$. The product $xy$ of elements of $S$ is some linear combination of elements of $S$, say
The notation \( \sum c_{xyz}^2 \) where the sum is over \( S \) and only finitely many \( c_{xyz} \) are nonzero. The values \( c_{xyz} \) are called structure constants for the algebra.

If \( \{ M_i : i \in I \} \) is a collection of submodules of an \( R \)-module \( M \), their join, which is the submodule generated by their union, is those \( x \in M \) which can be written as \( \sum_{i \in J} x_i \) for some finite subset \( J \subseteq I \) and some \( x_i \in M_i \). This is also called the sum of the submodules.

Suppose \( M \) is the sum of the \( M_i \). It readily follows that the expression of any \( x \in M \) as a finite sum of members of the \( M_i \) is unique iff each \( M_i \) has only the 0 element in common with the sum of the \( M_j \), \( j \neq i \); and also iff whenever \( \sum_{i \in J} x_i = 0 \), \( J \) finite, \( x_i \in M_i \), then all \( x_i \) are 0. When this is so, clearly \( M \) is the direct sum of the \( M_i \) (i.e., the diagram property is satisfied). If \( M \) can be written this way as a direct sum of submodules \( M_i \), it is called their “internal” direct sum; we also say that the \( M_i \) form a direct sum decomposition of \( M \). The subset of the Cartesian product of the \( M_i \) is sometimes called the “external” direct sum. The notation \( \oplus_i M_i \) may be used for either, with the ambiguity resolved by context.

Given submodules \( M_1, \ldots, M_k \subseteq M \) the map \( \langle x_1, \ldots, x_k \rangle \mapsto x_1 + \cdots + x_k \) is a homomorphism from \( M_1 \times \cdots \times M_k \) to \( M_1 + \cdots + M_k \). \( M \) is the sum of the \( M_i \) iff the map is surjective, and the direct sum iff it is bijective. In particular, if \( M \) is the direct sum of finitely many submodules \( M_i \) then it is isomorphic to the product of the \( M_i \). This is no longer true if \( I \) is infinite.

A submodule \( N \subseteq M \) of a module \( M \) is called a summand if there is a module \( L \) such that \( M = N \oplus L \). It is readily seen that \( M = N \oplus L \) iff \( L \) is a system of representatives of the cosets of \( N \).

Observe that a free \( R \)-module \( M \) is the direct sum of principal submodules; indeed, if \( \{ x_i \} \) is a basis for \( M \) then \( M \) is the direct sum of \( \{ Rx_i \} \). On the other hand, \( M \) can be the direct sum of \( \{ Rx_i \} \) without \( \{ x_i \} \) forming a basis, because the map \( r \mapsto rx_i \) might not be injective.

5. Noetherian modules. Noetherian \( R \)-modules have certain closure properties, as follows.

- If \( M \) is Noetherian and \( N \) is a submodule of \( M \) then \( N \) and \( M/N \) are Noetherian;
- if \( N \) and \( M/N \) are Noetherian so is \( M \);
- if \( M_1, \ldots, M_k \) are Noetherian so is \( M_1 \times \cdots \times M_k \).

We leave the proof as an exercise. As a corollary, if \( M \) is finitely generated over a Noetherian ring \( R \) then \( M \) is Noetherian, since for some \( k \) it is the image of \( R^k \) under the \( R \)-module homomorphism which maps the unit vector \( e_i \) to the generator \( g_i \in M \).

6. Module of fractions. Suppose \( R \) is a commutative ring, \( S \subseteq R \) is a multiplicative subset, and \( M \) is an \( R \)-module. For \( m_i \in M \) and \( s_i \in S \), \( i = 1, 2 \), and \( r \in R \), define

\[
m_1/s_1 \equiv m_2/s_2 \text{ iff } \exists s \in S(s(s_1m_2 - s_2m_1) = 0),
m_1/s_1 + m_2/s_2 = (m_1s_2 + m_2s_1)/s_1s_2,
(r/s_1)(m/s_2) = (rm)/(s_1s_2).
\]

It is a routine verification that \( \equiv \) is an equivalence relation respected by the necessary operations, so that the resulting structure is an \( R_S \)-module, called the module of fractions; we denote it as \( M_S \).

Via the homomorphism from \( R \) to \( R_S \), \( M_S \) is an \( R \)-module (multiplication by \( [r/1] \) is a special case of multiplication by \( [r/s] \)). The map \( \mu(m) = [m/1] \) is an \( R \)-module homomorphism from \( M \) to the \( R \)-module \( M_S \). It is not necessarily an embedding (\( \mu \) is an embedding iff \( m = 0 \) whenever \( sm = 0 \) for \( s \in S \) and \( m \in M \), for example if \( M \) is torsion-free).

If \( N \) is a submodule of \( M \) the map taking \( [n/s] \) in \( N_S \) to \( [n/s] \) in \( M_S \) is readily verified (since \( [n/s]_N = [n/s]_{M_S} \cap (N \times S) \)) to be well-defined and an \( R_S \)-monomorphism. In this way \( N_S \) may be considered a submodule of \( M_S \), namely \( [\mu[N]] \), the \( R_S \)-submodule of \( M_S \) generated by \( \mu[N] \), which can be characterized as those classes which contain \( n/s \) for some \( n \in N \) and \( s \in S \).
On the other hand for $K \subseteq M_S$ $\mu^{-1}[K]$ is an $R$-submodule of $M$. The maps $N \mapsto N_S$ and $K \mapsto \mu^{-1}[K]$ have many properties of interest. These are better given after some additional machinery has been developed, and are left to chapter 20. As an example of computing with fractions one fact is proved here, namely $(\mu^{-1}[K])_S = K$. If $[m/s] \in K$ then $[m/1] \in K$, so $m \in \mu^{-1}[K]$, so $[m/s] \in (\mu^{-1}[K])_S$. If $[m/s] \in (\mu^{-1}[K])_S$ then $[m/s] = [m'/s']$ where $[m'/1] \in K$; then $[m'/s'] \in K$, and so $[m/s] \in K$.

Finally, suppose $B \subseteq M$. If $\sum_i r_i b + i = 0$ for $r_i \in R$ and $b_i \in B$ then $\sum_i [r_i/1] [b_i/1] = 0$; thus, if $\mu[B]$ is linearly independent over $R_\mathbb{F}$ then $B$ is linearly independent over $R$. The converse holds if $R$ is an integral domain and $M$ is torsion-free, because if $\sum_i [r_i/s_i] [b_i/t_i] = 0$ then multiplying by $s_1 \cdots s_n t_1 \cdots t_n$ yields a linear combination over $R$ of the $b_i$ which equals 0.

7. Linear algebra of modules. The linearly independent subsets of an $R$-module $M$ form an inductive family of subsets of $M$. If $C$ is a chain, any linear combination of elements of $\bigcup C$ is a linear combination of elements of some member of $C$, and hence if it equals 0 is trivial. By the maximal principle there are maximal linearly independent subsets; indeed any linearly independent subset can be extended to one.

A collection $C$ of finite sets is called a matroid if
1. $C$ is nonempty;
2. $C$ closed under subset, that is, $S \in C$ and $T \subseteq S$ imply $T \in C$;
3. every member of $C$ is contained in a maximal element; and
4. whenever $S, T$ are sets in $C$ and $|S| < |T|$ then for some $t \in T$ $S \cup \{t\} \in C$.

It is a theorem that in a matroid all maximal elements of $C$ have the same cardinality. Indeed, let $S$ be a maximal subset, of size $n$, and let $T$ be a larger set. Then $T$ cannot be in $C$, else $S$ would not be maximal. Hence any maximal subset has size $\leq n$, and since $S$ was arbitrary all maximal subsets have the same size. The requirements 1-4 can be relaxed slightly; if $C$ is a collection of set having properties 1-3, and property 4 for finite sets in $C$, and if $C$ has a finite maximal element, then all maximal elements have the same finite cardinality. The proof is identical.

LEMMA 3. Suppose $R$ is an integral domain and $M$ is an $R$-module. Let $S \subseteq M$ be a linearly independent subset of size $k > 0$, and let $T$ be a subset of size $k + 1$. If $S \cup \{t\}$ is linearly dependent for each $t \in T$ then $T$ is linearly dependent.

PROOF: By induction on $k$. Let $S = \{s_1, \ldots, s_k\}$, $T = \{t_1, \ldots, t_{k+1}\}$. If $k = 1$ then $a_1 t_1 + b_1 s_1 = 0$, $a_2 t_2 + b_2 s_1 = 0$ for some $a_i, b_i \in R$, and $a_1 b_{i_1} t_1 - a_2 b_{i_2} t_2 = 0$. For $k > 1$, by hypothesis $a_i t_i + \sum b_{ij} s_j = 0$, $1 \leq i \leq k + 1$, for some $a_i, b_{ij} \in R$. Certainly $a_i \neq 0$, so some $b_{ij}$ must be nonzero. Let $j$ be such that $b_{k+1,j} \neq 0$; by multiplying equations $i$ and $k + 1$ by appropriate constants and adding we obtain $(a_i' t_i + c_i t_{k+1}) + u_i = 0$, $1 \leq i \leq k$, where $u_i \in \text{Span}(S - \{s_j\})$. By induction the elements $a_i' t_i + c_i t_{k+1}$ are linearly dependent, which shows that $T$ is.

By lemma 3, if $R$ is an integral domain and there is a finite maximal linearly independent subset, of size $n$, then the linearly independent subsets form a matroid, and the maximal ones all have size $n$. This will be the case if $M$ is finitely generated, since an infinite linearly independent subset could not be finitely generated, by lemma 3. The case when $R$ is a field is particularly simple; a module over a field is called a vector space.

THEOREM 4. A maximal linearly independent subset of a vector space is a basis, and conversely. In particular, a vector space $V$ over a field $F$ is a free module; any linearly independent set is contained in a basis; and if $V$ is finitely generated its bases all have the same finite size $n$.

PROOF: If $S$ is a maximal linearly independent subset and $x \notin S$ then there is a linear combination $a x + t$ where $a \in F, a \neq 0$ and $t \in \text{Span}(S)$; but then since $F$ is a field, $x = -t/a \in \text{Span}(S)$. If $S$ is a basis it
is a linearly independent set, and cannot be enlarged since any other element is a linear combination. The remaining claims are immediate from the preceding discussion.

The size of a basis for a finite dimensional vector space $V$ called its dimension, denoted $\dim(M)$, or $\dim_k(M)$ when the field must be specified. In the infinite case, any two bases have the same cardinality (recall that this is so iff they can be put in one-to-one correspondence). The dimension can in fact be defined, and is an infinite cardinal number.

The proof that two infinite bases $B$ and $C$ have the same cardinality is as follows. For each $c \in C$ there is a finite set $S_c$ of elements of $B$, such that $c$ is a linear combination of the elements of $S_c$. The union of the $S_c$ must be all of $B$, else a subset of $B$ would generate all of $C$ and hence the entire vector space, which is impossible since any element of $B$ is not a linear combination of other elements of $B$. The cardinality of $\cup c \in C S_c$ is no greater than the cardinality of $C$ by set theory (see appendix 1). Symmetrically, $|C| \leq |B|$ as well.

In a vector space $V$, any linearly independent subset can be extended to a basis. Given a subspace $W$ of the space $V$, not necessarily finite dimensional, let $S$ be a basis of $W$, and let $T$ be a set of elements such that $S \cup T$ is a basis for $V$. $T$ generates a subspace $U$ such that $V$ is the direct sum of $U$ and $W$. In particular $\dim(W) \leq \dim(V)$.

**Theorem 5.** If $M$ is a free $R$-module where $R$ is commutative, its bases all have the same cardinality.

**Proof:** Let $K$ be a maximal ideal of $R$, so that $R/K$ is a field. Let $S$ be a basis for $M$, and consider the free $R/K$-module over with basis $S$; this is a vector space, which we denote $M/K$. The canonical epimorphism $h : R \rightarrow R/K$ can be extended to an epimorphism from $M$ to $M/K$, which we also denote as $h$, by defining $h(a_1 x_1 + \cdots + a_k x_k) = h(a_1)x_1 + \cdots + h(a_k)x_k$. This map is a commutative group epimorphism and satisfies $h(ax) = h(a)h(x)$. If $T$ is a generating set of $M$ then $h(T)$ is a generating set of $M/K$; any $x' \in M/K$ equals $h(x)$ for some $x \in M$, and the linear combination of elements of $T$ yielding $x$ maps to a linear combination of elements of $h[T]$ yielding $x'$. Thus, $|T| \geq |h(T)| \geq |S|$.

In this case as well we define the dimension of $M$ to be the size of a basis. A basis is certainly a maximal linearly independent set, but a maximal independent set need not be a basis. In the case of finite dimension, a maximal linearly independent set is the same size as a basis.

### 8. Modules over principal ideal domains.

**Theorem 6.** Suppose $R$ is a principal ideal domain, and $M$ is a free $R$-module. If $N$ is a submodule of $M$ then $N$ is free, and $\dim(N) \leq \dim(M)$.

**Proof:** We give the proof only for $M$ finitely generated; the proof in general requires transfinite induction and is given in appendix 1. Let $\{x_1, \ldots, x_n\}$ be a basis for $M$, let $M_i = \text{Span}(x_1, \ldots, x_i)$, and let $N_i = N \cap M_i$. We will construct inductively a basis $S_i$ for $N_i$, of size $\leq i$. Let $S_0 = \emptyset$. For $i > 0$ let $I = \{a \in R : \exists u \in M_{i-1}(ax_i + u \in N_i)\}$. If $I$ is trivial let $S_i = S_{i-1}$. Otherwise $I$ equals $bR$ for some nonzero $b \in R$, and there is a $v \in M_{i-1}$ with $bx_i + v \in N_i$; let $S_i = S_{i-1} \cup \{bx_i + v\}$. Clearly $\text{Span}(S_i) \subseteq N_i$. If $x \in N_i$ then $x = u + ax_i$ for some $u \in M_{i-1}$ and $a \in R$. Further, $a$ is divisible by $b$, and $x \in \text{Span}(S_i)$ follows. Finally, $x_i$ is not in $\text{Span}(S_{i-1})$, so $S_i$ is linearly independent.

In the case of $\mathbb{Z}$ theorem 6 states that a submodule of $\mathbb{Z}^n$ consists of the integer linear combinations of some linearly independent set of elements. These elements are in fact linearly independent over $\mathbb{Q}$; this
follows by remarks at the end of section 6. It is also true that a set of elements in \( \mathbb{Q}^n \) is linearly independent over \( \mathbb{Q} \) iff it is linearly independent over \( \mathbb{R} \); Gaussian elimination (see appendix 2) provides a proof of this.

There is another important elementary theorem which has a generalization to the context of modules over principal ideal domains, namely the decomposition theorem for finitely generated commutative groups. Since the more general theorem is little more difficult, we will give it.

If \( R \) is a commutative ring a homomorphism of \( R \)-modules \( h : M \rightarrow N \) is called a split epimorphism if it has a right inverse which is a homomorphism. It is easily see that this is so iff \( \text{Ker}(h) \) is a summand, i.e., iff there is a submodule \( L \) of \( M \) which is a system of coset representatives of \( \text{Ker}(h) \).

**Lemma 7.** Suppose \( R \) is a commutative ring. If \( h : M \rightarrow N \) is an epimorphism of \( R \)-modules and \( N \) is free then \( h \) splits.

**Proof:** Choose a basis \( B \) for \( N \), and choose an element in each \( h^{-1}(x) \), \( x \in B \), yielding a set \( C \). Let \( L \) be the submodule of \( T \) generated by \( C \). \( L \) has exactly one element in each coset of \( \text{Ker}(h) \); indeed a linear combination from \( C \) is in the coset given by the corresponding linear combination from \( B \).

If \( M \) is free and \( M = N \oplus L \) then \( N, L \) are free, and the union of bases for \( N, L \) is a basis for \( M \). From all this, it follows that if \( M \) is free and finite dimensional, \( N \) is free, and \( h : M \rightarrow N \) is an epimorphism, then \( \text{Dim}(M) = \text{Dim}(\text{Ker}(h)) + \text{Dim}(N) \).

**Lemma 8.** Suppose \( R \) is an integral domain. If \( M \) is an \( R \)-module then \( M/\text{Tor}(M) \) is torsion-free.

**Proof:** Let \( h \) be the canonical epimorphism and suppose \( ah(x) = 0 \), \( a \in R, a \neq 0 \); then \( h(ax) = 0 \), so \( ax \in \text{Tor}(M) \), so \( bax = 0 \) for some \( b \in R \), so \( x \in \text{Tor}(M) \), so \( h(x) = 0 \).

Finally, if \( M \) is a finitely generated torsion module over a commutative ring then \( \text{Ann}(M) \neq 0 \), (that is, there is a nonzero \( a \in R \) with \( aM = \{0\} \)). To see this, let \( S \) be a finite generating set; for each \( s \in S \) there is a nonzero \( a_s \in R \) with \( a_s s = 0 \). Let \( a \) be the product of these. For the next theorem, recall the notation \( M_r \) for \( \psi^{-1}r[0] \) mentioned in section 1.

**Theorem 9.** Let \( R \) be a principal ideal domain and \( M \) a finitely generated \( R \)-module. Then \( M \) is the direct sum of \( \text{Tor}(M) \) and a free submodule. If \( M \) is isomorphic to \( K \oplus L \) where \( K \) is a torsion module and \( L \) is torsion-free, then \( K \) is isomorphic to \( \text{Tor}(M) \).

**Proof:** First, \( M/\text{Tor}(M) \) is finitely generated, in fact Noetherian. We claim that a finitely generated torsion-free \( R \)-module is free. Granting this, by lemma 7 the canonical epimorphism to \( M/\text{Tor}(M) \) splits. The last claim follows because \( K \) is the torsion submodule of \( K \times L \). To prove the claim, let \( S \) be a finite generating set and \( T \) a maximal linearly independent subset of \( S \). For each \( s \in S \) there is an \( a_s \) such that \( a_s s \in \text{Span}(T) \); let \( a \) be the product of the \( a_s \). The map \( x \mapsto ax \) is injective since \( M \) is torsion-free; the image of \( M \) under this map is a submodule of \( \text{Span}(T) \), which is free; hence \( M \) is free, by theorem 6.

When \( R = \mathbb{Z} \) the theorem states that a finitely generated commutative group is the direct sum of a finite number of infinite cyclic groups, and a finite commutative group. The generalization of a finite commutative group is a finitely generated torsion module.

**Theorem 10.** Suppose \( R \) is a principal ideal domain, \( M \) is an \( R \)-module, and \( aM = 0 \). If \( a = bc \) where \( \gcd(b, c) = 1 \) then \( M = M_b \oplus M_c \).

**Proof:** Let \( 1 = b_1 b + c_1 c \); then for \( x \in M \), \( x = b_1 bx + c_1 cx \), and \( bx \in M_c, cx \in M_b \). Further, if \( x \in M_b \cap M_c \) then \( x = 1x = b_1 bx + c_1 cx = 0 \).

It remains to determine the structure of \( M \) when \( \text{Ann}M = \{0\} \) and \( M \) is finitely generated. For \( x \in M \) the annihilator of \( Rx \) is some (principal) ideal \( Ra, a \in R \). The element \( a \) is called the period of \( x \), or of \( Rx \);
it is determined up to multiplication by a unit. For example, if $R = \mathbb{Z}$ the period is just the order of $x$, or equivalently of the cyclic group it generates.

**Theorem 11.** Suppose $R$ is a principal ideal domain, $p$ is a prime element of $R$, $M$ is a finitely generated $R$-module, and $p^rM = 0$. Then $M = \bigoplus_{i=1}^n M_i$ where $M_i$ is a principal submodule of period $p^{r_i}$ for some $r_i \leq r$. The number $n$ and the sequence $r_1 \geq \cdots \geq r_n$ are determined.

**Proof:** First note that if $p^rN = 0$ for submodules or quotient modules $N$ of $M$. Also the period of any nonzero $x \in M$ is $p^s$ for some $s \leq r$, since if $a$ is the period then $p^s \in aR$, or $a|p^s$. Define a set $\{x_i : i \leq k\}$ of elements of $M$ to be weakly independent if whenever $a_1x_1 + \cdots + a_kx_k = 0$, $a_i \in R$, then $a_ix_i = 0$, all $i$. Clearly the module generated by $\{x_i\}$ is the direct sum of the $Rx_i$ iff the $x_i$ are weakly independent. We prove the theorem by induction on the size $k$ of a generating set; the basis is $k = 0$, the trivial module. Let $\{x_i : i \leq k\}$ be a generating set; we may assume $r$ is as small as possible, and that the period of $x_1$ is $p^r$. Let $K = Rx_1$; clearly $M/K$ is generated by $\{x_i + K : 2 \leq i \leq k\}$, so inductively $M/K$ has a weakly independent generating set $\{y_1, \ldots, y_t\}$. We next claim that for any $w \in M$ there is a $y \in w + K$ with the same period as $w$, and $\bar{w}$ is determined up to multiplication by a unit. For example, if $R = \mathbb{Z}$, while the second has elements of order 6 but none of order 12. As we will see in chapter 10, corollary 12 has another very important use in basic algebra.

**Corollary 13.** If $M$ is a finitely generated torsion module over a principal ideal domain $R$, then $M$ is the direct sum of principal submodules $M_{ij}$, $1 \leq i \leq s$, $1 \leq j \leq rs$. The period of $M_{ij}$ is $p_i^{e_{ij}}$ for some prime element $p_i$ and positive integer $e_{ij}$. The $p_i^{e_{ij}}$, which are called the elementary divisors, are unique up to rearrangement.

**Proof:** Suppose $aM = 0$, $a \neq 0$; $a$ has a factorization $p_1^{e_1} \cdots p_k^{e_k}$ where $p_i$ is prime element. By theorem 11 $M$ is the direct sum of the $M_i = M_i^{(e_i)}$. Further $p_iM_i = 0$, and $M_i$ is finitely generated ($M$ is Noetherian since it is finitely generated over a Noetherian ring). It is now an immediate consequence of theorem 11 that the $M_{ij}$ and $p_i^{e_{ij}}$ exist. Let $e_i$ be the largest of the $e_{ij}$, and let $a$ be the product of the $p_i^{e_i}$; then $aM = 0$, so in any other set $q_i^{e_{ij}}$, $q_i|a$ and $q_i = p_j$ for some $j$. By symmetry the $p_i$ and the $q_i$ are the same. Similarly the $e_i$ are the same, so the $M_i$ are, and uniqueness follows by the uniqueness claim of theorem 11.

**Corollary 14.** A finite commutative group is the direct sum of cyclic subgroups of prime power order; the orders of the subgroups are uniquely determined.

**Proof:** This follows immediately, since a finite commutative group is the same thing as a finitely generated torsion module over $\mathbb{Z}$, and the submodules are the subgroups. Corollary 13 may be proved directly (see for example [Hall]). For an example, there are two commutative groups of order 12, $\mathbb{Z}_4 \times \mathbb{Z}_3$ and $\mathbb{Z}_2^2 \times \mathbb{Z}_3$. The direct product of cyclic groups of relatively prime order is cyclic (see exercise 6), so the first is the same as $\mathbb{Z}_{12}$, while the second has elements of order 6 but none of order 12. As we will see in chapter 10, corollary 12 has another very important use in basic algebra.

**Exercises.**

1. Show that $(M_1 \times M_2)/M_1$ is isomorphic to $M_2$, where $M_1, M_2$ are $R$-modules.
2. Show that the $R$-module homomorphisms between two $R$-modules are closed under pointwise addition. Show that they are closed under the action of $R$ if $R$ is commutative.

3. Suppose $M$ is an $R$-module and $N$ is a submodule; show the following.
   a. If $M$ is Noetherian so are $N$ and $M/N$. Hint: That $N$ is Noetherian is trivial. For $M/N$, consider the inverse image of an ascending chain under the canonical epimorphism.
   b. If $N$ and $M/N$ are Noetherian so is $M$. Hint: A chain $K_0 \subseteq K_1 \subseteq \cdots$ in $M$ yields the chains $K_0 \cap N \subseteq K_1 \cap N \subseteq \cdots$ in $N$ and $h(K_0) \subseteq h(K_1) \subseteq \cdots$ in $M/N$, where $h$ is the canonical epimorphism. These must eventually become constant, so it suffices to show that a submodule $K$ is uniquely determined by $K \cap N$ and $K + N$ (in any module).
   c. Show that if $M_1, \ldots, M_k$ are Noetherian so is $M_1 \times \cdots \times M_k$. Hint: Use part b.

4. Give an example of a torsion-free $\mathbb{Z}$-module which is not free.

5. Complete the proof of theorem 6.

6. Show that, in a commutative group, $o(xy) = \text{lcm}(o(x)o(y))$. 

1. Basic definitions. If $F \subseteq E$ where $F$ and $E$ are fields, then $F$ is of course called a subfield of $E$. A field $F$ contains a smallest subfield. Indeed, if $F$ is of characteristic 0 the smallest subfield is $\mathbb{Q}$; and if $F$ is of characteristic $p$ for some prime $p$ the smallest subfield is $\mathbb{Z}_p$. This subfield is called the prime subring; it equals the prime subring if the characteristic is nonzero, else its field of fractions.

If $F \subseteq E$ is a subfield of $E$, we also say that $E$ is an extension of $F$. An extension of fields may be considered as the pair $F \subseteq E$, or $E \supseteq F$; we may say “the extension $E \supseteq F$”, or “the extension $E$” if $F$ is understood. A sequence of extensions $F_1 \subseteq \cdots \subseteq F_k$ is called a tower; note that each $F_i \subseteq F_j$, $i < j$, is an extension.

An extension $E \supseteq F$ is a vector space over $F$, with addition the addition of $E$ and $cx$ for $c \in F$, $x \in E$ being simply $cx$ in $E$. $E$ is called a finite extension if it has a finite basis; the size of a basis is called the degree of the extension, and denoted $[E : F]$.

Suppose $F_1 \subseteq F_2 \subseteq F_3$ is a tower; we claim that $F_1 \subseteq F_3$ is finite iff $F_1 \subseteq F_2$ and $F_2 \subseteq F_3$ are, in which case $[F_3 : F_1] = [F_3 : F_2][F_2 : F_1]$. If $F_1 \subseteq F_3$ is finite then $F_1 \subseteq F_2$ is, since $F_2$ is a subspace of $F_1$; and $F_2 \subseteq F_3$ is, since a basis over $F_2$ is certainly a generating set over $F_2$. On the other hand, if $u_1, \ldots, u_l$ is a basis for $F_2$ over $F_1$, and $v_1, \ldots, v_m$ is a basis of $F_3$ over $F_2$, we claim that $\{u_1v_j : 1 \leq i \leq l, 1 \leq j \leq m\}$ is a basis for $F_3$ over $F_1$. Certainly every element in $F_3$ can be written as a linear combination over $F_1$ of the $u_iv_j$; first write it as a linear combination over $F_2$ of the $v_i$, and then write the coefficients as linear combinations over $F_1$ of the $u_i$. If $a_{i1}v_1 + \cdots + a_{im}v_m = 0$, we have $b_1v_1 + \cdots + b_nv_m = 0$ where $b_j = a_{ij}u_1 + \cdots + a_{lj}u_l$. The $b_i$ must all be 0; from this it follows that the $a_{ij}$ are all 0.

2. Algebraic extensions. In $R[x]$, $R$ a commutative ring, the notation $(p(x))$ is frequently used to denote the ideal generated by $p(x)$. Recall that if $E \supseteq F$ is an extension of fields and $a \in E$, the map $p(x) \mapsto p(a)$ from $F[x]$ to $E$ is a homomorphism, called the evaluation homomorphism. If its kernel is trivial, $F[a]$ is isomorphic to $F[x]$, and $a$ is called transcendental over $F$. Otherwise, the kernel is a nontrivial ideal $I$, and $a$ is called algebraic over $F$.

For a field $F$, $F[x]$ is a principal ideal domain and $F[a]$, being a subring of the field $E$, is an integral domain. Hence when $a$ is algebraic $I = (p(x))$ for some polynomial $p(x)$ of degree at least 1, and $I$ is prime, so $p(x)$ is irreducible. Hence $F[x]/(p(x))$ is a field, so $F[a]$ is, since it is isomorphic. Since the generator of a principal ideal in a commutative ring is unique up to a unit, the polynomial $p(x)$ is unique up to multiplication by nonzero members of $F$. It is often required to be monic, and called the irreducible polynomial of $a$.

If the irreducible polynomial has degree $n$ then $\{a^i : 0 \leq i < n\}$ form a basis for the extension $F[a]$ of $F$. These generate $F[a]$, since $a^n$ and hence $a^i$ for $i \geq n$ is a linear combination. Further they are linearly independent, else $q(a)$ would equal 0 for some polynomial $q$ of degree less than $n$, which could not be in $(p(x))$. The degree of the irreducible polynomial for $a$ thus equals the degree $[F[a] : F]$ of the extension; this number is also called the degree of $a$.

The notation $F(a)$ is used for the smallest field containing $F$ and $a$ (in some field containing $F$ and $a$). This equals $F[a]$ iff $a$ is algebraic. In one direction, we have just seen that if $a$ is algebraic then $F[a]$ is a field, and clearly any field containing $F$ and $a$ must contain it. Conversely if $F(a) = F[a]$ then $a^{-1} = p(a)$ for some $p \in F[x]$, whence $a$ is algebraic. If $a$ is transcendental over $F$ then $F(a)$ is isomorphic to $F(x)$, the field of rational functions over $F$, which is by definition the field of fractions of $F[x]$. Indeed, the map taking $p/q$ to $p(a)/q(a)$ is an isomorphism from $F(x)$ to $F(a)$.

If $F \subseteq E$ is finite and $a \in E$ then $F \subseteq F[a] \subseteq E$, so $F \subseteq F[a]$ is finite and $a$ is algebraic. Call an extension $E \supseteq F$ algebraic if every element of $E$ is algebraic over $F$; we have just shown that finite extensions are algebraic. Suppose $F_1 \subseteq F_2 \subseteq F_3$: we claim that $F_1 \subseteq F_3$ is algebraic iff both $F_1 \subseteq F_2$ and $F_2 \subseteq F_3$. If $a \in F_3$, we have just shown that $F_1 \subseteq F_3$ is algebraic iff both $F_1 \subseteq F_2$ and $F_2 \subseteq F_3$.
are. Certainly an element algebraic over \( F_1 \) is algebraic over \( F_2 \); indeed its irreducible polynomial over \( F_2 \) is a factor of its irreducible polynomial over \( F_1 \). One direction follows; for the other, if \( a \) is algebraic over \( F_2 \) and \( p \) is its irreducible polynomial, the field \( F' \) generated by the coefficients of \( p \) is a finite extension of \( F \) (see exercise 1). \( F_2[a] \) is a finite extension of \( F' \), so \( a \) is algebraic over \( F_1 \).

Recall from chapter 7 that a field \( F \) is said to be algebraically closed if every \( p \in F[x] \) splits (has all its roots) in \( F \). If \( F \) is an algebraically closed field and \( E \supseteq F \) is a finite extension then \( E = F \), because if \( x \in E \) then \( x \) is a root of a polynomial over \( F \). It follows that if \( E \) is an algebraic extension then \( E = F \). Conversely if \( E \) has no proper finite extensions then \( E \) is algebraically closed, since given any polynomial there is a finite extension in which it has a root.

For the remainder of this chapter, all extensions will be algebraic.

3. Splitting fields. Let \( F \) be a field and \( F[x] \) the polynomial ring over \( F \). \( F \) is a subring of \( R \) in a canonical manner, namely the constant polynomials. An ideal \( I \subseteq F[x] \) is trivial iff it contains a constant polynomial, as is easily seen. From this, if \( I \) is a proper ideal, the canonical homomorphism is injective on \( F \), so that \( F \) may be considered a subring of \( F[x]/I \) in a canonical manner.

If \( p(x) \) is irreducible over \( F \) then it has a root in \( E = F[x]/(p(x)) \), namely \( h(x) \) where \( h \) is the canonical epimorphism, since \( p(h(x)) = h(p(x)) = 0 \). In the above manner \( E \) may be considered an extension of \( F \). If \( p \) does not factor to linear factors in \( E \), it may be extended to \( E[x]/(q(x)) \) where \( q \) is some irreducible factor of \( p \) of degree greater than 1. This step may be repeated, eventually yielding an extension \( K \) of \( F \) in which \( p \) factors into linear factors, where further \( [K : F] \leq n! \) where \( n = \deg(p) \). Recall that a polynomial \( p \in F[x] \) is said to split in a field \( E \supseteq F \) if it factors into linear factors; \( E \) is called a splitting field for \( p \) if \( p \) splits in \( E \), and \( E \) is generated by the roots of \( p \).

Suppose \( F \) and \( K \) are fields and \( \sigma : F \rightarrow K \) is an embedding; we will use the exponential notation \( a^\sigma \) for \( \sigma(a) \), and \( p^\sigma \) for the polynomial \( a_0^\sigma + \cdots + a_n^\sigma x^n \) where \( p \) equals \( a_0 + \cdots + a_n x^n \).

Suppose \( a \) is algebraic over \( F \), with irreducible polynomial \( p \); and \( p \) has some root \( a' \) in \( K \). There is a unique extension of \( \sigma \) to (any copy of) \( F[a] \) which maps \( a \) to \( a' \) and embeds \( F[a] \) in \( K \). Indeed, there is an “evaluation map” taking \( q \in F[x] \) to \( q^\sigma(a) \) in \( K \); and it maps \( p \) to 0, so induces a homomorphism on \( F[x]/(p) \cong F[a] \); further, the isomorphism maps \( x + (p) \) to \( a \). Clearly, this homomorphism is uniquely determined by the given requirements.

**Theorem 1.** Suppose \( E \) is a splitting field of the irreducible polynomial \( p \in F[x] \). Suppose \( K \) is a field, \( \sigma \) is an embedding of \( F \) in \( N \), and \( p^\sigma \) splits in \( K \). Then there is an extension of \( \sigma \) to an embedding of \( E \); further for any two such the image is the same.

**Proof:** Given a root \( a \) of \( p \) in \( E \) and a root \( a' \) of \( p^\sigma \) in \( K \), there is an extension of \( \sigma \) to \( F[a] \) taking \( a \) to \( a' \). If this is not yet an embedding of \( E \), the process may be continued over \( F[a] \) with an irreducible factor of \( p \). For the last claim, if \( \sigma \) denotes also the extension, \( S \) is the roots of \( p \) in \( E \), then \( S^\sigma \) is the roots of \( p^\sigma \) in \( K \) and generates \( E^\sigma \).

In particular the splitting field is unique up to isomorphism. In an extension of \( F \) containing all the roots of an irreducible polynomial \( p(x) \), the roots are said to be conjugate, over \( F \). Let \( \operatorname{Aut}_F(E) \), for an extension \( E \supseteq F \), be the automorphisms of \( E \) which fix \( F \). The proof of the theorem shows that if \( p(x) \) is irreducible over \( F \) and splits in \( E \) then \( \operatorname{Aut}_F(E) \) acts transitively on the roots, so that the conjugates of a root \( a \) are exactly the orbit of \( a \) under \( \operatorname{Aut}_F(E) \).

4. Galois extensions. An extension \( E \supseteq F \) is called normal if any polynomial over \( F \) with a root in \( E \) splits. Equivalently, if \( a \in E \) then all \( a \)'s conjugates are in \( E \). If \( E_0 \supseteq F \) is any extension then there is an extension \( E_2 \supseteq E_0 \) such that \( E_2 \) contains the conjugates of the elements of \( E_0 \). This follows by transfinite recursion and the fact that (by theorem 1) the conjugates of a single \( a \in E_0 \) can be added. Details are left
to the reader who has read appendix 1; if $E_0 \supset F$ is finite only a finite induction is required. The subfield $E_1$ of $E_2$ generated by the conjugates of the elements of $E_0$ is the smallest extension of $E_0$ which is a normal extension of $F$ (exercise 2); we call it the normal closure.

Given a tower $F \subseteq E \subseteq K$, an embedding of $E$ into $K$ which leaves each element of $F$ fixed is said to be an embedding over $F$. If $E \supseteq F$ is normal and $K \supseteq E$ is an extension of $E$, any embedding of $E$ in $K$ over $F$ maps $E$ onto itself (indeed maps each class of conjugates onto itself). We claim that conversely, if $E \supseteq F$ has this property then it is normal. To prove this, let $K \supseteq E$ contain the conjugates of the elements of $E$. $\text{Aut}_E(K)$ acts transitively in the conjugates of elements of $E$, and stabilizes $E$ by hypothesis, so $E$ is normal.

An element $a$ is called separable over $F$ if its irreducible polynomial has no multiple roots. Fields $F$ such that any element algebraic over $F$ is separable are called perfect. Many fields of interest are perfect, for example fields of characteristic 0 (exercise 3). Call an extension separable if all its elements are. Exercise 4 shows that if $\alpha \in F$ are distinct, exercise 4 shows that $[F : F]_s := \sum_{\sigma \in \text{Aut}_F(F)} \sigma(\alpha)$.

It follows that $E \supseteq F$ is separable iff $[E : F]_s = [E : F]$. Suppose $E \supseteq F$ is finite and $L \supseteq E$ is the normal closure. By exercise 4, there are at most $[E : F]$ embeddings of $E$ over $F$ in $L$. Since $L$ is the join of the images of these embeddings, $L$ is finite. If $E$ is separable then $L$ is, since the conjugates of members of $E$ are separable and each member of $L$ lies in a field obtained by a tower of extensions by such elements.

**Lemma 2.** If $E \supseteq F$ is finite and separable there is a $c \in E$ such that $E = F[c]$.

**Proof:** It suffices to show the lemma for $E = F[a, b]$, since then we may apply induction. Let $\alpha_1, \ldots, \alpha_n$ be the embeddings over $F$ of $F[a, b]$ into its normal closure. If we can find a $c \in F[a, b]$ such that the $\alpha_i(c)$ are distinct, exercise 4 shows that $[F[c] : F] = n = [F[a, b] : F]$, so $F[c] = F[a, b]$. We assume $F$ is infinite; the case of $F$ finite is proved in section 6. Given $i, j$, at most one $d \in F$ can satisfy $\alpha_i(a) + d\alpha_j(b) = \alpha_j(a) + d\alpha_j(b)$, noting that if $\alpha_i(b) = \alpha_j(b)$ then $\alpha_i(a) \neq \alpha_j(a)$. Hence there is a $d$ satisfying none of them, and we let $c = a + db$.

**Corollary 3.** If $E \supseteq F$ is separable and its elements all have degree $\leq n$ then $E \supseteq F$ is finite and $[E : F] \leq n$.

**Proof:** If not, choose $L$ with $F \subseteq L \subseteq E$ and and $[L : F] > n$, and let $c$ be such that $L = F[c]$; then $c$ has degree greater than $n$, contradicting the hypothesis.

An extension $E \supseteq F$ is called Galois if it is normal and separable. The group $\text{Aut}_F(E)$ is called the Galois group of the extension. The following are readily verified.

- If $F \subseteq E \subseteq K$ and $F \subseteq K$ is Galois, then $E \subseteq K$ is Galois. We have already seen that is is separable.
- It is also normal, since the conjugates of an element of $K$ over $E$ are among those over $F$.
- If $E \supseteq F$ is finite and Galois then $[E : F]$ equals the order of $\text{Aut}_F(E)$. This follows because if $[E : F] = n$ then there are $n$ embeddings of $E$ over $F$ into itself, and these are just the automorphisms.

**5. Galois theory.** Suppose $E \supseteq F$ is an extension and $G \subseteq \text{Aut}_F(E)$ is a group of automorphisms of $E$ over $F$. Let $\text{Fixed}(G)$ be the elements of $E$ which are fixed by every element of $G$; clearly these form a subfield. If the extension is Galois, $\text{Fixed}(\text{Aut}_F(E)) = F$, since any element $a \in E$, $a \notin F$, is moved to its conjugates by $\text{Aut}_F(E)$, and has at least one such.
**Theorem 4 (Fundamental Theorem of Galois Theory).** Suppose \( K \supseteq F \) is a Galois extension, and consider the maps \( E \mapsto \text{Aut}_F(K) \) and \( G \mapsto \text{Fixed}(G) \) between the fields \( E \) with \( F \subseteq E \subseteq K \) and the subgroups \( G \) of \( \text{Aut}_F(K) \).

a. \( E_1 \subseteq E_2 \) implies \( \text{Aut}_{E_2}(K) \supseteq \text{Aut}_{E_1}(K) \); \( G_1 \supseteq G_2 \) implies \( \text{Fixed}(G_1) \subseteq \text{Fixed}(G_2) \); and \( \text{Aut}_E(K) \supseteq G \) if \( E \subseteq \text{Fixed}(G) \).

b. \( E \subseteq \text{Fixed}(\text{Aut}_E(K)) \); \( \text{Aut}_{\text{Fixed}(G)} \supseteq G \); \( \text{Aut}_{\cap_\alpha \text{Aut}_{E_\alpha}(K)} = \cap_\alpha \text{Aut}_{E_\alpha}(K) \); and
\[ \text{Fixed}(\cap_\alpha G_\alpha) = \cap_\alpha \text{Fixed}(G_\alpha). \]

c. In fact, \( \text{Fixed}(\text{Aut}_E(K)) = E \) (so that \( E \mapsto \text{Aut}_E(K) \) is injective), and \( E_1 \subseteq E_2 \) iff \( \text{Aut}_{E_2}(K) \supseteq \text{Aut}_{E_1}(K) \).

d. \( E \) is a normal extension of \( F \) iff \( \text{Aut}_E(K) \) is a normal subgroup of \( \text{Aut}_F(K) \); in this case the map taking \( \alpha \in \text{Aut}_F(K) \) to its restriction to \( E \) is an epimorphism from \( \text{Aut}_F(K) \) to \( \text{Aut}_E(K) \) whose kernel is \( \text{Aut}_E(K) \), and \( \text{Aut}_E(E) \) is isomorphic to \( \text{Aut}_F(E)/\text{Aut}_E(K) \).

e. If \( K \supseteq F \) is finite the map \( E \mapsto \text{Aut}_E(K) \) is bijective.

**Proof:** Parts a and b are left to the reader. Part a states that the two maps form a Galois adjunction (see chapter 13), and part b then follows by partial order theory. For part c, by remarks above \( K \supseteq E \) is Galois so if \( \alpha \notin E \) then \( \alpha \) is moved by some element of \( \text{Aut}_E(K) \). The second claim follows by partial order theory. For part d, if \( E \) is normal then \( E \) is mapped to itself by every element of \( \text{Aut}_F(K) \). The restriction map is a homomorphism, and its kernel is \( \text{Aut}_E(K) \). To see that its image is \( \text{Aut}_F(E) \), given \( \alpha \in \text{Aut}_F(E) \) and \( a \in K \) there is an \( \beta \in \text{Aut}_F(E(a)) \) which fixes \( a \) and extends \( \alpha \); repeating this step, using a transfinite argument in the infinite case, yields an extension of \( \alpha \) to \( K \). If \( E \) is not normal, some \( \alpha \in \text{Aut}_F(K) \) moves \( E \) to a distinct \( E' \). Then \( \text{Aut}_{E'}(E') = \alpha \text{Aut}_F(E) \alpha^{-1} \), and must be distinct from \( \text{Aut}_F(E) \) since \( E, E' \) are distinct. For part e consider \( H \subseteq \text{Aut}_F(K) \) with \( |H| = n \); let \( E = \text{Fixed}(H) \). Given \( a \in K \) let \( S \) be its orbit under \( H \); let \( p(x) = \prod_{s \in S} (x - s) \). Since \( H \) maps \( p(x) \) to itself, its coefficients lie in \( E \), and the degree of \( a \) over \( E \) is at most \( n \). By corollary 3 \( |\text{Aut}_E(K)| = [K : E] \). Since \( H \subseteq \text{Aut}_F(K) \) we must have \( H = \text{Aut}_E(K) \).

In sum, if \( K \supseteq F \) is finite and Galois, the maps \( E \mapsto \text{Aut}_E(K) \) and \( G \mapsto \text{Fixed}(G) \) are inverse.

**6. Finite fields.** A finite field \( F \) is a finite dimensional vector space over the prime subfield, which has order \( p \) for some prime \( p \), so \( F \) has \( q = p^k \) elements for some integer \( k \geq 1 \). This number is called the order of the field. The order of the multiplicative group of nonzero elements of \( F \) is \( q - 1 \), so \( a^q = a \) for all \( a \in F \). \( F \) is thus the splitting field of \( x^q - x \) over the prime field. On the other hand, given \( q \) there is a splitting field \( F \), and if its order is \( q' \) then we cannot have \( q' < q \), else \( x^{q'} - x \) cannot have enough roots in \( F \), noting that \( x^{q'} - x \) has no multiple roots; and we cannot have \( q' > q \), since the roots of \( x^{q'} - x \) form a subfield (exercise 5).

Thus, there is a unique finite field \( F \) of order \( p^k \) for each prime \( p \) and integer \( k > 0 \). This field is denoted by various notations; we use \( F_{p^k} \). If \( E \subseteq F_{p^k} \) is a subfield, of order \( p^l \), then \( p^{ln} = p^k \) where \( n \) is the degree of the extension, so \( l|k \). On the other hand, if \( l|k \) then \( x^{p^l} - x \mid x^{p^k} - x \) (see exercise 7), so the roots of \( x^{p^l} - x \) in a field in which both polynomials split lie in the splitting field of \( x^{p^k} - x \). It follows that \( F_{p^k} \) contains \( F_{p^l} \) as a subfield. The elements of this subfield are the roots of \( x^{p^l} - x \), so it is unique. One can show directly that such elements form a subfield; see exercise 5.

The automorphisms of \( F_{p^k} \) over \( F_p \) are easily determined. The map \( \alpha_j(x) = x^{p^j} \) is such, by exercise 5. The only such map with \( 0 \leq j < k \) which fixes every element is \( \alpha_0 \) since \( x^{p^j} - x = 0 \) has at most \( p^j \) roots. The \( \alpha_j \) thus form a cyclic group of order \( k \), generated by \( \alpha_1 \). This must be the entire group of automorphisms, since any automorphism fixes \( F_p \) and the number of automorphisms cannot exceed the degree of the extension.

65
In particular, the extension \( \mathcal{F}_p \subseteq \mathcal{F}_{p^k} \) is separable. It is also normal, since given any finite extension \( E \) of \( \mathcal{F}_p \), and any automorphism \( \alpha \) of \( E \) over \( \mathcal{F}_p \), if \( \mathcal{F}_{p^k} \subseteq E \) then \( \alpha \) stabilizes \( \mathcal{F}_{p^k} \). It follows that an extension \( \mathcal{F}_{p^l} \subseteq \mathcal{F}_{p^k} \) is separable (and hence finite fields are perfect) and normal. It also follows that the automorphisms of \( \mathcal{F}_{p^k} \) over \( \mathcal{F}_{p^l} \) are the \( \alpha_d \) with \( l | d \), since these are the ones that fix \( \mathcal{F}_{p^l} \).

**Theorem 5.** A finite subgroup \( G \) of the multiplicative group of a field is cyclic.

**Proof:** We may write \( G = G_1 \times \cdots \times G_k \) where \( G_i \) has order \( p_i^{e_i} \) and the \( p_i \) are distinct primes. By exercise 8.6 it suffices to show that each \( G_i \) is cyclic. Let \( a \in G_i \) be of maximal order \( p_i^{e_i} \), so that every \( x \in G_i \) satisfies \( x^{p_i^{e_i}} = 1 \). If \( r < e_i \) then \( x^{p_i^{e_i} - 1} \) has more than \( p_i^{e_i} \) roots in \( F \), a contradiction.

The theorem has various other proofs. One uses exercise 8.6, by finding an element of order \( q_i = p_i^{e_i} \) for each \( i \), where \( n = |G| \) has prime factorization \( q_1 \cdots q_k \). Let \( m = n/p_i \); the polynomial \( x^m - 1 \) has at most \( m \) roots in the field, and since \( m < n \) there is some \( a \in G \) which is not a root. Let \( b = a^{n/q_i} \); then \( b \) has order \( q_i \) in \( G \), as is easily seen. Another uses the fact that, if in a commutative group \( G \) of order \( n \) there are at most \( d \) elements satisfying \( x^d = 1 \) for any \( d | n \) then \( G \) is cyclic. This follows similarly to the theorem, except \( r < e_i \) violates the assumption that there are at most \( p_i^{e_i} \) elements in \( G \) with \( x^{p_i^{e_i}} = 1 \). There are yet further proofs.

**Corollary 6.** The multiplicative group of a finite field \( \mathcal{F}_q \) is cyclic.

The corollary also has various other proofs. For example, observe that, for \( d | q - 1 \) there are exactly \( d \) elements \( x \) of \( \mathcal{F}_q^\times \) such that \( x^d = 1 \). This follows because \( x^{q-1} - 1 \) has \( q - 1 \) roots, \( x^{q-1} - 1 = \sum_{c \neq 1} \psi(c) \) (exercise 7), and the quotient has at most \( q - 1 - d \) roots. Writing \( \psi(c) \) for the number of elements of order \( c \), it follows that for \( d | q - 1 \) \( \sum_{c \neq 1} \psi(c) = d \). By Möbius inversion \( \psi(d) = \phi(d) \) for \( d | q - 1 \).

A generator of the multiplicative group of \( \mathcal{F}_q \) is called a primitive element. If \( q \) is a prime \( p \), \( g \) is a primitive element of \( \mathcal{F}_p \), and \( h \equiv g \mod p \), then \( h \) is called a primitive element, mod \( p \). Since \( (g + p)^i \equiv g^i \mod p \), the powers of \( h \) run through a reduced system of residues mod \( p \).

Corollary 6 provides another proof that \( \mathcal{F}_{p^k} \) has only one subfield of a given order. The multiplicative group of \( \mathcal{F}_{p^k} \) has a subgroup of order \( t \) exactly when \( t = p^l - 1 \) where \( l | k \); as we have seen this is the multiplicative group of a subfield. Finally, lemma 2 for \( F \) finite is a trivial consequence of the corollary.

The existence of \( \mathcal{F}_{p^k} \) shows that there is an irreducible polynomial of degree \( k \) over \( \mathcal{F}_p \), \( p \) prime; in fact we can obtain an expression for the number \( N_{p^k} \) of monic irreducible polynomials \( f(x) \) of degree \( k \) over \( \mathcal{F}_q \).

**Lemma 7.** Suppose \( f \) is an irreducible polynomial of degree \( d \) over \( \mathcal{F}_q \), which is of characteristic \( p \). The roots of \( f \) lie in \( \mathcal{F}_{q^k} \) iff \( d | k \), iff \( f(x) | x^{q^k} - x \). If they do, and \( a \) is one root, the conjugates are \( a^q^j \), \( 0 \leq j < d \).

**Proof:** The roots of \( f \) lie in \( \mathcal{F}_{q^k} \), since this is the unique extension of \( \mathcal{F}_q \) of degree \( d \); thus, they lie in \( \mathcal{F}_{q^k} \) iff \( d | k \). If this is so that \( f(x) | x^{q^k} - x | x^{q^k} - x \); conversely if \( f(x) | x^{q^k} - x \) the roots of \( f \) lie in \( \mathcal{F}_{q^k} \). The last claim follows since this is the orbit of the automorphism group.

Now, \( x^{q^k} - x \) has no multiple roots, and it follows that \( x^{q^k} - x \) is the product of the monic irreducible polynomials over \( \mathcal{F}_q \) whose degree divides \( k \). Thus, \( \sum_{d | n} dN_{q^d} = q^n \), whence by Möbius inversion \( N_{q^k} = (1/k) \sum_{d | k} \mu(k/d)q^d \).

**Exercises.**

1. Show that the subfields of a field \( E \) form a complete lattice. If \( E_1, E_2 \) are two finite extensions of \( F \), show that their join is finite.

2. Suppose \( F \subseteq E_0 \subseteq E_2 \) is a tower where \( E_2 \) contains the conjugates of the elements of \( E_0 \), and \( E_1 \) is the subfield generated by these. Suppose \( K \supseteq F \) is an extension containing the roots of any polynomial over \( F \) with a root in \( E_0 \). Show that any embedding of \( E_1 \) in \( K \) over \( F \) maps to the same subfield of \( K \),
namely that generated by the roots. Show that \( E_1 \) is normal, and so that the splitting field of an irreducible polynomial is normal.

3. Show that if an irreducible polynomial \( f \) has a multiple root then its derivative \( f' \) must be identically 0. This cannot happen if the field has characteristic 0. If the characteristic is \( p \), then the polynomial must be of the form \( f(x^p) \) for some \( p \), and in fact of the form \( f(x^p) \) \( f' \) is not identically 0. Hint: If \( f \) is the irreducible polynomial of \( x \), and \( f'(x) = 0 \), then \( f' \) must be identically 0. If the characteristic is \( p \), only terms \( cx^ip \) can appear, so \( f = f_1(x_p) \), and we may continue inductively if \( f_1'(x) = 0 \).

4. Show the following.

a. Suppose \( F \subseteq E \subseteq K \) and \( F \subseteq K \) is normal. The cardinality of the set of embeddings of \( E \) in \( K \) over \( F \) does not depend on \( K \). Note that the infinite case is included; simply establish a one-to-one correspondence.

b. Denoting the cardinality of part a as \([E : F]_s\), for a tower \( F_1 \subseteq F_2 \subseteq F_3 \), \([F_3 : F_1]_s = [F_3 : F_2]_s[F_2 : F_1]_s\).

Again the infinite case is included; establish a one-to-one correspondence between the embeddings of \( F_3 \) over \( F_1 \), and the pairs of embeddings of \( F_3 \) over \( F_2 \) and \( F_2 \) over \( F_1 \). Hint: Extend each embedding of \( F_2 \) over \( F_1 \) to one of \( F_3 \), and consider the products of the pairs.

c. For finite \( E \subseteq F \), \([E : F]_s[E : F]\); for nonzero characteristic the quotient is a power of the characteristic \( p \). Hint: Show this first for \( E = F[a] \), using exercise 2.

d. In a tower \( F_1 \subseteq F_2 \subseteq F_3 \) of finite extensions \( F_1 \subseteq F_3 \) is separable iff \( F_1 \subseteq F_2 \) and \( F_2 \subseteq F_3 \) are.

e. Remove the finiteness condition from part d. Hint: Let \( a \in F_3 \); let \( F'_2 \) be the subfield of \( F_2 \) generated by the coefficients of the irreducible polynomial of \( a \) over \( F_2 \). Replace \( F_2 \) by \( F'_2 \) and \( F_3 \) by \( F'_2(a) \).

5. Let \( p \) be a prime and \( F \) a field of characteristic \( p \).

a. Show that for \( 0 < k < p \), \( p \left( \binom{k}{2} \right) \).

b. Show that if \( a, b \in F \) and \( 1 \leq t \), \( (a + b)^t = a^t + b^t \).

6. Verify the claims regarding the algebra of functions from an order \( \langle S, \leq \rangle \) to a field \( F \). Hint: For the inverse, show that left and right inverses exist. Verify the claims when the order is \( \langle N, \mid \rangle \).

7. Show that in \( F[x] x^m - 1 | x^n - 1 \) iff \( m | n \). Hint: Give an explicit expression for the quotient when \( m | n \), and write \( n = km + r, x^{km+r} - 1 = x^r(x^{km} - 1) + (x^r - 1) \) for the converse.

8. Show the following. Let \( F(x_1, \ldots, x_n) \) be the field of rational functions over \( F \), i.e., the field of fractions of \( F[x_1, \ldots, x_n] \). The group Sym \(_n\) acts on this by permuting the variables; these maps are automorphisms of the field. \( F(x_1, \ldots, x_n) \) is the splitting field over \( F(\sigma^n_1, \ldots, \sigma^n_n) \) (where \( \sigma^n_i \) denotes the elementary symmetric function) of \( x^n - \sigma^n_1x^{n-1} - \cdots + (-1)^n\sigma^n_n \), \( F(\sigma^n_1, \ldots, \sigma^n_n) \) is the fixed field of Sym \(_n\), and \( [F(x_1, \ldots, x_n) : F(\sigma^n_1, \ldots, \sigma^n_n)] = n! \).

9. Show that for any prime \( p \) and \( k \geq 1 \) there is a polynomial \( f(x) \in \mathbb{F}_p[x] \), irreducible of degree \( k \), such that \( x \) is the generator of the multiplicative group of nonzero elements of \( \mathbb{F}_p/(f(x)) \). Such a polynomial is called primitive. Hint: Consider the irreducible polynomial of a primitive element.
10. Linear algebra.

1. Basic definitions. The basic linear algebra of vector spaces is covered in appendix 2 for readers unfamiliar with it. In this section, all proofs will be left to the reader; many can be found in appendix 2. Many basic facts of linear algebra carry over to the context of finite dimensional free modules over a commutative ring \( R \). In particular, the \( m \times n \) matrices form an \( mn \)-dimensional free \( R \)-module with componentwise addition and scalar multiplication. The product \( C = AB \) is defined if \( A \) is \( m \times n \) and \( B \) \( n \times p \); \( C \) is \( m \times p \), and

\[
C_{ij} = \sum_k A_{ik}B_{kj}.
\]

where \( A_{ij} \) denote the entry in row \( i \) and column \( j \) of the matrix \( A \), etc. The \( n \times n \) matrices form an \( R \)-algebra with this operation. The multiplicative identity has 1’s on the diagonal and 0’s elsewhere, and is called the identity matrix.

There is an \( R \)-module isomorphism between the \( R \)-linear transformations \( L_R(\mathbb{R}^n; \mathbb{R}^m) \) and the \( m \times n \) matrices with entries from \( R \). One such is obtained by identifying \( R^m \) with the column vectors; the columns of the matrix corresponding to a map are then the images of the standard unit vectors. If \( m = n \) the isomorphism is an \( R \)-algebra isomorphism. If \( x \) is a column vector the matrix product \( Ax \) gives the result of applying the map \( A \) to the vector \( x \).

Some facts hold even when \( R \) is not commutative. The matrices are a left \( R \)-module with componentwise addition and left scalar multiplication. The definition of matrix multiplication is as before. In particular the \( A_{ik} \) might be \( m_i \times p_k \) matrices, and the \( B_{kj} \) \( p_k \times n_j \) matrices, with entries \( C_{ij} \) being \( m_i \times n_j \) matrices. One easily checks that the same result for the product is obtained, whether the matrices are considered to consist of submatrices or ring elements. For the rest of the chapter, \( R \) is commutative.

The transpose of an \( m \times n \) matrix \( A \) is the \( n \times m \) matrix \( A^t \) where \( A^t_{ij} = A_{ji} \). The transpose satisfies \((AB)^t = B^tA^t\) and \((A^t)^t = A \). If \( M \) is an \( m \times n \) matrix, \( S \subseteq \{1, \ldots, m\} \), and \( T \subseteq \{1, \ldots, n\} \), let \( M_{ST} \) denote the matrix obtained by keeping only the rows of \( S \) and the columns of \( T \), in order. Such a matrix is called a minor of \( M \).

For the following, let \( i \) denote \( \{1, \ldots, m\} \setminus i \). The determinant \( \det(M) \) of an \( n \times n \) matrix \( M \) is defined by

\[
\det(M) = \sum_{\sigma \in S_n} \text{sg} (\sigma) \prod_{i=1}^n M_{i, \sigma(i)},
\]

where \( S_n \) is the symmetric group of permutations of \( \{1, \ldots, n\} \) and \( \text{sg}(\sigma) \) is the sign of sigma. The determinant satisfies \( \det(MN) = \det(M)\det(N) \) and \( \det(I) = 1 \) (and is a homomorphism of the multiplicative monoids). Also,

\[
\det(M) = \sum_{j=1}^n (-1)^{i+j} M_{ij} \det(M_{ij})
\]

for any row \( i \), and

\[
\det(M) = \sum_{i=1}^n (-1)^{i+j} M_{ij} \det(M_{ij})
\]

for any column \( j \); this is called expansion by minors. The adjugate \( A^{\text{adj}} \) of \( A \) is defined by

\[
A^{\text{adj}}_{ij} = (-1)^{i+j} \det(M_{ji}).
\]

Then \( AA^{\text{adj}} = A^{\text{adj}}A = \det(A)I \).

An \( n \times n \) matrix \( A \) is defined to be invertible iff it is invertible in the multiplicative monoid; equivalently if the corresponding linear transformation on \( \mathbb{R}^n \) has a two-sided inverse which is a linear transformation.
By universal algebra this is so iff it has a two-sided inverse, iff it is bijective. By the preceding this is so iff \( \det(A) \) is a unit; indeed \( A^{-1} = \det(A)^{-1} A^{\text{adj}} \), and \( \det(A^{-1}) = \det(A)^{-1} \).

The linear transformation of an \( m \times n \) matrix \( A \) is injective iff the kernel is trivial; \( A \) is said to be nondegenerate (on the right). For \( n \times n \) matrices this is so iff \( \det(A) \) is not a zero divisor, and when \( R \) is an integral domain iff \( \det(A) \neq 0 \). In the latter case the matrix is also called nonsingular. Unlike fields, a square matrix may be nondegenerate without being invertible.

The linear transformation of an \( m \times n \) matrix \( A \) is surjective iff \( A \) has an \( n \times n \) right inverse \( A^R \) (to obtain the columns of \( A^R \) choose linear combinations of the columns of \( A \) equaling the unit vectors in \( R^n \)). If \( A \) is \( n \times n \) and the linear transformation is surjective then it is bijective; this follows because \( \det(A) \) is a unit (\( \det(A^R) \) is its inverse).

A matrix \( M \) is said to be in upper (lower) triangular form if \( M_{ij} = 0 \) for \( j < i \) \((j > i)\), and diagonal if \( M_{ij} = 0 \) for \( i \neq j \). The matrices of any of these forms are easily seen to form a subalgebra, which contains the multiplicative inverse if it exists. Further if \( C = AB \) then \( C_{ii} = A_{ii}B_{ii} \). We call a matrix triangular if it is in one of these forms; for a matrix \( M \) in triangular form \( \det(M) \) is the product of the entries on the diagonal.

An elementary row operation (on an \( m \times n \) matrix) consists of adding \( c \) times row \( s \) to row \( r \). This corresponds to multiplying on the left by an \( m \times m \) matrix which has 1’s on the diagonal, and \( c \) in the \( r, s \) entry and 0’s elsewhere. Multiplying on the right by the similarly defined \( n \times n \) matrix adds \( c \) times column \( s \) to column \( r \); this is called an elementary column operation. As noted in appendix 2, an elementary row operation or an an elementary column operation does not change the determinant of a square matrix.

Multiplying on the left by the matrix which has 1’s on the diagonal, except the \( r,r \) entry is \( c \), multiplies row \( r \) by \( c \); multiplying on the right multiplies column \( r \) by \( c \). These operations multiply the determinant of a square matrix by \( c \).

Again letting \( S_n \) denote the symmetric group of degree \( n \), given \( \sigma \in S_n \) let \( P_\sigma \) denote the matrix where \((P_\sigma)_{ij} = \delta(j, \sigma(i)) \). The map \( \sigma \mapsto P_\sigma \) is a bijection between \( S_n \) and the matrices whose columns (equivalently rows) are the standard unit vectors in some order; such matrices are called permutation matrices. The following are readily verified.

- \( P_{\sigma \tau} = P_\tau P_\sigma \).
- \( P_{\sigma^{-1}} = P_\sigma^{-1} = P_{\sigma^t} \).
- If \( B = P_\sigma A \) then \( B_{\sigma(i)j} = A_{ij} \) (i.e., row \( i \) gets moved to row \( \sigma(i) \)).
- If \( B = A P_\sigma \) then \( B_{i\sigma(j)} = A_{ij} \) (i.e., column \( j \) gets moved to column \( \sigma(j) \)).
- \( \det(P_\sigma) \) equals +1 if \( \sigma \) is even and −1 if \( \sigma \) is odd. It follows that if the rows or columns of a matrix are permuted then the determinant is unchanged if the permutation is even, and multiplied by −1 if the permutation is odd.

The trace \( \text{Tr}(M) \) of an \( n \times n \) matrix \( M \) is defined to be \( \sum_i M_{ii} \), the sum of the elements on the diagonal. The trace obeys the identity \( \text{Tr}(AB) = \text{Tr}(BA) \); indeed, both sides equal \( \sum_{i,k} A_{ik} B_{ki} \). Also, \( \text{Tr}(A^t) = \text{Tr}(A) \).

Suppose \( R \) is a principal ideal domain. Given an \( m \times n \) matrix \( M \), the image and kernel modules of the linear transformation specified by \( M \) are free, and their dimensions sum to \( n \). This follows from results in section 8.8, in particular the remarks following lemma 8.7. The dimension of the image module is called the rank, and that of the kernel module the nullity. For an \( n \times n \) matrix \( M \) the rank is \( n \) iff the matrix is nondegenerate.

Over any commutative ring the image space is the column space of the matrix, i.e., the subspace of \( R^n \) generated by the columns. Over a field the row space has the same dimension. This follows by Gaussian elimination (see appendix 2); another proof is given in section 2. This holds over a principal ideal domain; indeed, over any integral domain the dimension of a free \( R \)-module equals the dimension of the corresponding vector space over the field of fractions.
Suppose \( R \) is a principal ideal domain and \( S \) is a basis of \( R^n \); the matrix \( T \) whose columns are the elements of \( S \) is invertible. A column vector \( x' \) specifies the element \( x = Tx' \) of \( R^n \). More abstractly we may think of \( R^n \) as the representations of the members of an \( n \)-dimensional free module, so that \( x \) represents the element in one chosen basis, and \( x' \) in a second basis, where the elements of the second basis are given in the first basis by the columns of \( T \). If \( M \) is a matrix, representing some linear transformation in the first basis, we claim that \( M' = T^{-1}MT \) is the matrix representing the transformation in the second basis. Indeed \( x = Tx' \), so \( MTx' \) are the coordinates of the transformed vector in the first basis, so \( T^{-1}MTx \) are the coordinates in the second basis.

Over a commutative ring a matrix \( T^{-1}MT \) is said to be similar to \( M \). From \( \det(AB) = \det(A)\det(B) \) it follows that similar matrices have the same determinant. From \( \text{Tr}(AB) = \text{Tr}(BA) \) it follows that they have the same trace. Over a principal ideal domain similar matrices have the same rank; this follows by exercise 4.

A homomorphism of commutative rings \( \phi : R \to S \) induces an \( R \)-module homomorphism from the \( m \times n \) matrices over \( R \) to the \( m \times n \) matrices over \( S \) (where the latter are considered an \( R \)-module in the usual way). Further, \( \phi(AB) = \phi(A)\phi(B), \phi(A^t) = \phi(A)^t, \phi(\det(A)) = \det(\phi(A)) \), and \( \phi(\text{Tr}(A)) = \text{Tr}(\phi(A)) \).

If \( M \) is a finitely generated \( R \)-module it may still be convenient to use matrix notation for an \( R \)-linear transformation on \( M \), even if \( M \) is not free. Given an epimorphism \( \eta : R^n \to M \) and a homomorphism \( \phi : M \to M \), for each unit vector \( e_i \in R^n \) choose a vector \( c_i \in R_m \) so that \( \phi(\eta(e_i)) = \eta(c_i) \) and let \( \Phi(e_i) = c_i \). This yields a matrix \( \Phi : R^n \to R^m \), and by linearity \( \eta\Phi = \phi\eta \).

2. **Dual space.** Suppose \( V \) is a vector space over the field \( F \). The space \( L_F(V; F) \) is called the dual space, and its members are called linear functionals. It is often denoted \( V^* \). Given a basis \( \{e_i : i \in I\} \) for \( V \) let \( f_i \) be the linear functional such that \( f_i(e_j) = \delta(i, j) \). These are linearly independent; if \( \sum_i a_if_i = 0 \), applying the left side to \( e_i \), we get \( a_i = 0 \). The map \( e_i \to f_i \) thus induces an embedding of \( V \) in \( V^* \).

An element \( x \in V \) induces a linear functional \( \phi_x \) on \( V^* \), where \( \phi_x(f) = f(x) \) (the evaluation map). The map \( x \mapsto \phi_x \) is a linear transformation from \( V \) to \( V^{**} \). It is injective; if \( f(x) = 0 \) for all \( f \) then \( x = 0 \) (let \( f \) be the identity). This embedding of \( V \) in \( V^{**} \) may be called canonical; it may be defined without choosing a basis. The embedding of \( V \) in \( V^* \) does depend on the choice of basis.

If \( V \) is finite dimensional, with basis \( e_1, \ldots, e_n \), then the \( f_i \) form a basis for \( V^* \); given \( f \in V^* \), \( f = f(e_1)f_1 + \cdots + f(e_n)f_n \), since both sides have the same values on the \( e_i \). This basis is often called the . Letting \( \dim(V) \) denote the dimension of a vector space over its field, \( \dim(V^*) = \dim(V) \), and the canonical embedding of \( V \) in \( V^{**} \) is an isomorphism. In matrix terms, if \( e_i \) may be taken as the standard unit column vectors and the \( f_i \) as the standard unit row vectors.

Linear functionals arise naturally from a bilinear form, i.e., map \( * \) in \( L_F(W, V; F) \). Given \( w \in W \) let \( \phi_w \) be the map where \( \phi_w(v) = w * v \); this is an element of \( V^* \). Indeed the map \( w \mapsto \phi_w \) is a homomorphism from \( W \) to \( V^* \); let this be denoted \( \alpha_1 \). There is a similarly defined homomorphism \( \alpha_r : V \to W^* \).

Since \( \phi_w \) vanishes on \( \text{Ker}(\alpha_r) \), a map \( \phi'_w \) on \( V/\text{Ker}(\alpha_r) \) is induced. The map \( w \mapsto \phi'_w \) is a homomorphism from \( W \) to \( (V/\text{Ker}(\alpha_r))^* \). Since \( \phi'_w = 0 \) iff \( \phi_w = 0 \), its kernel is \( \text{Ker}(\alpha_1) \), whence an embedding \( \beta_1 : W/\text{Ker}(\alpha_1) \to (V/\text{Ker}(\alpha_r))^* \) is induced. By symmetry there is an embedding \( \beta_r : V/\text{Ker}(\alpha_r) \to (W/\text{Ker}(\alpha_1))^* \). If \( \dim(W/\text{Ker}(\alpha_1)) \) is finite \( \dim(V/\text{Ker}(\alpha_r)) \) is bounded by it, whence \( \beta_1 \) is an isomorphism.

When \( W = F^m \) and \( V = F^n \) the foregoing can be viewed in terms of matrices. The bilinear maps \( * \) in \( L_F(F^m, F^n; F) \) are in one-to-one correspondence with the \( m \times n \) matrices. Given a matrix \( M \), the corresponding bilinear map \( * \) takes \( (w, v) \) to \( w^tMv \) (or \( wMv \) if \( w \) is considered a row vector and \( v \) a column vector). Given a map \( * \), the corresponding matrix has \( M_{ij} = e_i * e_j \), where \( e_i \) are the standard unit vectors. Now, \( w \in \text{Ker}(\alpha_1) \) iff \( w^tM = 0 \), from which \( F^m/\text{Ker}(\alpha_1) \) is isomorphic to the row space; similarly \( F^n/\text{Ker}(\alpha_r) \) is isomorphic to the column space. This provides another proof that these have the same dimension.
Duality for modules is a more complex subject; however if $M$ is a module over a commutative ring $R$, let $M^*$ denote $L_R(M; R)$. With $\ast$ a map in $L_R(N, M; R)$, the maps $\alpha_1, \alpha_r$ are defined as before; and $\beta_1, \beta_r$ are embeddings. If $M$ is free let $e_i$ be a basis and define $f_i$ as above; these are linearly independent. If $M$ is finite dimensional they form a basis; also the canonical map from $M$ to $M^{**}$ is bijective, as is readily seen by a direct argument. If $R$ is a principal ideal domain and $M, N$ are free then $N/\text{Ker}(\alpha_1)$ and $M/\text{Ker}(\alpha_r)$ have the same dimension when one is finite dimensional. The bilinear maps $\ast$ in $L_R(R^m, R^n; R)$ are in one-to-one correspondence with the $m \times n$ matrices over $R$.

3. **Representations.** Suppose $R$ is a ring and $M$ an $R$-module. Let $L$ denote the $R$-module $L_R(M; M)$; as observed in section 8.2, it is a ring with with composition as multiplication. If $R$ is commutative then $L$ is closed under scalar multiplication and is an $R$-algebra; in particular there is a canonical embedding of $R$ in $L$, which maps $r$ to $x \mapsto rx$ (the scalar multiple $r$ of the identity). Further the action of $L$ on $M$ is consistent with that of $R$, in that the action of the map is that same as that of $r$. (In matrix terms, the action of $R$ on $M$ is scalar times vector; that of $L$ on $M$ is matrix times vector; and that of $R$ on $L$ is scalar times matrix).

For the rest of this section $R$ is assumed to be a commutative ring. If $B$ is an $R$-algebra, an $R$-algebra homomorphism $\phi : B \rightarrow L_R(M; M)$ is called a representation of $B$. The terminology derives from the fact that an abstractly given $R$-algebra $B$ is “represented” as an algebra of $R$-linear maps, indeed frequently as an algebra of matrices with entries in $R$. $\phi[B]$ is an $R$-subalgebra $A$ of $L = L_R(M; M)$, and $M$ is an $A$-module. It is also a $B$-module, with the action $\langle b, m \rangle \mapsto \phi(b)m$.

If $G$ is a monoid and $\phi : G \rightarrow L$ is a monoid homomorphism to the multiplicative group, $\phi$ extends canonically to the free $R$-algebra generated by $G$; $\phi$ is called a representation of $G$. Note that $G$ acts on $M$; $M$ is sometimes called an $R, G$-module. In a case of greatest importance, $G$ is a group and $\phi$ is a group representation.

If $M$ is an $S$-module, a submodule $N \subseteq M$ is also said to be $S$-invariant, or simply invariant, especially if $M$ is also an $R$ module for some other ring $R$. An $S$-module $M$ is called simple if it has no invariant submodules other than $\{0\}$ and $M$. In the case of a representation, $N$ is $A$-invariant iff it is $B$-invariant. The action, or the representation, is said to be irreducible, or simple, if $M$ is a simple $A$-module.

If $N_1$ is invariant there may or may not be a second invariant submodule $N_2$ such that $M = N_1 \oplus N_2$. For $M = R^n$ there is a basis in which the matrices of $A$ have the form

\[
\begin{bmatrix}
X_1 & 0 \\
0 & X_2
\end{bmatrix}
\text{ or }
\begin{bmatrix}
X_1 & 0 \\
X & X_2
\end{bmatrix}
\]

according to whether or not $N_2$ exists. Note that $X_1$ acts on $N_1$, and if $N_2$ exists then $X_2$ acts on it.

An $R$-module $M$ is said to be semisimple if it is the direct sum of a family of invariant simple submodules. In the case of a representation, $A$ (or the representation) is said to be completely reducible, or semisimple, if $M$ is a semisimple $A$-module.

If $B$ is any $R$-algebra and $b \in B$ the evaluation map $p(x) \mapsto p(b)$ is an $R$-algebra homomorphism from $R[x]$ to $B$. Its image may be denoted $R[b]$; it is the (commutative) subring of $B$ generated over the homomorphic image $R'$ of $R$ by the element $b$ (in cases where this map is considered, usually the map of $R$ into $B$ is injective). The kernel of the map is the polynomials $p$ such that $p(b) = 0$. As above, $M$ is an $R[b]$-module and an $R[x]$-module.

4. **Rational canonical form.** Suppose $F$ is a field, and let $B$ be a fixed member of $L_F(F^n; F^n)$. The polynomials such that $p(B) = 0$ form a nontrivial ideal in $F[x]$, since $F[B]$ is a subspace of the finite dimensional vector space $L(F^n; F^n)$. They form a principal ideal, generated by a unique monic polynomial which we denote $p_M$; $p_M$ is called the minimal polynomial of $B$. Note that the degree $d$ of $p_M$ equals the dimension of $F[x]/(p_M)$ (the cosets of $1, \ldots, x^{d-1}$ form a basis), and so of $F[B]$ ($1, \ldots, B^{d-1}$ form a basis).
Since $p_M F^n = 0$, $F^n$ is a finitely generated torsion module over the principal ideal domain $F[x]$. By corollary 8.12 there are principal submodules $V_{ij}$ with $F^n = \oplus_{ij} V_{ij}$, where the period of $V_{ij}$ is $p_i^{e_{ij}}$ for some monic irreducible polynomial $p_i$ and positive integer $e_{ij}$. The $V_{ij}$ are a fortiori $F$-submodules, i.e., subspaces. Assuming $e_{i1} \geq \cdots \geq e_{ir}$, the $p_i^{e_{ij}}$ (the elementary divisors) are unique up to rearrangement of the $p_i$.

Since a principal submodule $W = F[x]v$ of $F^n$ is invariant, $W$ is an $F[x]$ module itself. Let $q(x)$ be the period, and suppose it is of degree $m$. We claim that $v, xv, \ldots, x^{m-1}v$ form a basis for $W$. Any $p(x)$ may be written as $d(x)q(x) + r(x)$ where $\deg(r) < m$, and $p(x)v = r(x)v$, so the $x^i v, i < m$, generate $W$. Further they are linearly independent, else there would be a polynomial $q'(x)$ of degree less than $m$ with $q(x)v = 0$.

Letting $n_i = \deg p_i$, it follows that $n = \sum_{i,j} e_{ij} n_i$. The polynomial $\prod_{i,j} p_i^{e_{ij}}$ is called the characteristic polynomial; its degree is $n$. We will denote it as $p_C$. It is easily seen that $p_M = p_1^{e_{11}} \cdots p_r^{e_{rr}}$; indeed the minimal polynomial on the direct sum of submodules is the least common multiple of the minimal polynomials on the submodules. Thus, $p_M|p_C$; in particular $\deg(p_M) \leq n$.

Choosing a basis for each $V_{ij}$, in the basis for $F^n$, which is their union (appropriately ordered) every matrix of $F[B]$ has “blocks” (square submatrices) $B_{ij}$ down the diagonal, where $B_{ij}$ acts on $V_{ij}$. The minimal polynomial of $B_{ij}$ equals the period of $V_{ij}$. It remains to choose a basis for $V_{ij}$.

Given any monic polynomial $q \in F[x]$, say $q(x) = a_0 + \cdots + a_{m-1} x^{m-1} + x^m$, let $C_q$ be the $m \times m$ matrix where (indexing rows and columns from 0) $(C_q)_{i,i+1} = 1$ for $0 \leq i \leq m - 2$, $(C_q)_{m-1,j} = -a_j$ for $0 \leq j \leq m - 1$, and $(C_q)_{ij} = 0$ otherwise; i.e.,

$$C_q = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -a_{m-2} \\ 0 & 0 & \cdots & 1 & -a_{m-1} \end{bmatrix}.$$  

$C_q$ is called the companion matrix of $q$. It is readily verified that $q(C_q) = 0$ (exercise 2).

If $W = F[x]v$ is a principal module, with period $q(x)$ of degree $m$, in the basis $v, xv, \ldots, x^{m-1}v$ the matrix for the action of $x$ is clearly $C_q$. Choosing such a basis for each $V_{ij}$, $B_{ij}$ is the companion matrix of $p_i^{e_{ij}}$. The matrix $B'$ corresponding to the action of $x$ in this basis is called the rational canonical form of the matrix $B$. It is similar to $B$, and is unique up to rearrangement of the $p_i$, given the other choices we have made.

Clearly two matrices with the same rational canonical form (or the same elementary divisors) are similar. The converse is also true. Suppose $B' = T^{-1}BT$; then for any $p \in F[x]$ $p(B') = T^{-1}p(B)T$, as is easily seen. Thus, for any $v \in F^n$ $T(p(B')v) = p(B)Tv$, so $v \mapsto Tv$ is an $F[x]$-module isomorphism from the action induced by $B'$ on $F^n$ to that induced by $B$. Isomorphic modules clearly have the same elementary divisors.

Similar matrices, having the same rational canonical form, have the same minimal and characteristic polynomials. Note that the nullity of a rational canonical form matrix equals the number of elementary divisors having $p_i = x$. In particular a matrix is nonsingular iff its minimal polynomial is not divisible by $x$.

5. Jordan canonical form. The polynomials $p_M$ and $p_C$ have the same irreducible factors, so one splits in $F$ iff the other does. When $p_C$ splits, its roots are called the characteristic roots, or the eigenvalues, of the matrix. In this case, the blocks $B_{ij}$ of the rational canonical form may be put in another form. The elementary divisor belonging to the block will be $(x - \alpha_i)^{e_{ij}}$ for some $\alpha_i \in F$ and positive integer $e_{ij}$.

For this paragraph write $V$ for $V_{ij}$, etc. Let $v$ be such that $V = F[x]v$. The elements $v, (x - \alpha)v, \ldots, (x - \alpha)^{e-1}v$ are linearly independent over $F$, because a linear dependence would yield one between the $x^i v$; hence they form a basis for $V$. Now, $B(B - \alpha)^i = \alpha(B - \alpha)^i + (B - \alpha)^{i+1}$. Thus, in this basis, $B_{ii} = \alpha$, $1 \leq i \leq e$.
\[B_{i+1,i} = 1, \quad 1 \leq i < e, \quad B_{ij} = 0 \text{ otherwise}; \quad \text{i.e.,} \]
\[
B = \begin{bmatrix}
\alpha & 0 & \cdots & 0 & 0 \\
1 & \alpha & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \alpha & 0 \\
0 & 0 & \cdots & 1 & \alpha \\
\end{bmatrix}
\]

Such a matrix is called a Jordan block, and \(B\) is said to be in Jordan canonical form when each block is a Jordan block.

The Jordan canonical form of \(B\) is similar to \(B\), and unique up to ordering the Jordan blocks. Two matrices are similar iff they have the “same” Jordan canonical form. A matrix in Jordan canonical form is upper triangular; \(\det(B)\) is thus the product of the eigenvalues, each taken with its multiplicity (i.e., its multiplicity as a root of \(p_C\), namely \(\sum_j e_{ij}\), which is the number of times it occurs on the diagonal in the Jordan canonical form). Similarly \(\text{Tr}(B)\) is the sum of the eigenvalues.

A matrix can be put in Jordan canonical form iff all its eigenvalues lie in \(F\). When they do not, we may take an extension \(E \supseteq F\); the inclusion map extends to an inclusion of the matrices. It follows that two matrices with entries from \(F\) are similar over \(E\) if they have the same Jordan canonical form. But it then follows that they are similar over \(F\). Say \(B' = T^{-1}BT\) where \(B, B'\) are matrices over \(F\) and \(T\) is a matrix over \(E\). Then \(T^{-1}\) is an \(F[x]\)-module homomorphism from the action \((p, x) \mapsto p(B)x\) on \(F^n\) to the action \((p, y) \mapsto T^{-1}p(B)Ty\) on \(T^{-1}[F^n]\). \(B, B'\) thus have the same elementary divisors in \(F[x]\) and are similar.

Given a matrix \(B\) over \(F\), let \(T\) be a matrix over \(E \supseteq F\) transforming \(B\) to Jordan canonical form \(B'\). Consider the matrix \(xI - B\) over \(F[x]\) (where \(I\) is the identity matrix); this transforms to \(xI - B'\). The determinant of \(xI - B'\) is the characteristic polynomial; hence so is the determinant of \(xI - B\). That is, \(p_C = \det(xI - B)\). It follows easily that the diagonal entries of a triangular matrix are the eigenvalues. Also, a matrix is similar to a triangular matrix over \(F\) iff its eigenvalues are all in \(F\).

It should be remarked that we have defined \(p_M\) and \(p_C\) in terms of the rational canonical form, and proved various properties. Frequently \(p_C\) is defined as \(\det(xI - B)\). The Cayley-Hamilton theorem then states that \(B\) satisfies this polynomial; and \(p_M\) is defined to be the monic polynomial of lowest degree satisfied by \(B\).

It is easy to see that a matrix \(B\) is singular iff \(Bv = 0\) for some \(v \neq 0\). Thus, for \(\alpha \in F\), \(\det(\alpha I - B) = 0\) iff \(\alpha I - B\) is singular iff \((\alpha I - B)v = 0\) for some \(v \neq 0\) iff \(Bv = \alpha v\) for some \(v \neq 0\). Such a \(v\) is called an eigenvector of, or belonging to, the eigenvalue \(\alpha\).

We suppose for the rest of the section that the eigenvalues of any matrix lie in \(F\). The eigenvectors belonging to an eigenvalue \(\alpha\) form a subspace, called the eigenspace of \(\alpha\). If \(B_\alpha\) is a Jordan block for eigenvalue \(\alpha\), \(B_\alpha x = \beta x\) has a nonzero solution iff \(\beta = \alpha\), in which case it has a one-dimensional subspace of solutions. It follows (considering the basis where \(B\) is in Jordan canonical form) that the eigenspace for \(\alpha\) has dimension equal to the number of Jordan blocks with eigenvalue \(\alpha\); this is called the geometric multiplicity of \(\alpha\), and the usual multiplicity is also called the algebraic multiplicity. Note that vectors from distinct eigenspaces are linearly independent; for a basis independent proof of this see exercise 5.

An eigenvalue is called defective if its geometric multiplicity is less than its algebraic multiplicity, and a matrix is called defective if it has a defective eigenvalue. A matrix \(B\) is called diagonalizable iff there is a basis consisting of eigenvectors of \(B\). Clearly this is so iff the Jordan canonical form is diagonal, iff \(B\) is similar to a diagonal matrix, iff \(B\) has no defective eigenvalues. It is easy to see that \(B\) is diagonalizable iff \(p_M\) splits into distinct linear factors, since the minimal polynomial of a Jordan block has degree equal to the dimension of its subspace.

Say that two \(n \times n\) matrices \(A, B\) commute if \(AB = BA\). Say that a set \(\{A_i\}\) commutes if each pair \(A_i, A_j\) does. If this is so then the subspace of the space of matrices generated by \(\{A_i\}\) commutes, so we
assume that \{A_i\} is finite. Note that a commuting set of matrices remains commuting after a change of basis. If \(A, B\) commute and \(W\) is an eigenspace of \(A\) then \(W\) is an invariant subspace of \(B\); indeed, if \(Aw = aw\) then clearly \(ABw = aBw\).

It is now easy to see that if \(\{A_i\}\) commutes then there is a vector \(w\) which is an eigenvector of every \(A_i\). Let \(W_1\) be any eigenspace of \(A_1\); inductively, let \(W_i\) be the vectors in \(W_{i-1}\) which belong to some eigenvalue of \(A_i\) acting on \(W_{i-1}\).

If \(\{A_i\}\) commutes then there is a basis in which every \(A_i\) is upper triangular. To see this, choose a basis in which the first column of every matrix is \(a_1e_1\) where \(e_1\) is the standard unit vector and \(A_ie_1 = a_ie_1\). We may continue inductively on the subspace spanned by the remaining standard unit vectors. As a corollary, if \(AB = BA\) it is possible to order the eigenvalues \(a_1, \ldots, a_n\) of \(A\) and \(b_1, \ldots, b_n\) of \(B\) so that \(a_1b_1, \ldots, a_nb_n\) are the eigenvalues of \(AB\). It is also clear that the eigenvalues of \(p(A)\) are \(p(a_1), \ldots, \); and if \(A\) is invertible (i.e., its eigenvalues are all nonzero) then the eigenvalues of \(A^{-1}\) are \(a_i^{-1}, \ldots\).

If \(\{A_i\}\) commutes and each \(A_i\) is diagonalizable then there is a basis in which every \(A_i\) is diagonal. Let \(a_{it}\) be the eigenvalues of \(A_i\), and \(W_{it}\) the corresponding eigenspaces. Fix \(r\) and let \(X\) be any set of the form \(W_{1r_1} \cap \cdots \cap W_{rr_r}\). Distinct \(X\) clearly have trivial intersection, and each \(X\) is an invariant subspace of each \(A_i\). Since the \(A_i\) are diagonalizable, \(F^n\) is the direct sum of the nontrivial \(X\). Choose any basis for each nontrivial \(X\) and take the union of these; the result is a basis for \(F^n\) in which each \(A_i\) is diagonal. For \(r\) sufficiently large the \(X\) will no longer change as \(r\) is increased.

6. Invariants. Let \(M\) be a module over a principal ideal domain \(R\). As mentioned in chapter 8 a principal submodule \(Rv\) is isomorphic to \(R/\alpha R\) where \(\alpha\) is the period. Conversely if a submodule is isomorphic to \(R/\alpha R\) for some \(\alpha\) it is a principal submodule of period \(\alpha\), since \(R/\alpha R\) is. It is easily verified that if \(M\) is torsion and \(M = M_1 \oplus M_2\) then the period of \(M\) is the least common multiple of the periods of \(M_1\) and \(M_2\). Also, if \(p^i\) is the highest power of the prime \(p\) dividing the period of \(M\) then \(p^i\) is the period of \(M_{p^i}\), where recall the latter is \(\{v \in M : \alpha p^i v = 0\}\). If further \(M\) is principal so is \(M_{p^i}\).

**Theorem 1.** Suppose \(M\) is a finitely generated torsion module over an integral domain \(R\). There are unique (up to multiplication by units) elements \(q_1, \ldots, q_r \in R\) such that \(q_r | \cdots | q_1\), and there are \(M_1, \ldots, M_r\) such that \(M_j\) is a principal submodule of period \(q_j\) and \(M = M_1 \oplus \cdots \oplus M_r\).

**Proof:** Let \(p_i^{e_{ij}}\), \(1 \leq i \leq s\), \(1 \leq j \leq r_i\), \(e_{ij} \geq \cdots \geq e_{i1}\), be the elementary divisors. Let \(M_{ij}\) be such that \(M = \bigoplus_{ij} M_{ij}\) and \(M_{ij}\) is a principal submodule of period \(p_i^{e_{ij}}\). Let \(r = \max\{r_1\};\) for \(1 \leq j \leq r\) let \(q_j = \prod_{i}, \alpha^{e_{ij}}\) where \(e_{ij}\) is taken as \(0\) for \(j > r_i\), and let \(M_j = \bigoplus_{ij} M_{ij}\) where \(M_{ij}\) is taken as \(0\) if \(j > r_i\). This shows that the \(q_j\) exist. Conversely, given such \(q_j\) write \(q_j\) as \(q_j = \prod_i p_i^{e_{ij}}\), and let \(M_{ij} = (M_j)^{e_{ij}}\). By uniqueness of the elementary divisors, these are the \(p_{ij}\).

The \(q_j\) are called the invariants of \(M\). In the case of the action of \(F[x]\) on the \(n \times n\) matrices induced by a matrix \(B\), \(q_1\) is the minimal polynomial; the product of the invariants is the characteristic polynomial. In this case the invariants do not change if the field is enlarged to an extension \(E \supseteq F\). This is so because the subspace of \(E^n\) generated by a basis for \(M_j\) is generated as a principal submodule over \(E[x]\) by the same vector as over \(F[x]\).

7. Bilinear forms. Suppose \(R\) is a commutative ring and \(M\) an \(R\)-module. As noted in section 8.2 the elements of \(L_R(M, M; R)\) are called bilinear forms on \(M\). If \(M = R^n\) these are in one-to-one correspondence with the \(n \times n\) matrices over \(R\). \(M\) together with a bilinear form on it is called a bilinear space. If \(N \subseteq M\) is a submodule the form may be considered as one on \(N\), by taking the restriction. Recalling the maps \(\alpha_l\) and \(\alpha_r\) of section 2, a bilinear form is called nondegenerate if \(\text{Ker}(\alpha_l) = \text{Ker}(\alpha_r) = 0\). When \(R\) is an integral domain and \(M = R^n\) this is so iff the matrix in nonsingular.
Suppose $R$ is equipped with an automorphism of order 2, which we will denote by $a \mapsto a^*$. A map $\psi : M \mapsto N$ between $R$-modules is called antilinear if $\psi(x + y) = \psi(x) + \psi(y)$, and $\psi(ax) = a^* \psi(x)$. A map $*: M \times M \mapsto R$ is called sesquilinear if it is linear in its first argument and antilinear in its second, i.e., if
\[
(\begin{align*}
(1 + x_2) * y &= (x_1 * y) + (x_2 * y), \\
(ax) * y &= a(x * y) \\
x * (y_1 + y_2) &= (x * y_1) + (x * y_2), \\
x * (ay) &= a^*(x * y).
\end{align*}
\]
The remarks of the first paragraph apply to sesquilinear forms also. Indeed, defining $M_\sigma$ to be $M$ with the action $(a, x) \mapsto a^* x$, a sesquilinear map is a bilinear map in $L_R(M, M_\sigma; R)$.

The identity $(x + y) * (x + y) = x * x + x * y + y * x + y * y$ (observed in chapter 8 for bilinear forms) holds also for sesquilinear forms. Replacing $y$ by $-y$ and adding yields $(x + y) * (x + y) + (x - y) * (x - y) = 2(x * x + y * y)$; this is called the parallelogram law, because in an inner product space (defined below), where $x * x = |x|^2$, it expresses the fact that the sum of the squares of the lengths of the diagonals of a parallelogram equals the sum of the squares of the lengths of the sides. Replacing $y$ by $-y$ and subtracting yields $(x + y) * (x + y) - (x - y) * (x - y) = 2(x * y + y * x)$; this (and closely related identities) is called the polarization identity.

A bilinear form $\ast$ is called
- symmetric if $y * x = x * y$; and
- alternating if $x * x = 0$.

These are special cases of a definition in section 8.2, where it is also observed that if $\ast$ is alternating then $y * x = -x * y$; and conversely, provided 2 is not a zero divisor. A sesquilinear form $\ast$ is called Hermitian if
- Hermitian if $y * x = (x * y)^*$.

If $M = R^n$ and $A$ is the matrix of $\ast$ it is easily seen that $\ast$ is symmetric iff $A^t = A$; alternating iff $A^t = -A$ and diagonal entries are 0 (which follows from $A^t = -A$ if 2 is not a zero divisor); and Hermitian iff $A^{*t} = A$.

In the case of a symmetric bilinear form some identities given above simplify, namely $2(x * y) = (x + y) * (x + y) - x * x - y * y$, and $4(x * y) = (x + y) * (x + y) - (x - y) * (x - y)$.

If $M = R^n$ a quadratic form is a function $f : M \mapsto R$ of the form
\[
f(x_1, \ldots, x_n) = \sum_{i \leq j} a_{ij} x_i x_j.
\]
Its matrix $A$ is upper triangular. If $\ast$ is a symmetric form then $x * x$ is a quadratic form. If 2 has a multiplicative inverse then conversely given a quadratic form $f$ let $B = (A + A^t)/2$; this is the matrix of a symmetric form $\ast$ such that $f$ equals $x * x$. 2 may fail to have a multiplicative inverse, even if it is not a zero divisor, the standard example being $\mathcal{Z}$. The distinction between symmetric forms and quadratic forms is thus necessary in various contexts. In such cases, $(A + A^t)$ is sometimes called the bilinear form associated with $f$. Its quadratic form is $2f$; it is readily verified to be given by the expression $f(x + y) - f(x) - f(y)$.

Facts which hold for bilinear forms over $\mathcal{R}$ or $\mathcal{C}$ can be shown to hold for bilinear forms over fields $F$ satisfying appropriate restrictions. Say that $F$ is weakly $\mathcal{R}$-like if it is ordered; say that it is $\mathcal{R}$-like if in addition every positive element has a square root, and let $\sqrt{x}$ denote the positive one. Say that $F$ is (weakly) $\mathcal{C}$-like if it is $F_\sigma(\sqrt{-1})$ for some (weakly) $\mathcal{R}$-like subfield $F_\sigma$. In this case, the map $a + b\sqrt{-1} \mapsto a - b\sqrt{-1}$ is an automorphism of order 2, which we denote $x \mapsto x^*$; it satisfies $x \in F_\sigma$ iff $x^* = x$. We always assume $F$ is equipped with this automorphism when considering a sesquilinear form $\ast$ in a vector space $\mathcal{V}$ over $F$.

To give some facts uniformly, it is convenient to equip a weakly $\mathcal{R}$-like field with the trivial map $x^* = x$, and also to let $F_\sigma = F$. For example we may define $|x|$ to be the quantity $\sqrt{x^2}$, in both the $\mathcal{C}$-like and $\mathcal{R}$-like case. This map is readily verified to satisfy the identities $|x| \geq 0, |x| = 0$ iff $x = 0, |x + y| \leq |x| + |y|$, and $|xy| = |x||y|$. 75
Although some facts hold in other cases, we suppose for the rest of the section that the ring is a field of one of the above types; and a module is a vector space $V$. In the weakly $C$-like case, if $*$ is Hermitian then $x * x \in F$ (because $(x * x)^* = x * x$). If $F$ is weakly $R$-like and $*$ is symmetric, or $F$ is weakly $C$-like and $*$ is Hermitian, say that $*$ is positive definite bilinear form if $x * x > 0$ whenever $x \neq 0$. A null vector is a nonzero vector $x$ such that $x * x = 0$; $*$ is positive definite iff there are no null vectors.

To avoid repetition, a (weak) inner product inner product is defined to be a symmetric form with the axioms for a norm and the axioms for a norm $R$-like, or a Hermitian form with the field (weakly) $C$-like, which is positive definite. If $*$ is an inner product the quantity $|x|$ may be defined as above. It satisfies the Cauchy-Schwarz inequality. To prove the triangle inequality, square both sides (noting that both sides are nonnegative) and use the Cauchy-Schwarz inequality. From the proof it easily follows that equality holds iff $x * y = |x||y|$, and that this holds for nonzero $x, y$ iff $x/|x| = y/|y|$, or equivalently $y = cx$ for some scalar $c > 0$.

8. Orthogonal sums. In this section $*$ is a symmetric, alternating, or a Hermitian form on a vector space $V$ over a field $F$. Then $x * y = 0$ iff $y * x = 0$. If $S \subseteq V$ define $S^\perp$ to be $\{v \in V : v * x = 0, \text{ all } x \in S\}$; or equally well $\{v \in V : x * v = 0, \text{ all } x \in S\}$. The following are easily verified.

- If $S_1 \subseteq S_2$ then $S_1^\perp \supseteq S_2^\perp$, and $S \subseteq T^\perp$ iff $S^\perp \supseteq T$.
- $S \subseteq S_1^\perp$, $S_1^\perp \perp \perp = S_1^\perp$, and for subspaces $U_a \perp (\cup_a U_a)^\perp = \cap_a U_a^\perp$ (these follow from the previous facts by partial order theory; see section 13.7).
- $S^\perp$ is a subspace, and equals $\text{Span}(S)^\perp$.

The kernel on the left or right equals $V^\perp$, and is called simply the kernel of $*$; $*$ is nondegenerate iff this is trivial.

Lemma 2. If $W_1, W_2$ are subspaces of a finite dimensional vector space $V$ then $\dim(W_1 + W_2) = \dim(W_1) + \dim(W_2) - \dim(W_1 \cap W_2)$.

Proof: This follows because $(W_1 + W_2)/W_1$ and $W_2/(W_1 \cap W_2)$ are isomorphic.

Lemma 3. Suppose $V = F^n$, $*$ is nondegenerate, and $W \subseteq V$ is a subspace.

a. $\dim(W) + \dim(W^\perp) = n$.
b. $W^\perp = W$.
c. $*$ is nondegenerate on $W$ iff $W \cap W^\perp = 0$.
d. $*$ is nondegenerate on $W$ iff $V = W \oplus W^\perp$.
e. If $*$ is nondegenerate on $W$ then $*$ is nondegenerate on $W^\perp$.

Proof: Let $A$ be the matrix of $*$. For part a, let $w$ denote $\dim(W)$; choose a basis in which $W$ has the first $w$ standard unit vectors as a basis, and let $A_1$ be the leftmost $w$ columns of $A$. Since $*$ is nondegenerate, the column space of $A_1$ has dimension $w$. The kernel of the map from $F^w$ to $F^n$ determined by $A_1$ is $W^\perp$, and the claim follows. Part b follows from part a and $W \subseteq W^\perp$. For part c, let $A_{11}$ be the top $w$ rows of $A_1$; $*$ is nondegenerate on $W$ iff $A_{11}$ is nonsingular, which is so iff for any nonzero row vector $y^t$ which is 0 outside the first $w$ components, $y^t A_{11} \neq 0$. For part d, by lemma 2 and part a $\dim(W + W^\perp) = n - \dim(W \cap W^\perp)$; the claim follows by part c. For part e, by symmetry $V = W^\perp \oplus W^\perp \perp$ iff $*$ is nondegenerate on $W^\perp$, and the claim follows by part b.
For nondegenerate * say that a subspace W is nondegenerate if * is nondegenerate on W. If * is an inner product on V then any subspace of V is clearly nondegenerate. Call a set S ⊆ V of nonzero vectors orthogonal if v * w = 0 for v, w ∈ S, v ≠ w. If in addition none of the vectors is null, S is linearly independent (exercise 1). Call a vector space equipped with an inner product an inner product space; in such a space a vector of norm 1 is called a unit vector. If V is an inner product space and each v in an orthogonal set S is a unit vector call S orthonormal.

Let K ⊆ V be the kernel of *, and let V1 be such that V = V1 ⊕ K. It is readily verified that V1 is nondegenerate. An orthogonal basis for V1 may be extended by any basis for K to an orthogonal basis for V. On the other hand any vector in an orthogonal basis for V is null iff it is in K. In the next two theorems we may thus suppose * is nondegenerate.

**Theorem 4.** Suppose V = F^n, and * is symmetric or Hermitian, and nondegenerate. Suppose also that any nondegenerate subspace contains a nonnull vector. Then V has an orthogonal basis.

**Proof:** V must contain some nonnull v. If n = 1 we are done, else we proceed by induction, as we may since the hypotheses continue to hold (e.g. the definition of symmetric or Hermitian is universal, so holds in a substructure). Let W be the subspace generated by v; clearly * is nondegenerate on W, so by lemma 3 V = W ⊕ W⊥, and * is nondegenerate on W⊥. Let {v2, ..., vn} be a basis for W⊥; then {v1, v2, ..., vn} is a basis for V, as is readily verified. Further it is orthogonal.

The hypothesis that a nondegenerate subspace contains a nonnull vector is met if * is symmetric and the characteristic of F is not 2. Indeed, if x * y ≠ 0 then x * y + y * x ≠ 0, whence one of x, y, or x + y is nonnull. It is also met if F is weakly C-like and * is Hermitian. In this case if x * y + y * x = 0 replace x by ix. Finally we note that if the characteristic of F is not 2, V = W ⊕ W⊥, and * is nondegenerate on W, W⊥, then * is nondegenerate on V, as is readily verified.

In an inner product space, an orthogonal basis can be transformed to an orthonormal one by “normalizing”, i.e., replacing each bj by bj/|bj|. There is a procedure for transforming any basis {v1, ..., vn} into an orthonormal basis {v′1, ..., v′n}, called Gram-Schmidt orthonormalization. If u is a unit vector the projection of any vector x on u is defined to be x * u. Given an orthonormal basis for a subspace, the projection of x onto the subspace is the sum of its projections onto the basis vectors. The Gram-Schmidt procedure subtracts from vi its projection onto the subspace generated by the v′j, j < i, leaving a vector normal to the subspace.

In more detail, let v′i = vi/|vi|. Inductively, for i > 1 let v′′i = vi − ∑j<i(vi * v′j)v′j, and let v′i = v′′i/|v′′i|. One readily verifies that v′′i is nonzero and orthogonal to v′j, j < i. By exercise 1 {v′1, ..., v′n} is linearly independent; it is readily seen that it has the same span as {v1, ..., vn}. Note that if v1, ..., vk are orthonormal then one may start at vk+1. It follows that any orthonormal set of vectors can be expanded to an orthonormal basis. Note also that the Gram-Schmidt procedure yields an orthonormal basis in the case that V is of countably infinite dimension, from any basis {vi : i ∈ N}, by simply continuing for all i.

A two dimensional subspace on which * is nondegenerate is called a hyperbolic plane if it contains a null vector. It always does if * is alternating.

**Theorem 5.** If V = F^n and * is alternating and nondegenerate then V is an orthogonal sum of hyperbolic subspaces. In particular n is even.

**Proof:** Let v be any nonzero vector, and let w be any second vector such that w * v ≠ 0. Let W = Span(w, v); then W is a hyperbolic plane. The rest of the proof is as in theorem 4.

Suppose V is finite dimensional, and x1 is the coordinates of a vector in basis B1, i = 1, 2; then x2 = Tx1 where T is a matrix whose columns are are the coordinates in B1 of the vectors in B2. If A is such that y1 * x1 = y1′ Ax1 in B1, then y1 * x1 = (Ty1) * (Tx2) = Ty1(T′AT)x2, so that the matrix for * in B2 is T′AT.

We leave it as an exercise to verify the following. 77
- If \( V = U \oplus U^\perp \) then there is a basis where the matrix of \(*\) has the form

\[
\begin{bmatrix}
A_1 & 0 \\
0 & A_2
\end{bmatrix}
\]

- In an orthogonal basis the matrix for \(*\) is diagonal. The number of 0’s on the diagonal equals the dimension of the kernel.

- A hyperbolic plane has a basis in which the matrix is of the form

\[
\begin{bmatrix}
0 & 1 \\
1 & 0
\end{bmatrix} \quad \text{or} \quad \begin{bmatrix}
0 & 1 \\
-1 & 0
\end{bmatrix}
\]

in the symmetric or Hermitian, and alternating, cases respectively.

**Theorem 6 (Sylvester’s Law of Inertia).** Suppose \( V = F^n; \) \( F \) is weakly \( R \)-like and \(*\) is symmetric, or \( F \) is weakly \( C \)-like and \(*\) is Hermitian; and \(*\) is nondegenerate. Then there is an integer \( r \) such that in any orthogonal basis \( S, \{ v \in S : v \ast v > 0 \} = r. \)

**Proof:** Let \( v_1, \ldots, v_n \) be an orthogonal basis with \( v_i \ast v_i > 0 \) iff \( i \leq r \), and let \( w_1, \ldots, w_n \) be an orthogonal basis with \( w_i \ast w_i < 0 \) iff \( i \leq s \). It suffices to show that \( v_1, \ldots, v_r, w_1, \ldots, w_s \) are linearly independent; for then \( r + s \leq n \) and by symmetry \((n - r) + (n - s) \leq n. \) Write a linear combination summing to 0 as \( \sum_i a_i v_i = -\sum_j b_j w_j \) and apply \( x \mapsto x \ast x \) to both sides, yielding \( \sum_i a_i^2 v_i \ast v_i = \sum_j b_j^2 w_j \ast w_j. \) The left side is \( \geq 0, \) and the right side \( \leq 0, \) so both sides are 0 and the claim follows.

In terms of matrices, if \( M \) is nonsingular then in any way of writing \( M \) as \( S' DS \) where \( D \) is diagonal and \( S \) has orthogonal columns, the number of + entries in the diagonal of \( D \) is determined. By the remarks preceding theorem 4, for any \( M \) the triple of numbers of +0− entries is determined; this triple is called the inertia of \( M. \) Clearly the number of 0 entries equals the nullity of \( M. \)

9. Witt’s theorem. An isomorphism of bilinear or sesquilinear spaces is a linear map \( T \) which is bijective, and preserves the form, that is, such that \((Ty) \ast (Tx) = y \ast x. \) If \( M = F^n \) and \( A_1, A_2 \) are the matrices of the two forms, the second requirement may be written \( A_2 = T^t A_1 T, \) or \( A = T' A T \) in the case of an automorphism. For the rest of the section, suppose \( V = F^n; \) \( F \) is a field of characteristic other than 2 and \(*\) is a symmetric, or \( F \) is \( C \)-like and \(*\) is Hermitian; and \(*\) is nondegenerate. As observed in the previous section, every nondegenerate subspace contains a nonnull vector.

**Lemma 7.** If \( u, v \in V \) with \( u \ast u = v \ast v \neq 0 \) there is an automorphism \( T \) of \( V \) such that \( T(u) = v. \)

**Proof:** If \( w \) is nonnull let \( T_w \) be defined by

\[
T_w(x) = x - 2 \frac{\langle x, w \rangle}{\langle w, w \rangle} w.
\]

It is readily verified that \( T_w \) is linear, in fact an automorphism, \( T_w(w) = -w, \) and \( T_w(x) = x \) if \( w \ast x = 0. \) \( T_w \) may be described as the “reflection” in the “hyperplane normal to” \( w. \) If \(*\) is symmetric then since \( u \ast u = v \ast v, (u - v) \ast (u - v) = 2u \ast (u - v). \) If \( u - v \) is nonnull then we may take \( T = T_{u-v}. \) Otherwise, since \( (u - v) \ast (u - v) + (u + v) \ast (u + v) = 4u \ast u, u + v \) is nonnull and we may take \( T = T_v T_{u+v}(u) = v. \) In the Hermitian case, note that if \( a = \pm |u \ast v|/u \ast v \) then \( av \ast u = u \ast av. \) We may proceed as in the symmetric case to map \( u \) to \( \pm av. \) Since multiplication by a scalar \( a \) with \( |a| = 1 \) is an automorphism, we are done.

**Theorem 8 (Witt’s theorem).** If \( W_1, W_2 \) are nondegenerate subspaces and \( T : W_1 \mapsto W_2 \) is an isomorphism then \( T \) can be extended to an automorphism of \( V. \)
Lemma 10. are hyperbolic planes $P_i$ for $i > 1$. Write $W_i$, $i = 1, 2$, as $U_i \oplus K_i$ where $K_i$ is the kernel of $*$ on $W_i$. By the following lemma expand $W_i$ to a subspace of the form $U_i \oplus P_{i1} \oplus \cdots \oplus P_{is}$ where the direct sum is orthogonal, $P_{ij}$ is a hyperbolic plane, and $s = \dim(K_i)$. Since $T[U_1] = U_2$, and two hyperbolic planes are isomorphic, the claim follows using the theorem.

Lemma 9. Suppose $W \subseteq V$ is a subspace, and $W = U \oplus K$ where $K$ is the kernel of $*$ on $W$. Then there are hyperbolic planes $P_i$, $1 \leq i \leq s$ where $s = \dim(K)$, such that $W$ is contained in the orthogonal direct sum $U \oplus P_1 \oplus \cdots \oplus P_s$.

Proof: Let $\{v_1, \ldots, v_s\}$ be a basis for $K$. We claim there is a $w_1 \in W^\perp$ such that $w_1 \not\in \{v_1, \ldots, v_s\}$ for $i > 1$. Indeed, $U \oplus \text{Span}(v_2, \ldots, v_s) \subseteq W$, so by lemma 3 $W^\perp \subset (U \oplus \text{Span}(v_2, \ldots, v_s))^\perp$, which proves the claim. Now, $P_1 = \text{Span}(w_1, v_1)$ is a hyperbolic plane, and replacing $U_1 \oplus \text{Span}(v_1)$ by $U \oplus P_1$ we may proceed inductively.

10. Forms and operators. Suppose $R$ is a commutative ring and $M$ is an $R$-module, and $*$ is a bilinear or sesquilinear form on $M$. Let $M^*$, $\alpha_l$, and $\alpha_r$ be as in section 2. Call $*$ $l$-nondegenerate if $\alpha_l$ is injective, and $r$-regular if $\alpha_l$ is bijective; and similarly for $\alpha_r$.

The form $*$ induces a map $\rho_r : L_R(M; M) \rightarrow L_R(M, M; R)$ (the codomain is $L_R(M, M; \sigma; R)$ if $*$ is sesquilinear), which maps the linear map ("operator") $T$ to the form $x \ast T(y)$. Similar claims hold for $l$.

Lemma 10. If $*$ is $r$-nondegenerate then $\rho_r$ is injective, and if $*$ is $r$-regular then $\rho_r$ is bijective. Similar claims hold for $l$.

Proof: Suppose $x \ast T(y) = 0$ for all $x, y$; then for all $y \in K_r$, so if $*$ is $r$-nondegenerate then $T(y) = 0$ for all $y$. This proves that $\rho_r$ is injective. Given any form $x \circ y$, the map $\psi_y = x \circ y$ is in $M^*$, so if $*$ is $r$-regular $\psi_y$ equals $\alpha_r(w) = \phi_w$ for some $w \in M$. The map $T : y \mapsto w$ is readily verified to be linear, and $x \circ y = x \ast T(y)$. This proves that $\rho_r$ is surjective.

If $*$ is symmetric, alternating, or Hermitian, one readily verifies that $*$ is $r$-nondegenerate (r-regular) iff it is $l$-nondegenerate (l-regular), in which case we call it simply nondegenerate (regular). As noted previously, an arbitrary form is called nondegenerate if it is both $r$- and $l$-nondegenerate; the same applies for regular. For $M = R^n$ $*$ is regular iff the corresponding matrix is invertible.

In many contexts one fixes a nondegenerate symmetric, alternating, or Hermitian form $*$, for example the dot product on $F^n$ or $\int f(x)g(x)dx$ on some vector space of functions. Various definitions can then be given, with respect to the form, for example that of an orthogonal set of vectors. Say that the operator $T^\dagger$ is adjoint to the operator $T$ if $y \ast T(x) = T^\dagger(x) \ast y$ (the term is usually used for a Hermitian form, but we use it in all three cases). By lemma 10, since $*$ is nondegenerate there is at most one adjoint, and if $*$ is regular then one exists.

Suppose $S^\dagger$ and $T^\dagger$ exist; the following are readily verified.
- $(cT)^\dagger = c^*T^\dagger$ (if $c^*$ is weakly R-like), $(S + T)^\dagger = S^\dagger + T^\dagger$, $(ST)^\dagger = T^\dagger S^\dagger$, and $T^{\dagger\dagger} = T$ (in particular these exist).
- If $*$ is symmetric then $\rho_r(T)$ is symmetric iff $T^\dagger = T$, in which case $y \ast T(x) = T(y) \ast x = x \ast T(y) = T(x) \ast y$. 

79
- If $*$ is Hermitian then $\rho_r(T)$ is Hermitian iff $T^\dagger = T$, in which case $(y * T(x))^* = \langle T(y) * x \rangle^* = x * T(y) = T(x) * y$.

Exercise 7 gives additional such facts. An operator $T$ is called self-adjoint if $T^\dagger = T$. If $*$ is Hermitian a self-adjoint operator is also called Hermitian; exercise 8 shows that when $F$ is weakly C-like this is so iff $x * T(x) \in F_r$ for all $x$. It is easily seen that $T$ is an automorphism of $F$ iff $T^\dagger$ exists and equals $T^{-1}$.

Restricting now $M$ to a vector space, in an inner product space, an automorphism of $*$ clearly maps any orthonormal basis to an orthonormal basis; conversely if $T$ maps some orthonormal basis $B$ to an orthonormal basis then $T$ is an automorphism, since for any linear combination $\sum a_i x_i$ of elements of $B$, $T(\sum a_i x_i) * T(\sum a_i x_i)$ and $\sum a_i x_i * \sum a_i x_i$ both equal $\sum |a_i|^2$. The terms orthogonal, symplectic, and unitary are used to refer to the group of automorphisms of a symmetric, alternating, and Hermitian form respectively; one also refers to orthogonal or unitary operators. If the field has characteristic other than 2, an operator is orthogonal iff $T(x) * T(y) = x * y$ for all $x, y$, as can be seen by considering $T(x+y) * T(x+y)$.

An operator $T$ is called normal if $T^\dagger$ exists, and $TT^\dagger = T^\dagger T$. Included in this class are operators with $T^\dagger = T, -T, T^{-1}$. For the rest of this section we abbreviate “orthogonal basis of eigenvectors” as o.b.e. For an operator to have an o.b.e. it is necessary that its eigenvalues be in $F$.

**Theorem 11.** Suppose $T$ has an o.b.e.; then $T$ is normal. If $*$ is symmetric or alternating, or $*$ is Hermitian and the eigenvalues satisfy $a^* = a$, then $T$ is self-adjoint.

**Proof:** Let $\{x_i : i \in I\}$ be the basis, with $T(x_i) = a_i x_i$. In the symmetric or alternating case,

$$T(x_i) * x_j = a_i (x_i * x_j) = a_j (x_i * x_j) = x_i * T(x_j);$$

this follows in the Hermitian case also if $a_i^* = a_i$. In general, let $S$ be the unique operator where $S(x_i) = a_i^*$; then

$$S(x_i) * x_j = a_i (x_i * x_j) = a_j (x_i * x_j) = x_i * T(x_j).$$

Thus, $T^\dagger$ exists and $T^\dagger(x_i) = a_i^* x_i$; and so

$$x_i * TT^\dagger(x_j) = T^\dagger(x_i) * T^\dagger(x_j) = a_i a_j^* (x_i * x_j) = a_i^* a_j (x_i * x_j) = T(x_i) * T(x_j) = x_i * T^\dagger T(x_j),$$

and $TT^\dagger = T^\dagger T$.

For the rest of this section we suppose $*$ is a weak inner product. In particular $x = 0$ if $x * x = 0$.

**Lemma 12.** For any operator $T$,

a. $T^\dagger T$ is self-adjoint;

b. $T^\dagger T(x) = 0$ iff $T(x) = 0$;

c. $x * T^\dagger T(x) \geq 0$.

If $T$ is normal then

d. $T - aI$ is normal for any $a \in F$;

e. if $T(x) = 0$ for $x$ nonnull then $T^\dagger(x) = 0$;

f. if $T(x) = ax$ for $x$ nonnull then $T^\dagger(x) = a^* x$.

g. if $T^2(x) = 0$ for some $k \geq 1$ then $T(x) = 0$;

h. eigenvectors belonging to distinct eigenvalues are orthogonal.

**Proof:** Part a is a trivial computation; part b follows because $x * T^\dagger T(x) = T(x) * T(x)$; and part c for the same reason. Part d is a straightforward calculation. For part e,

$$T^\dagger(x) * T^\dagger(x) = x * TT^\dagger(x) = x * T^\dagger T(x) = 0,$$
so $T^\dagger(x) = 0$. Part f follows from parts e and d. For part g, it suffices to show that $(T^\dagger T)(x) = 0$ if $(T^\dagger T)^2(x) = 0$, which follows by induction on $j$. For part h, suppose $T(x) = ax$, $T(y) = by$, $a \neq b$, $x, y \neq 0$; then

$$a(x * y) = (a^* x) * y = T^\dagger(x) * y = x * T(y) = x * (by) = b(x * y),$$

so $x * y = 0$.

Suppose $T(x) = ax$, $x \neq 0$. If $T$ is self-adjoint then by the lemma $T(x) = T^\dagger(x) = a^* x$, so $a = a^*$ and $a \in F_r$. Say that $T$ is positive definite with respect to $*$ if $x * T(x) > 0$ whenever $x \neq 0$; in this case if $a \in F_r$ then $a > 0$, since $a(x * x) = x * T(x)$ and $x * x > 0$, $x * T(x) > 0$. Note also that $T$ is injective. If $T^\dagger = T^{-1}$ then $x = T^{-1}T(x) = a^* ax$, so $|a| = 1$.

**Theorem 13.** Suppose $V = F^n$, $*$ is positive definite, and $T$ is normal and has its eigenvalues in $F$. Then $T$ has an o.b.e.

**Proof:** For an eigenvalue $a$ let $W_a$ be the sum of the principle submodules of the elementary divisors of the form $(x - a)^{e_i}$. If $e_i$ is the largest $e_j$ then $(A - aI)^{e_i}(x) = 0$ for any $x \in W_a$. By lemma 12, $x$ is thus an eigenvector belonging to $a$. We may thus choose any orthogonal basis for $W_a$, and take the union of these to obtain a basis for $V$, which is orthogonal by lemma 12.

The above may be applied to matrices by taking $*$ as the dot product and $V = F^n$; for simplicity we consider $F = C$, so $F_r = R$. Note that, if a matrix $T$ has real entries then $T^\dagger = T^\iota$, and a Hermitian (unitary) real matrix is the same as a symmetric (orthogonal) one. Also, the rank of $T$ is the same whether the field is considered to be $C$ or $R$. Clearly, a matrix has an o.b.e. iff it is similar to a diagonal matrix via a unitary transformation. An example of a diagonalizable matrix which is not normal is

$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

A symmetric real or Hermitian matrix has an o.b.e. (using the fundamental theorem of algebra), and has real eigenvalues; if the matrix is positive definite (i.e., $x^\iota T x = 0$ for all $x$) the eigenvalues are positive. An orthogonal real or unitary matrix has an o.b.e., and has eigenvalues of absolute value 1. An obvious question is whether the eigenvector can be taken as real when the matrix $T$ and the eigenvalue $a$ are; indeed, in such a case the real and imaginary parts of any eigenvector satisfy $T(x) = ax$, and one at least is nonzero. Note that a normal real matrix with real eigenvalues is symmetric.

**11. Projection operators.** Suppose $V$ is a vector space over a field $F$. As noted in chapter 8, if $V = U_1 \oplus U_2$ for $U_1, U_2 \subseteq V$ we may call $\langle U_1, U_2 \rangle$ a direct sum decomposition of $V$. Define a projection operator on $V$ to be a linear transformation $p : V \mapsto V$ such that $p^2 = p$. Given a projection operator $p$, let $\iota - p$ map $x$ to $x - p(x)$. We leave the following to the reader.

- $x \in \text{Im}(p)$ iff $p(x) = x$, and $\text{Im}(p) \cap \text{Ker}(p) = \{0\}$.
- $\iota - p$ is a projection operator, and $\text{Ker}(p) = \text{Im}(\iota - p)$.
- $\langle \text{Im}(p), \text{Ker}(p) \rangle$ is a direct sum decomposition of $V$.

**Theorem 14.** The map $p \mapsto \langle \text{Im}(p), \text{Ker}(p) \rangle$ maps the projection operators bijectively to the direct sum decompositions. The inverse map maps $\langle U_1, U_2 \rangle$ to the map (with obvious notation) $u_1 + u_2 \mapsto u_1$.

**Proof:** The map $u_1 + u_2 \mapsto u_1$ maps $u_1 = u_1 + 0$ to $u_1$ so is a projection. Since $x = p(x) + (\iota - p)(x)$, $\langle \text{Im}(p), \text{Ker}(p) \rangle$ gets mapped to $p$. Given $\langle U_1, U_2 \rangle$ let $p$ be the map $u_1 + u_2 \mapsto u_1$; clearly $\text{Im}(p) = U_1$ and $\text{Ker}(p) = U_2$.  

81
Now suppose $*$ is an inner product on $V$. The map $U \mapsto U^\perp$ is readily seen to have the following properties.

- $U \cap U^\perp = 0$, so if $U_1 \subseteq U_2^\perp$ then $U_1 + U_2 = U_1 \oplus U_2$. 
- If $U_1 \oplus U_2 = V$ and $U_1 \subseteq U_2^\perp$ then $U_1 = U_2^\perp$. For suppose $y \in U_2^\perp$ and write $y = u_1 + u_2$ as its decomposition under the direct sum. Then $0 = u_2 \ast y = u_2 \ast u_1 + u_2 \ast u_2$; since $u_2 \ast u_1 = 0$, it follows that $u_2 = 0$. 
- In particular if $U \oplus U^\perp = V$ then $U = U^\perp \perp$. This claim also follows because $U_1 \subseteq U_2$ and $U_2 \cap W = 0$ then $U_1 \oplus W \subseteq U_2 \oplus W$.

Call a direct sum decomposition $(U_1, U_2)$ orthogonal if $U_2 = U_1^\perp$. Call a projection operator $p$ orthogonal iff $(v - p)(x) \ast p(x) = 0$ for all $x$, that is, if $\text{Im}(p) \subseteq \text{Ker}(p)^\perp$ (note that this is distinct from the use of the term for operators in general, which must be injective to preserve an inner product). Then clearly $p$ is orthogonal iff $(\text{Im}(p), \text{Ker}(p))$ is; that is, the map $p \mapsto (\text{Im}(p), \text{Ker}(p))$ induces a bijection from the orthogonal projection operators to the orthogonal decompositions.

**Theorem 15.** A projection operator is orthogonal iff it is self-adjoint.

**Proof:** Suppose $p$ is self-adjoint. If $x \in \text{Ker}(p)$ and $p(u) = v$ then $v \ast x = p(u) \ast x = u \ast p^\dagger(x) = u \ast p(x) = 0$; thus $\text{Ker}(p) \subseteq \text{Im}(p)^\perp$. If $x \ast p(u) = 0$ for all $u$ then $p^\dagger(x) \ast u = 0$ for all $u$, so $p^\dagger(x) = 0$, so $p(x) = 0$; thus $\text{Im}(p)^\perp \subseteq \text{Ker}(p)$. Now suppose $\text{Ker}(p) = \text{Im}(p)^\perp$. Then

\[ p(x) \ast y = p(x) + (p(y) + (v - p)(y)) = p(x) \ast p(y) + p(x) \ast (v - p)(y) = p(x) \ast p(y); \]

similarly $x \ast p(y) = p(x) \ast p(y)$.

Suppose $V$ is finite dimensional. Then $U \oplus U^\perp$ always equals $V$, by lemma 3 (in section 24.8 it will be seen that this is not the case in general). The map $p \mapsto \text{Im}(p)$ is a bijection from the orthogonal projections to the subspaces. If $U$ is an $n \times k$ matrix whose columns are orthonormal, and $p$ is the map with matrix $UU^\dagger$, it is readily verified that $p$ is idempotent and self-adjoint, and $\text{Im}(p)$ is the column space of $U$.

12. **Factorizations.** If a matrix $A$ is written as a product of matrices having certain properties, the product is called a factorization or decomposition of $A$. For simplicity we mostly consider $F = \mathcal{R}$ or $F = \mathcal{C}$ in this section, although many facts hold more generally. An equivalence of matrices is a factorization $A = TBS$ where $S, T$ are square and nonsingular; this can be considered as changing the bases for the map $B : V \mapsto W$. A similarity is a factorization $A = S^{-1}BS$ where $S$ is nonsingular and all matrices are square; this can be considered as changing the basis for the map $B : V \mapsto V$.

Unless otherwise specified a rectangular matrix will be taken as $m \times n$, and a square one as $n \times n$. For $A$ square let $A^t$ denote the adjoint matrix $A^{t^*}$, which reduces to $A^t$ if $A$ is real. Several types of matrices may be defined, some of which we have already seen.

- Upper ($i > j \Rightarrow A_{ij} = 0$) and lower ($i < j \Rightarrow A_{ij} = 0$) triangular. Here $A$ may be rectangular. If $A$ is square and has $1$’s on the diagonal it is called unit triangular.

For the remaining types $A$ is square.

- $k$-band ($|i - j| < k \Rightarrow A_{ij} = 0$, some $k \geq 0$); diagonal if $k = 0$, tridiagonal if $k = 1$.
- Upper Hessenberg if $i > j + 1 \Rightarrow A_{ij} = 0$.
- Symmetric ($A^t = A$), alternating ($A^t = -A$), or orthogonal ($A^t = A^{-1}$); usually for real matrices. Alternating matrices are also called antisymmetric or skew-symmetric.
- Hermitian ($A^H = A$), skew-Hermitian ($A^H = -A$), or unitary ($A^H = A^{-1}$).
- Normal ($AA^t = A^tA$).
- Positive definite ($x^tAx = 0$ for $x \neq 0$). If $x^tAx \geq 0$ $A$ is called nonnegative definite; in this case it is positive definite if it is nonsingular.
For a square matrix $A$ we have the following similarities $A = S^{-1}BS$.

- $B$ in Jordan canonical form. $B$ is essentially unique. $S$ may be chosen real if $A$ and $B$ are.
- $A$ diagonalizable, $B$ diagonal. This is a special case of Jordan canonical form.
- $B$ upper triangular, $S$ unitary. This is called a Schur decomposition; we will show that it exists in theorem 16.
- $A$ normal, $B$ diagonal, $S$ unitary. This is a special case of both Jordan canonical form and Schur decomposition. $B$ is real if $A$ is Hermitian.

**Theorem 16.** A Schur decomposition exists.

**Proof:** If $n = 1$ the claim is trivial. Otherwise let $\lambda, x$ be such that $Ax = \lambda x$ and $|x| = 1$. Let $S$ be any unitary matrix whose first column is $x$. One readily checks that

$$S^\dagger AS = \begin{bmatrix} \lambda & A_1 \\ 0 & A_2 \end{bmatrix}$$

for some $A_1, A_2$. The theorem now follows by induction. Specifically, let $S_2^\dagger B_2 S_2$ be a Schur decomposition of $A_2$. Expand $S_2$ to a block diagonal matrix $S_3$ with leading block $1$. Then $S_3^\dagger A S S_3^\dagger$ is upper triangular.

If $A$ is normal then in any Schur decomposition $B$ is diagonal. This is a consequence of two readily verified facts. First, if $A$ is normal and $S$ is unitary then $S^\dagger AS$ is normal. Second, if $B$ is normal and upper triangular then $B$ is diagonal.

Suppose $a = A_{11}$ is nonzero, $A$ rectangular; then

$$\begin{bmatrix} 1/a & 0 \\ -A_1/a & I \end{bmatrix} \begin{bmatrix} a & A_2 \\ A_1 & A_3 \end{bmatrix} = \begin{bmatrix} a & A_2 \\ 0 & -A_1 A_2/a + A_3 \end{bmatrix}. $$

The leftmost matrix is called a Gauss transformation; it accomplishes a pivot step of Gaussian elimination in the entry $A_{11}$. It is unit lower triangular, as is its inverse, which has $A_1/a$ in the lower left block.

Suppose $A$ is rectangular and has rank $m$, or “full row rank”, whence $m \leq n$. A factorization $A = LUP$ where $L$ is $m \times m$ and unit lower triangular, $U$ is upper triangular with nonzero diagonal entries, and $P$ is an $n \times n$ permutation matrix, is called an LUP decomposition.

**Theorem 17.** If $A$ has full row rank it has an LUP decomposition.

**Proof:** Since $A$ has rank $m$ there is a permutation matrix $P_0$ such that the 1,1 entry of $AP_0$ is nonzero. If $m = 1$ we are done. Otherwise let $L_0$ be the Gauss transformation which accomplishes the pivot on the 1,1 entry of $AP_0$. We may now proceed inductively, similarly to theorem 16.

One use of the factorization $A = LUP$ is the solution of $Ax = b$, with several right sides $b$; the problem reduces to successively solving $Ly = b$, $Uz = y$, $Px = z$. Only one Gaussian elimination stage is necessary. Another use is obtaining theoretical algorithms for various matrix operations, including inverse and determinant, given one for matrix multiplication; these make use an algorithm for obtaining the factorization.

For a discussion of this topic, including some matrix multiplication algorithms, see [AHU].

If $S = \{1, \ldots, k\}$ where $k \leq m$ the minor $A_{SS}$ is called a leading principal minor. It is easily seen that a Gauss transformation does not alter the determinant of any of these. If the leading principal minors of $A$ are all nonsingular then the $i, i$ entry will never be 0, as $A$ is modified by applying the Gauss transformations. In this case we may omit any column permutations, and arrive at an $LU$ decomposition. We claim that an $LU$ decomposition is unique if it exists; indeed if $LU = L_1 U_1$ then $L_1^{-1} L = U_1 U^{-1}$ is unit lower triangular and upper triangular, so is the identity. It follows that the condition that the leading principal minors be nonsingular is also necessary for the existence of an $LU$ decomposition.
If $A$ is square and positive definite then all its principal minors (i.e., minors $A_{ST}$ where $S = T$) have positive determinant. Indeed, let $X$ be the matrix whose $i$th column is the standard unit vector $e_j$ where $j$ is the $i$th largest element of $S$. Then $X^tAX$ is $A_{SS}$; also, it is positive definite (this is easily seen to follow since $X$ has full column rank). In particular, $A$ has an $LU$ decomposition.

An $LU$ decomposition of a square matrix may be written as $LDM^\dagger$ where $M$ is unit lower triangular and $D$ is diagonal. This is unique if it exists. If $A$ is Hermitian then $M = L$; indeed $M^{-1}LD = M^{-1}A(M^{-1})^\dagger$ is Hermitian and lower triangular, so is diagonal; $M^{-1}L$ is therefore diagonal, so is the identity.

If $A$ is Hermitian and positive definite then the $LDL^\dagger$ factorization exists, and the diagonal elements of $D$ are positive. We may thus take their square roots, and obtain a factorization $A = LL^\dagger$ where $L$ is lower triangular and has positive diagonal entries, which is clearly unique. This is called the Cholesky factorization.

Suppose $A$ has full column rank. We claim that there is a factorization $A = Q_1R_1$, where $Q_1$ is $m \times n$ and has orthonormal columns; and $R_1$ is $n \times n$ and nonsingular upper triangular. Indeed, by the Gram-Schmidt procedure $Q_1 = AR_1^{-1}$ where $R_1^{-1}$ is upper triangular because column $i$ of $Q_1$ is a linear combination of the first $i$ columns of $A$. It is readily seen that $Q_1$ is unique up to right multiplication by a diagonal matrix with $\pm 1$’s on the diagonal.

The computation of $Q_1$ and $R_1$ can be organized as a sequence of column normalizations and elementary column operations. At stage $k$ normalize column $k$; then subtract from each column $j$, $j > k$, its projection onto column $k$. $R_1$ may be kept track of by the usual methods. This is called the modified Gram-Schmidt procedure.

This factorization is useful in least squares problems. A solution $x$ to the system $Ax = b$, $A$ rectangular and rank $n$, which minimizes $|Ax - b|$ is called a least squares solution. Equivalently, $x$ minimizes $(Ax - b)^\dagger(Ax - b)$; or letting $B = A^\dagger A$ and $c = A^\dagger b$,

$$x^\dagger Bx - (x^\dagger c + cx^\dagger) = (x - B^{-1}c)^\dagger B(x - B^{-1}c) - c^\dagger B^{-1}c.$$ 

Clearly $x = B^{-1}c$; that is, the least squares solution to $Ax = b$ is the unique solution to $A^\dagger Ax = A^\dagger b$. If $A = Q_1R_1$ as above, the latter equation becomes $R_1x = Q^\dagger b$.

The matrix $A^\dagger A$ is called the Gram matrix of $A$. It is self-adjoint, has the same kernel, rank, and row space as $A$, and is nonnegative definite. $A$ has full column rank iff it is positive definite. These facts hold more generally for matrices of the form $A^\dagger CA$ where $C$ is self-adjoint and positive definite. Another use of Gram matrices is the following, called Hadamard’s inequality; note that the bound is met by any matrix whose columns are orthogonal.

**Theorem 18.** Let $v_1, \ldots, v_n$ be the columns of the rank $n$ rectangular matrix $A$. Then $\det(A^\dagger A) \leq |v_1|^2 \cdots |v_n|^2$. In particular, if $A$ is square then $|\det(A)| \leq |v_1| \cdots |v_n|$.

**Proof:** Let $B$ be the matrix consisting of the first $n - 1$ columns of $A$; $W$ the column space of $B$; $v_n = u + w$ where $u \in W^\perp$ and $w \in W$; and $C$ the matrix obtained from $A$ by replacing $v_n$ by $u$. There is a vector $X$ such that

$$[B \ v_n] = [B \ u] \begin{bmatrix} I & X \\ 0 & 1 \end{bmatrix},$$

and $\det(A^\dagger A) = \det(C^\dagger C)$ follows. Further

$$C^\dagger C = \begin{bmatrix} B^\dagger \\ u^\dagger \end{bmatrix} \begin{bmatrix} B & u \\ 0 \ 0 \end{bmatrix} = \begin{bmatrix} B^\dagger B & 0 \\ 0 & |u|^2 \end{bmatrix}.$$ 

Thus, $\det(A^\dagger A) = |u|^2 \det(B^\dagger B)$; since $|u| \leq |v_n|$ the theorem follows by induction.
For $A$ of full column rank there is also a factorization $A = QR$, where $Q$ is $m \times m$ and unitary; and $R$ is $m \times n$ and upper triangular. This may be obtained from the $Q_1R_1$ factorization by completing the columns of $Q_1$ to an orthonormal basis of $\mathbb{R}^m$ and appending rows of 0’s to $R_1$.

A $QR$ factorization may be computed using reflections, which are also called Householder transformations if $F = \mathbb{R}$. If $u$ is a unit vector the reflection $Q_u$ in the hyperplane normal to $u$ may be written as $I - 2uu^t$. Note also that $Q_u$ is Hermitian and unitary (see exercise 10 for a converse). As in appendix 2 for $F = \mathbb{C}$, if $|x| = |y| \neq 0$ then for some $a$ of absolute value 1, $Q_{x-ay}x = ay$; $a$ may be 1 in the real case. In particular, if $x$ is any nonzero vector there is a unitary transformation mapping $x$ to $|x|e_1$ where $e_1$ is the standard unit vector with a 1 in row 1. We may proceed inductively as usual to obtain a $QR$ factorization.

The $QR$ factorization can be used to solve a least squares problem. Since $Q$ is orthogonal the minimum distance solution to $QRx = b$ is that to $Rx = Q^{-1}b$. Both the solution and the distance are easy to obtain from this system of equations.

An important use of the $QR$ factorization is in the “QR” algorithm for finding eigenvalues. A complete discussion of the QR algorithm, and of computational and numerical aspects of factorizations discussed in this section, can be found in [GvL].

The $Q_1R_1$ factorization can be constructed even when $A$ does not have full column rank, assuming $m \geq n$. If in step $i$ of the orthonormalization a linearly dependent column is encountered, leave the column of $Q_1$, “blank”, and set the column $R_i$ to the linear combination of the columns $Q_j$ for $j < i$ yielding $A_i$, with diagonal entry 0. In a final step, replace the blank columns with arbitrary unit vectors which result in an orthonormal set of columns. A $QR$ factorization can be obtained as before.

Still assuming $m \geq n$ let $U^tDU$ be the diagonalization of $A^tA$; the diagonal entries of $D$ are nonnegative. The columns of $AU$ are orthogonal, so the orthonormalization of the preceding paragraph results in a factorization $V_1E_1$ of $AU$ where $V_1$ is $m \times n$ and has orthonormal columns, and $E_1$ is diagonal; also $E_1^t = D^{1/2}$. This yields a factorization $VE$ where $V$ is $m \times m$ and unitary, and $E$ is $m \times n$ with $E_{ij} \neq 0$ only if $i = j$.

The factorization $VEU^t$ of $A$ is called a singular values decomposition (SVD) of $A$. The diagonal entries are called the singular values, and are determined up to sign. Any two signs can be flipped; if $A$ is square and has nonnegative determinant then all signs can be taken as +. If $m \leq n$ an SVD of $A$ is obtained by taking the adjoint of an SVD of $A^t$.

Suppose $m \geq n$ and let $E_1$ be as above, i.e., the first $n$ rows of $E$, and let $U'$ be $U^t$ with $m - n$ rows of 0’s appended below. Then $A = QP$ where $Q = VU'$ and $P = UE_1U^t$; $Q$ is $m \times n$ and $Q^tQ = I$, and $P$ is $n \times n$ and Hermitian. This is called a polar decomposition of $A$. Note that $A^tA = P^2$. If $m \leq n$ a polar decomposition of $A$ is obtained by the adjoint of a polar decomposition of $A^t$. This has the factors in the opposite order; if $A$ is square there are polar decompositions in either order.

13. Compound matrices. In this section matrices are over a commutative ring $R$. Let $A = \{1, \ldots, a\}$, $B = \{1, \ldots, b\}$, $C = \{1, \ldots, c\}$. Let $M$ be an $a \times b$ matrix and $N$ a $b \times c$ one. The subsets of $A$ (or $B,C$) of a given size are ordered “lexicographically”, that is, where $S_1$ precedes $S_2$ if, in the first place where the sequences in increasing order of the elements differ, the $S_1$ entry is lesser (although any fixed order would suffice). As in appendix 2 for $S \subseteq A$, $T \subseteq B$ let $M_{ST}$ be the “minor” with rows $S,T$, that is, where $M_{ST}[i,j] = M[\rho_S(i), \rho_T(j)]$ where $\rho_S, \rho_T$ enumerate $S,T$ in increasing order.

We may consider the entries $M_{ij}$ to be indeterminates. For square $M$ det$(M)$ is then a polynomial in the $M_{ij}$ with coefficients in $R$. If entries from a commutative ring $R$ are substituted for the $M_{ij}$, the determinant of the resulting matrix is the value of the polynomial. Identities involving determinants may be considered as polynomial identities.

Lemma 19. For $b = a$ the polynomial det$(M)$ is irreducible.
Proof: Let \( p \) be the polynomial, and suppose it factors as \( qr \) where \( q \) contains \( M_{ij} \) in some term and \( r = M_{kl} \). Clearly \( i \neq k, j \neq l \), whence \( M_{il} \neq q, r \), a contradiction.

Lemma 20. If \( c = a \leq b \) then
\[
\det(MN) = \sum_{U \subseteq B, |U| = a} \det(M_{U}) \det(M_{U^c}).
\]

Proof: Using notation of appendix 2,
\[
\det(MN) = \sum_{\sigma} \sgn(\sigma) \prod_{1 \leq l \leq a} (MN)_{l,\sigma(l)} = \sum_{\sigma} \sgn(\sigma) \prod_{1 \leq l \leq a} M_{tr} N_{r,\sigma(l)}
\]
\[
= \sum_{\sigma} \sgn(\sigma) \sum_{\mu : A \rightarrow B} \prod_{1 \leq l \leq a} M_{l,\mu(l)} N_{\mu(l),\sigma(l)}
\]
\[
= \sum_{U \subseteq B, |U| = k} \sum_{\sigma} \prod_{1 \leq l \leq a} M_{U}[l, \pi(l)] N_{U,A}[\pi(l), \sigma(l)].
\]

But it is easily seen that the sum over \( \sigma \) equals \( \det(M_{U}) \det(N_{U,A}) \) (cf. appendix 2).

For \( 0 \leq k \leq a, b \) let \( M^{(k)} \) be the \( \binom{a}{k} \) by \( \binom{b}{k} \) matrix where \( M^{(k)}_{ST} = \det(M_{ST}) \), where we consider the determinant of the empty matrix to be 1. This matrix is called the \( k \)-th compound matrix of \( M \). Its entries may be considered polynomials in the \( M_{ij} \). A minor \( M_{SS} \) is called principal. The determinants of the \( k \)-th order principal minors are the diagonal entries of \( M^{(k)} \), and \( \text{Tr}(M^{(k)}) \) is their sum. Let \( M^{[k]} \) be the \( \binom{a}{k} \) by \( \binom{b}{k} \) matrix where \( M^{[k]}_{ST} = (-1)^{S^{T}} \sum \det(M_{T-S'}) \) where \( \sum S \) denotes the sum of the elements of \( S \). This matrix is called the \( k \)-th compound conjugate of \( M \). Note that \( M^{(1)} \) is just the adjugate.

Theorem 21.

a. If \( b = a \) then \( M^{(k)} M^{[k]} = M^{[k]} M^{(k)} = \det(M) I \).

b. (Sylvester) If \( b = a \) then \( \det(M^{(k)}) = \det(M)^{\binom{k}{k} - 1} \) and \( \det(M^{[k]}) = \det(M)^{\binom{a}{a} - 1} \).

c. (Binet-Cauchy) If \( k \leq a, b, c \) then \( (MN)^{(k)} = M^{(k)} N^{(k)} \).

d. If \( k \leq b, c \) then \( (MN)^{[k]} = N^{[k]} M^{[k]} \).

e. (Jacobi) If \( b = a \) and \( 1 \leq k \leq a \) then \( (M^{(k)})^{(k)} = \det(M)^{k-1} M^{[k]} \).

f. If \( b = a \), in the characteristic polynomial for \( M \) the coefficient of \( x^k \) equals \( (-1)^{a-k} \det(M^{(a-k)}) \) for \( 0 \leq k \leq a \).

Proof: We leave part a as an exercise; see appendix 2. For part b, taking determinants of part a and using lemma 19 we see that \( \det(M^{(k)}) \) is a power of \( \det(M) \). The degrees of the polynomial entries are \( \binom{a}{k} \) and \( a \) respectively, and the first identity follows. The second is similar. For part c, \( (MN)^{(k)}_{ST} = \det((MN)_{ST}) \), and
\[
(M^{(k)} N^{[k]})_{ST} = \sum_{U \subseteq B, |U| = k} M^{(k)}_{SU} N^{[k]}_{U^c} = \sum_{U \subseteq B, |U| = k} \det(M_{SU}) \det(N_{U^c}).
\]

These are equal by lemma 20. Now,
\[
M^{(k)} N^{[k]} M^{[k]} = \det(M) \det(N) I = \det(MN) I = (MN)^{(k)} (MN)^{[k]};
\]
using part c and canceling polynomials yields the claim. For part e, take the \( k \)-th compound of part a with \( k = 1 \), apply part c to this, multiply part a for \( k \) by \( \det(M^{k-1}) \), and cancel. For part f, if \( S \subseteq A \) the terms of \( \det(xI - M) \) involving \( \prod_{i \in S} x - M_{ii} \) contribute \( \det(-M_{S-S'}) \) to the coefficient of \( x^k \).
Exercises.

1. Suppose \( S \) is an orthogonal set in a vector space with a symmetric or Hermitian form, and that no \( v \in S \) is null. Show that \( S \) is linearly independent.

2. Show that if \( C_q \) is the companion matrix to the monic polynomial \( q \) then \( q(C_q) = 0 \). Hint: Let \( X \) be the column matrix where \( X_i = x^i, \) \( 0 \leq i \leq m - 1, \) \( m = \deg(q) \). Over \( F[X], \) \( C_q X \equiv xX \mod q(x) \), whence \( q(C_q)X \equiv 0 \mod q(x) \). The degree of every polynomial on the left is less that \( m \).

3. Show directly that a \( n \times n \) matrix \( M \) over a field is invertible iff \( p_M \) is not divisible by \( x \). Hint: Write \( p_M \) as \( a_0 + xq \); if \( a_0 \neq 0 \) construct an inverse, and if \( a_0 = 0 \) derive a contradiction from assuming that \( M^{-1} \) exists.

4. Suppose \( M \) is \( m \times n \), \( L \) \( m \times m \), and \( R \) \( n \times n \). Show that (over a field) the rank of \( LM \) is at most the rank of \( M \), and equal to it if \( L \) is nonsingular; similarly for \( MR \).

5. Show directly that vectors from distinct eigenspaces of a linear map are linearly independent. Hint: Let \( b_1v_1 + \cdots + b_kv_k \) be a nontrivial linear combination equaling 0, with \( k \) minimal. Apply \( B \) to the linear combination, multiply it by \( \alpha_1 \), and subtract.

6. Show that the map \( \alpha \mapsto \alpha \) is an isomorphism from \( L_R(M,M;R) \) (or \( L_R(M_\sigma,M;R) \)) to \( L_R(M;M^*) \); the inverse map takes \( \alpha \) to the form \( \alpha(y)(x) \).

7. Suppose the characteristic of \( F \) is not 2; show the following.
   a. If \( * \) is symmetric then \( \rho_r(T) \) is alternating iff \( T^\dagger = -T \), in which case \(- (y*T(x)) = T(y)*x = x*T(y) = - (T(x)*y) \).
   b. If \( * \) is alternating then \( \rho_r(T) \) is alternating iff \( T^\dagger = T \), in which case \(- (y*T(x)) = - (T(y)*x) = x*T(y) = T(x)*y \).

8. Suppose \( F \) is weakly \( C \)-like and \( * \) is Hermitian. Show that \( T \) is Hermitian iff \( x*T(x) \in F_r \) for all \( x \). Hint: in the converse direction, consider \( w*T(w) \) with \( w = x,y,x+y,x+iy \).

9. Let \( q = p^c \) where \( p \) is a prime other than 2; let \( F = F_q^2 \) and \( x* = x^q \), so that \( F_r = F_q \). Suppose \( * \) is a Hermitian form on \( F^n \). Show that if \( x*x = 0 \) on some subspace \( W \) then \( x*y = 0 \) on \( W \).

10. If \( u_1, \ldots, u_r \) are orthonormal vectors let \( Q_{u_1 \ldots u_r} = I - 2 \sum u_i u_i^t \). Show that a linear operator on \( \mathbb{R}^n \) is symmetric and orthogonal iff it is a \( Q_{u_1 \ldots u_r} \). Show that \( Q_u Q_v = Q_v Q_u \) iff \( u^t v = 0 \), in which case \( Q_{uv} = Q_u Q_v \).
11. Model theory.

1. Syntax and semantics. We have seen that various categories of algebraic structures may be specified as those satisfying a set of axioms. It is useful to have more formal definitions. These are the starting point of the branch of mathematics known as model theory, which has found applications in algebra. First, a formal definition must be given of an axiom; for this we need the notion of a formal language.

Let \( A \) be a set, called the alphabet; its members are to be thought of as symbols, such as +, 2, \&, etc. \( A \) may be infinite; for example it might contain infinitely many variables \( x_0, x_1, \ldots \). In some contexts (for example if the symbols are to be read by an algorithm) the alphabet must be finite. In such cases, the infinite alphabet is merely a convenience, and may be transformed to a finite one by obvious methods, such as using the strings \( x, x', x'', \ldots \) for the variables. In other cases, such as when the alphabet contains a symbol for each member of a structure, this is not so.

A string over \( A \) is a finite sequence of members of \( A \); in full set theoretic formalism this is a function from some interval \( \{ i \in \mathbb{N} : i < n \} \) to \( A \). We write \( a_0 \cdots a_{n-1} \) to denote such a string; \( n \) is called the length. The set of strings over \( A \) is denoted \( A^* \). It contains a unique string of length 0, called the null string. There is a binary operation, concatenation, defined on \( A^* \); if \( a = a_0 \cdots a_{n-1} \) and \( b = b_0 \cdots b_{m-1} \) then \( ab = a_0 \cdots a_{n-1}b_0 \cdots b_{m-1} \). \( A^* \) forms a monoid with this operation, where the identity is the null string. This monoid obeys the left and right cancellation laws; that is, if \( xy = xz \) or \( yx = zx \) then \( y = z \). These obvious facts about strings follow readily from the formal definition.

A formal language over \( A \) is a finite subset of \( A^* \). The statements of mathematics can be defined as a formal language; doing so is an essential step in the development of mathematical logic. It has proved useful to consider the alphabet to depend on the context; it has a fixed part, called the logical symbols, and a variable part, called the nonlogical symbols. The logical symbols are as follows.

- propositional connectives \( \neg \land \lor \Rightarrow \Leftrightarrow \)
- quantifiers \( \forall \exists \)
- variables \( x, y, \) etc.
- punctuation marks (,)

The propositional connectives \( \top \) (true) and \( \bot \) (false) will also be used.

A propositional connective combines two statements to produce a new one; the truth or falsity of the compound statement depends only on the truth or falsity of its constituents. We use \( \top \) and \( \bot \) to denote the truth values, as well as the symbols representing them (often 1 and 0 are used). Similarly the propositional connectives are used to represent the function on truth values (“Boolean function”) that they express; these are as follows.

<table>
<thead>
<tr>
<th>( S )</th>
<th>( T )</th>
<th>( S \land T )</th>
<th>( S \lor T )</th>
<th>( S \Rightarrow T )</th>
<th>( S \Leftrightarrow T )</th>
<th>( S \oplus T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \top )</td>
<td>( \top )</td>
<td>( \top )</td>
<td>( \top )</td>
<td>( \top )</td>
<td>( \top )</td>
<td>( \bot )</td>
</tr>
<tr>
<td>( \top )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \top )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
</tr>
<tr>
<td>( \bot )</td>
<td>( \top )</td>
<td>( \bot )</td>
<td>( \top )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \top )</td>
</tr>
<tr>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
</tr>
</tbody>
</table>

The nonlogical symbols consist of
- a set of constant symbols;
- a set of function symbols, each with an arity;
- a set of relation symbols, each with an arity.

The intention is that the language applies for structures of a certain type. The nonlogical symbols will include symbols for the constants, functions, and relations of the type. We will primarily be interested in
the case of a finite set of nonlogical symbols, although many facts hold for an infinite set, and such will be considered when necessary.

In logic, the term “first order language” is used to refer to the non-logical symbols. A first order language may be “expanded” by adding symbols. As mentioned above often a constant symbol for each element of a structure under consideration is added. Another example of expanding the language is expansion by definitions; a symbol might be added for a constant, function, or relation which can be defined using the original language (we will give the formal definition of “defined” below). There are other common expansions which are useful in formal logic and model theory.

The strings in the formal language of mathematical logic are called well formed formulas, or simply formulas. These are most readily defined by a “recursive” definition. The nature of such definitions is obvious, but a more detailed discussion is worth the digression.

Let \( A \) be an alphabet, called the “terminal” symbols; augment this with an additional (disjoint) alphabet \( N \), called “nonterminal” symbols. A “production” is a pair \( X = \sigma \) where \( X \in N \) and \( \sigma \in (A \cup N)^* \). If \( A \), \( N \), and the set \( P \) of productions are finite, the triple \( A, N, P \) is called a “context free grammar”; however we allow infinite sets.

The idea is that each nonterminal \( X \) represents a syntactic category, and corresponds to a language \( L(X) \subseteq A^* \). A string \( x \) is in \( L(X) \) if there is a production \( X = \sigma \), and recursively for each occurrence \( Y_i \) of a nonterminal in \( \sigma \) a string \( y_i \in L(Y_i) \), such that when \( y_i \) is replaced by \( y_i \) in \( \sigma \) the result is \( x \). The production builds up a new instance of \( X \) from instances of the \( Y_i \) in \( \sigma \). We may define \( L_0(X) = \emptyset \) for all \( X \in N \); and recursively \( L_{i+1}(X) \) to be the strings derivable by some production using strings in the \( L_i(Y) \), \( Y \in N \); then \( L(X) = \cup_i L_i(X) \).

The recursive definition of the formulas of first order logic is as follows; nonterminals are underlined.

1. \( \text{term} = c \) where \( c \) is a constant.
2. \( \text{term} = x \) where \( x \) is a variable.
3. \( \text{term} = f(\text{term}, \ldots, \text{term}) \) where \( f \) is an \( n \)-ary function and there are \( n \) terms between the parentheses.
4. \( \text{formula} = R(\text{term}, \ldots, \text{term}) \) where \( R \) is an \( n \)-ary relation and there are \( n \) terms between the parentheses.
5. \( \text{formula} = \text{atomic formula} \)
6. \( \text{formula} = \exists \text{variable}(\text{formula}) \)
7. \( \text{formula} = \forall \text{variable}(\text{formula}) \)
8. \( \text{formula} = \neg(\text{formula}) \)
9. \( \text{formula} = (\text{formula}) \land (\text{formula}) \), and similarly for \( \lor \Rightarrow \Leftrightarrow \).

An important property of this specification is that there is only one way in which a formula may be built up. That is, for any \( x \) in any \( L(X) \) there is a unique production and unique \( y_i \in L(Y_i) \) such that \( x \) is derived from the \( y_i \) by the production. This property is referred to as “unique readability”; in the case of a context free grammar, the grammar is said to be “unambiguous”. The proof is left as an exercise. The importance of unique readability resides in the fact that functions may be defined recursively on the strings in the language; since there is only one way to build up the string, the value of the function will be well defined by applying the appropriate case of the recursive definition at each step.

The specification above retains unique readability if it is relaxed to allow omitting some pairs of parentheses. In cases 6-9 the parentheses may be omitted around a formula on the right if it is atomic, or begins with a quantifier or \( \neg \). Further parentheses are often omitted in writing formulas; the specification no longer has the unique readability property, and one of the alternatives is selected according to some rules. For example, \( F \lor G \lor H \) might be interpreted as \( F \lor (G \lor H) \), and \( F \land G \lor H \) as \( (F \land G) \lor H \).

In first order logic the fundamental function on formulas we wish to define is the meaning of a formula in a structure. This requires some preliminary definitions, which have other uses as well; the first of these...
is free and bound variables. All occurrences of variables in an atomic formula are free. In \( \exists x(F) \) or \( \forall x(F) \) any free occurrence of \( x \) in \( F \) becomes bound, by the quantifier; all other occurrences are free or bound as they are in \( F \). In a propositional combination of formulas, all occurrences of variables are free or bound as they are in the constituent subformulas.

The second definition is that of a structure for a first order language \( L \). In this chapter we will assume that a structure \( S \) is nonempty. \( S \) is a structure for \( L \) if each constant \( c \), function \( f \), or relation \( R \) of the language is interpreted as some constant \( c_S \), function \( f_S \), or relation \( R_S \) of the structure. The third definition is that of an assignment; if \( S \) is a structure than an assignment in \( S \) is a function whose domain is a finite set of variables and whose codomain is \( S \).

Suppose now that \( S \) is a structure for a first order language \( L \); \( t \) is a term over \( L \); and \( \nu \) is an assignment in \( S \) whose domain includes the variables of \( t \). The definition of \( [t]_{S\nu} \) (abbreviated \( [t] \)) is given recursively as follows. If \( t \) is a constant \( c \) then \( [t] = c_S \). If \( t \) is a variable \( x \) then \( [t] = \nu(x) \). If \( t = f(t_1, \ldots, t_n) \) where \( t_i \) is a term then \( [t] = f_S([t_1], \ldots, [t_n]) \).

Now we can define \( [F]_{S\nu} \) (abbreviated \( [F] \)) recursively, where the domain of \( \nu \) includes the free variables of \( S \); this will be a truth value. If \( F \) is an atomic formula \( R(t_1, \ldots, t_n) \) then \( [F] = R_S([t_1], \ldots, [t_n]) \). If \( F = \exists xG \) (resp. \( \forall xG \)) then \( [F]_{S\nu} \) is true iff \( [G]_{S\nu} \) is true for some (resp. all) \( \tau \) whose domain is that of \( \nu \), together with \( x \), and which assigns to variables other than \( x \) the same value as \( \nu \). If \( F = \neg G \) then \( [F] = \neg [G] \), and similarly for the other propositional connectives. By this we mean, apply the Boolean function \( \neg \), for which we use the same symbol as we use for its denotation in the formal language.

It is an easy induction to show that \( [F] \) is the same for all assignments which have given values on the free variables of \( [F] \). In particular if \( F \) contains no free variables then \( [F]_{S\nu} \) does not depend on \( \nu \), and may be denoted \( [F]_S \). A formula containing no free variables is called closed, or a sentence. If the language is expanded with a name for every element of the structure \( S \), an easy induction shows that \( [F]_{S\nu} \) will equal \( [F']_S \) where \( F' \) is \( F \), with free occurrences of \( x \) replaced by the symbol for \( \nu(x) \). Many authors give the recursive definition using this language and omit assignments altogether. Note that by \( [F]_S \) we mean \( [F']_S \) where \( S' \) is the structure \( S \) “expanded” by interpreting the name for each \( a \in S \) as \( a \).

A convenient notation is to write a formula as \( F(x_1, \ldots, x_n) \), to indicate that the argument variables are taken to be \( x_1, \ldots, x_n \) in order. It is understood that the free variables of \( F \) are among the \( x_i \), but it is not required that the \( x_i \) all occur in \( F \). \( F(a_1, \ldots, a_n) \) may then be written for the formula of the expanded language where \( x_i \) is replaced by \( a_i \). More generally \( F(t_1, \ldots, t_n) \) may be written, where the \( t_i \) are terms, possibly containing variables. In this case, it is understood that the variables of \( t_i \) do not become bound by any quantifier of \( F \) when an occurrence of \( x_i \) is replaced by \( t_i \), in which case \( t_i \) is said to be free for \( x_i \) in \( F \).

It is easily shown that in this case, \( [F(t_1, \ldots, t_n)]_{S\nu} = [F(x_1, \ldots, x_n)]_{S\nu'} \) where \( \nu'(x_i) = [t_i]_{S\nu} \). The formula \( F(t_1, \ldots, t_n) \) is called an instance of the formula \( F(x_1, \ldots, x_n) \).

In model theory the symbol \( = \) is usually assumed to be among the binary relation symbols, and \( =_S \) is assumed to be equality in any structure \( S \). When this is done \( = \) is often taken to be a logical symbol. This point will be discussed further in the next section. An equation is an atomic formula where the relation is equality.

The notion of a first order language has a generalization to that of a multisorted first order language; we consider only finitely many sorts. For each sort there are an infinite collection of variables. The same quantifiers can be used. The relation and function symbols have the sort of each argument position associated with them; the function symbols have a sort for the value. It should be clear how a formula of a multisorted language, and its meaning in a multisorted structure of the appropriate type, are defined.

A formula is called open if it contains no quantifiers, that is, if it is a propositional combination of atomic formulas. It is a straightforward induction to verify that, if \( T \subseteq S \) is a substructure and \( \nu \) is an assignment whose codomain is \( T \), then for any term \( t \) \( [t]_{S\nu} = [t]_{T\nu} \), and for any open formula \( F \) \( [F]_{S\nu} = [F]_{T\nu} \).
By a similar induction one also verifies that if $S = S_1 \times \cdots \times S_n$ is a product structure, then for any term $t$ \([t]_{S}\nu = ([t]_{S_{\nu_1}}, \ldots, [t]_{S_{\nu_n}})\); and for any atomic formula $F$ \([F]_{S\nu} = [F]_{S_{\nu_1} \land \cdots \land [F]_{S_{\nu_n}}}, \) where $\nu(x) = (\nu_1(x), \ldots, \nu_n(x))$ (recall that this is true for $F = R(x_1, \ldots, x_n)$ by definition). This generalizes to infinite products, the “infinite conjunction” being true iff each of the conjuncts is.

A formula $H_1 \Rightarrow \cdots \Rightarrow (H_k \Rightarrow C) \cdots$ may be abbreviated $H_1 \Rightarrow \cdots \Rightarrow H_k \Rightarrow C$. This is true if, whenever the hypotheses $H_i$ are true the conclusion $C$ is. Define such a formula to be a Horn formula if the $H_i$ and $C$ are atomic; the case $k = 0$ is allowed.

**Theorem 1.** If a universally quantified open formula is true in a structure it is true in any substructure. If a universally quantified Horn formula is true in every factor structure it is true in the product structure.

**Proof:** If the open formula $F$ is true everywhere in the structure, since the truth value at each assignment does not change in passing to the substructure, $F$ is true everywhere in the substructure. For products, if the hypotheses are true in the product they are true in each factor, whence the conclusion is true in each factor, whence the conclusion is true in the product.

Thus we have proved the claims made in chapter 2 about axioms holding in substructures or products; in fact we have proved more. Since equations are Horn formulas, a class of distinguished structures which has a set of equations as axioms is closed under substructure and product. Such classes include semigroups, monoids, groups, rings, commutative rings, $R$-modules, and $R$-algebras. Partially ordered sets have Horn formulas for axioms, so they are closed under substructure and product, as are semilattices, lattices, and Boolean algebras. Integral domains are closed under substructure but not product; the axiom $xy = 0 \Rightarrow (x = 0 \vee y = 0)$ is not a Horn formula, and indeed may be false in the product but true in the factors. Fields are not closed under substructure; indeed multiplicative inverse cannot be taken as a function symbol.

There is another “preservation” property which is sometimes useful. Define a formula to be positive if it involves only the logical symbols $\land \lor \forall \exists$. If $h : S \to T$ is an epimorphism of structures, $F(x_1, \ldots, x_k)$ is a positive formula, $a_i \in S$, and $F(a_1, \ldots, a_k)$ is true in $S$, then $F(h(a_1), \ldots, h(a_k))$ is true in $T$. The proof is by induction; for $\forall x F(x, a_1, \ldots, a_k)$, for any $b \in T$ choose $a$ with $b = h(a)$ and apply the induction hypothesis. The other logical symbols are straightforward. In particular if an equation (indeed a positive open formula) is true everywhere in $S$ then it is true everywhere in $T$.

**2. Validity.** The material in this section is more a topic in mathematical logic than model theory, though the main construction is sufficiently basic that it is worth giving. It has uses in model theory [CK]. A structure $S$ is said to be a model of, or satisfy, a sentence $F$ (of the expanded language) if \([F] = T\); the notation $|=S F$ is used for this. It is convenient to generalize the notion of satisfaction; for any formula $F$ we write $|=S F$ if $|=S \overline{F}$ where $\overline{F}$ is $F$ preceded by universal quantifiers for all its free variables. A formula $F$ is called valid if $|=S F$ for any structure $S$ (of the given type); we write $|= F$ for this. A formula $F$ is said to be a logical consequence of a set of formulas $T$, written $T \models F$, if $|=S F$ for any $S$ where $|=S G$ for all $G \in T$.

**Lemma 2.** Let $F, G, H$ be formulas, $x$ a variable, and $t$ a term.

a. $|= F \Rightarrow (G \Rightarrow F)$.

b. $|= (H \Rightarrow (F \Rightarrow G)) \Rightarrow ((H \Rightarrow F) \Rightarrow (H \Rightarrow G))$.

c. $|= (\neg F \Rightarrow G) \Rightarrow ((\neg F \Rightarrow \neg G) \Rightarrow F)$.

d. $|= \neg \bot$.

e. $F, F \Rightarrow G \models G$.

f. $F \Rightarrow \forall x G(x) |\models F \Rightarrow G(t)$, provided $t$ is free for $x$ in $G$.

g. $F \Rightarrow G \models F \Rightarrow \forall x G$, provided $x$ does not occur free in $F$.

**Proof:** Exercise.
We define a derivation system to consist of a set of rules $G_1, \ldots, G_n \Rightarrow F$, where $G_i, F$ are strings over some alphabet. A rule where $n = 0$ is called an axiom. A derivation is a sequence $F_1, \ldots, F_i$ of strings, such that for every $i$ there are $i_1, \ldots, i_n < i$ such that $F_{i_1}, \ldots, F_{i_n} \Rightarrow F_i$ is a rule ($n = 0$ is allowed, if $F_i$ is an axiom). If in addition $F_i$ may be a member of some set $T$ we call the sequence a derivation from $T$. $F$ is called derivable (from $T$) if it is the last formula of some derivation (from $T$), and we write $\vdash F$ ($T \vdash F$). If the derivation system $D$ is specified, we write $\vdash_D$.

Replacing $|$ by $\vdash$ in the formulas of lemma 2 yields a derivation system, which we call first order logic, or FOL for short. Note that the logical symbols of FOL are $\bot, \neg, \Rightarrow, \forall$. The converse statement is called the completeness theorem for first order logic. Its proof requires some preliminary lemmas, of interest in themselves.

Let PC denote the derivation system consisting of rules a-e of lemma 2. This can be viewed as a system for strings involving the "atoms" $F, G, \ldots$, and the propositional connectives; we call such strings PC-formulas. A PC-formula is called PC-valid if it is true for any assignment of true or false to the atoms; we write $\vdash_{PC} F$ for this, and $S \vdash_{PC} F$ for PC-logical consequence, whose definition we leave to the reader. One easily sees that if a PC-formula is PC-valid then any formula obtained from it by replacing atoms by formulas is valid.

**Lemma 3.** $\vdash_{PC} F \Rightarrow F$.

**Proof:** We have

\[
(F \Rightarrow ((F \Rightarrow F) \Rightarrow F)) \Rightarrow ((F \Rightarrow (F \Rightarrow F)) \Rightarrow (F \Rightarrow F)) \quad b
\]
\[
F \Rightarrow ((F \Rightarrow F) \Rightarrow F) \quad a
\]
\[
(F \Rightarrow (F \Rightarrow F)) \Rightarrow (F \Rightarrow F) \quad d
\]
\[
F \Rightarrow (F \Rightarrow F) \quad a
\]
\[
F \Rightarrow F \quad d
\]

**Lemma 4.** If $S, H \vdash_{PC} G$ then $S \vdash_{PC} H \Rightarrow G$.

**Proof:** The proof is by induction on the length of the first derivation. If $G = H$ the claim follows by lemma 3. If $G$ is an axiom the claim follows since $\vdash G \Rightarrow (H \Rightarrow G)$. If $G$ follows by rule e then inductively $S \vdash H \Rightarrow F$ and $S \vdash H \Rightarrow (F \Rightarrow G)$, and $S \vdash H \Rightarrow G$ follows using axiom b and rule e.

Lemma 4 is called the deduction theorem, for the axiom system PC. Suppose $T$ is a set of PC-formulas. Say that $T$ is PC-consistent if $\not\vdash_{PC} \bot$. An extension of $T$ is a superset, in some specified set of atoms which include those of $T$. Call $T$ maximal PC-consistent if in addition no proper extension of $T$ is PC-consistent.

If there is a derivation of $\not\vdash_{PC} \bot$ from $T$ then there is one which uses no new atoms; simply replace any new atoms in such by original ones. The notion of a maximal consistent extension does require specifying the set of atoms, though.

**Lemma 5.** A PC-consistent set $T$ of PC-formulas can be extended to a maximal PC-consistent set $U$ (in a specified set of atoms).
Proof: Given a chain of PC-consistent extensions of $T$, let $S$ be the union of the chain. Then $S$ is consistent; for if there were a derivation of $\bot$ from some formulas of $S$ the formulas would all be in some member of the chain. By the maximal principle there is a maximal set $U$ among the PC-consistent extensions of $T$.

Lemma 6.  

a. $F, \neg F \vdash_{PC} G$

b. $F \Rightarrow G, \neg F \Rightarrow G \vdash_{PC} G$

c. $F, \neg G \vdash_{PC} \neg (F \Rightarrow G)$

Proof: For part a the derivation is

$$F, \neg F, F \Rightarrow (\neg G \Rightarrow F), \neg G \Rightarrow F, \neg F \Rightarrow (\neg G \Rightarrow \neg F), \neg G \Rightarrow \neg F, G,$$

the last step using axiom c. For part b, first, $\neg \neg F \vdash F$; the derivation is

$$\neg \neg F, \neg \neg F \Rightarrow (\neg F \Rightarrow \neg F), \neg F \Rightarrow \neg \neg F, \neg F \Rightarrow \neg \neg F \Rightarrow ((\neg F \Rightarrow \neg F) \Rightarrow F),$$

$$(\neg F \Rightarrow \neg F) \Rightarrow F, \neg F \Rightarrow \neg F, F.$$

We then have $F \Rightarrow G, \neg F \vdash G$, so $F \Rightarrow G \vdash \neg F \Rightarrow G$. But then $F \Rightarrow G, \neg G \vdash \neg F \Rightarrow G, \neg F \Rightarrow \neg F \Rightarrow \neg F$, whence $F \Rightarrow G \vdash G \Rightarrow \neg F$. Also $\neg F \Rightarrow G \vdash \neg F \Rightarrow \neg F \Rightarrow G \Rightarrow F$. Thus, $F \Rightarrow G, \neg F \Rightarrow G \vdash \neg G \Rightarrow \neg F, \neg G \Rightarrow G \vdash F$, proving part b. Clearly $F \vdash (F \Rightarrow G) \Rightarrow G$; hence $F \vdash \neg G \Rightarrow \neg (F \Rightarrow G)$, and part c follows.

Lemma 7. Let $T$ be a set of PC-formulas and $F, G$ PC-formulas.

a. If $T$ is PC-consistent then either $T \cup \{F\}$ or $T \cup \{\neg F\}$ is.

b. If $T$ is maximal PC-consistent then either $F \in T$ or $\neg F \in T$ but not both.

c. If $T$ is maximal PC-consistent then $\neg F \in U$ if $F \notin U$ or $G \in U$.

Proof: For part a, if $T, F \vdash \bot$ and $T, \neg F \vdash \bot$ then $T \vdash \bot$ follows by lemmas 4 and 6. For part b, by part a one of $F, \neg F$ is in $T$, and both cannot be by lemma 6. Part c follows using part b, lemma 6, and axiom a.

Say that $T$ PC-satisfiable if there is an assignment of true or false to the atoms which makes every $F \in T$ true. If $T$ is PC-satisfiable then $T$ is PC-consistent; this follows because (as is readily verified) a truth assignment which satisfies $T$ can be extended to satisfy any formula derivable from $T$.

Theorem 8. If $T$ is PC-consistent then $T$ is PC-satisfiable.

Proof: Let $U$ be a maximal consistent extension of $T$ (in the atoms of $T$). For each atom $F$, assign true to $F$ if $F \in U$, else assign false. We claim that under this assignment, the truth value $[H]$ of any formula $H$ equals $\top$ iff $H \in U$. The claim is clear if $H$ is an atom by lemma 7, and if $H$ is $\bot$ by axiom d and lemma 7. If $H$ is $\neg A$, inductively $[\neg A] = \top$ iff $[A] = \bot$ iff $A \notin U$ iff $\neg A \in U$. If $H$ is $A \Rightarrow B$ then $[A \Rightarrow B] = \top$ iff $[A] = \bot$ or $[B] = \top$, iff $A \notin U$ or $B \in U$; by lemma 7 this is so if $A \Rightarrow B \in U$.

Corollary 9. If $T \vdash_{PC} F$ then $T \vdash_{PC} F$, where $T$ is a set of PC-formulas and $F$ is a PC-formula.

Proof: First note that $\neg F \Rightarrow \bot \vdash F$, by axiom c, since $\neg F \Rightarrow \bot \Rightarrow \bot$ follows by axioms d and a. So, if $T \not\vdash F$ then $T \cup \{\neg F\}$ is PC-consistent, since otherwise $T \not\vdash \neg F \Rightarrow \bot$ whence $T \vdash F$. Thus if $T \not\vdash F$ then $T \cup \{\neg F\}$ is PC-satisfiable, so $T \not\vdash F$.  

93
Corollary 9 is of interest in its own right; and the proof of the corresponding fact for FOL follows the same pattern. An important consequence is compactness for PC, which states that if every finite subset of $T$ is PC-satisfiable then $T$ is; for if $T$ is not PC-satisfiable then $T \models_{PC} \bot$, so $G_1, \ldots, G_r \models_{PC} \bot$ for some finite \{ $G_1, \ldots, G_r$ \} $\subseteq T$, which is therefore not PC-satisfiable. As the reader might expect, compactness for PC can be proved without appealing to a derivation system; see the exercises. Also, corollary 9 can be proved for finite $T$ by finitary methods; again see the exercises.

As another variation, lemma 5 can be proved for countable sets of formulas by a more constructive argument. Let $F_0, F_1, \ldots$ be an enumeration of the PC-formulas. Let $U_0 = T$. If $U_i \cup \{ F_i \}$ is consistent let $U_{i+1} = U_i \cup \{ F_i \}$; otherwise let $U_{i+1} = U_i$. Let $U = \bigcup U_i$. Clearly each $U_i$ is consistent, and the consistency of $U$ follows. Further if $U \cup \{ F_i \}$ is consistent then $U_i \cup \{ F_i \}$ is, so $F_i \in U$; thus, $U$ is maximal consistent.

**Lemma 10.** If $T, F \vDash G$ then $T \vDash F \Rightarrow G$, provided that in the first derivation, no free variable of $F$ is involved in a use of rule $g$. In particular, this is so if $F$ is a sentence.

**Proof:** We need only add rules $f$ and $g$ to the proof of lemma 4. If $G \Rightarrow H(t)$ follows from $G \Rightarrow \forall x H(x)$ then $F \Rightarrow (G \Rightarrow H(t))$ follows from $F \Rightarrow (G \Rightarrow \forall x H(x))$ by propositional logic and rule $f$. If $G \Rightarrow \forall x H$ follows from $G \Rightarrow H$ then $F \Rightarrow (G \Rightarrow \forall x H)$ follows from $F \Rightarrow (G \Rightarrow H)$ by propositional logic and rule $g$, since by hypothesis $x$ is not free in $F$.

Lemma 10 is the deduction theorem for first order logic. Define the set $T$ of FOL formulas to be consistent set of formulas if $T \not\vDash \bot$. The language is assumed to be specified; but as in the propositional case consistency does not depend on it, although the proof is more involved. A simple case is expansion by constants, where new constant symbols are added. In a proof of $\bot$ in the expanded language, the new constants may be replaced by distinct variables which do not occur in the proof. In a specified language, a consistent set of formulas is maximal consistent if no proper extension is consistent.

One refinement must be made to the method. A witness for a formula $\neg \forall x F(x)$ is defined to be a formula $\neg F(c)$ for some constant $c$.

**Lemma 11.** A consistent set $T$ of formulas is a subset of a maximal consistent set $U$. Further, there is a maximal consistent set $V$ in a language expanded with constants, where every formula $\neg \forall x F(x) \in V$ has a witness in $V$.

**Proof:** The existence of $U$ follows by the same argument as Lemma 5. Given a set $S$ of formulas over some language, for each sentence $\neg \forall x F(x)$ of the language add to $S$ a sentence $\neg \forall x F(x) \Rightarrow \neg F(c)$, where the constants $c$ are new and distinct; call the result $S'$. We claim that $S'$ is consistent if $S$ is, so suppose $X, W_1, \ldots, W_k \vdash \bot$ where $X \subseteq S$ and $W_i$ is $\neg \forall x F_i(x) \Rightarrow \neg F(c_i)$. Then $X \vdash W_1 \Rightarrow \cdots \Rightarrow W_k \Rightarrow \bot$; $c_i$ may be replaced by a new variable $y_i$, and it follows that $X \vdash \bot$. Let $T_0 = T$, $T_{i+1} = T_i'$, and $T_w = \bigcup T_i$. Let $V$ be a maximal consistent set in the expanded language containing $T_w$. Since $V$ is maximal consistent and contains $\neg \forall x F(x) \Rightarrow \neg F(c)$, if it contains $\neg \forall x F(x)$ it contains $\neg F(c)$.

No restriction need be placed on the cardinality of the initial language, although as mentioned earlier for most purposes in this text this will be finite.

**Theorem 12.** If $T$ is consistent it has a model.

**Proof:** Let $U$ be a maximal consistent extension as in lemma 11. By a closed term or atomic formula we mean one containing no variables. $U$ is certainly maximal PC-consistent, so for every closed atomic formula $A$, exactly one of $A$ or $\neg A$ is in $U$. Define a structure $S$ whose elements are the closed terms, where $[f(t_1, \ldots, t_k)] = f(t_1, \ldots, t_k)$, and $[R(t_1, \ldots, t_k)]$ is $\top$ or $\bot$ according to whether or not $R(t_1, \ldots, t_k) \in U$. We claim that for a sentence $F$ in the expanded language, $\models_S F$ if $F \in U$. This follows by definition for atomic formulas, and for propositional combinations by induction as in theorem 7. Finally $\models_S \forall x G(x)$ if
\[ \models_S G(t) \text{ for all closed terms } t. \] If \( \neg \forall x G(x) \in U \) then by construction \( \neg G(c) \in U \) for some \( c \); on the other hand if \( \forall x G(x) \in U \) then \( G(t) \in U \) for all closed terms \( t \) since \( U \) is maximal consistent. Thus, \( \models_S \forall x G(x) \) iff \( \forall x G(x) \in U \). Now suppose \( F \in T \); then \( \notmodels F \), where \( \notmodels \) is \( F \) preceded by universal quantifiers for the free variables, whence \( \models_S \notmodels \), and so \( \models_S F \).

**Corollary 13.** If \( T \models F \) then \( T \notmodels F \).

**Proof:** Exactly as in corollary 9.

Corollary 13 is known as the completeness theorem for predicate logic. Define \( T \) to be satisfiable if it has a model. The compactness theorem is the statement that if every finite subset of a set \( T \) of formulas is satisfiable then \( T \) is; this follows by theorem 12. There are proofs which do not use a derivation system; see the exercises.

One example of a consequence of compactness is that there are “nonstandard” models of the true statements \( T \) of arithmetic. We occasionally use \( \mathcal{N} \) to denote the structure for a first order language, here \( 0, \text{Suc, +, } \times, = \) (where \( \text{Suc}(x) = x + 1 \)). Let \( T \) be the theory of \( \mathcal{N} \). If new constant \( \infty \) is added to the language, and \( T' \) is \( T \) with the axioms \( i < \infty \) added, every finite subset of \( T' \) has a model, namely \( \mathcal{N} \) with \( \infty \) interpreted as a sufficiently large integer. Thus, \( T' \) has a model, which is a model of \( T \) other than the integers.

The foregoing illustrates the limitations of first order logic. In the next chapter a system of axioms called Peano’s axioms will be defined. Their logical consequences are a (proper) subset of the formulas true in \( \mathcal{N} \). On the other hand, an “informal” version of these uniquely characterizes the structure \( \mathcal{N} \).

The completeness theorem can be used to show that a set \( S \) of formulas remains consistent if the language is expanded. Indeed, \( S \) has a model, and any expansion of the structure to the expanded language remains a model. A constructive proof can be given using Gentzen systems, q.v. see [Smullyan] for example.

If \( = \) is a strong congruence relation in a structure \( S \) a straightforward induction shows that if \( a_i \equiv b_i \) then \( F(a_1, \ldots, a_k) \Leftrightarrow F(b_1, \ldots, b_k) \) for any formula \( F(x_1, \ldots, x_k) \). The axioms of equality are

\[- x = x \]
\[- x = y \Rightarrow y = x \]
\[- x = y \Rightarrow y = z \Rightarrow x = z \]
\[- x_1 = y_1 \Rightarrow \cdots \Rightarrow x_k = y_k \Rightarrow f(x_1, \ldots, x_k) = f(y_1, \ldots, y_k) \]
\[- x_1 = y_1 \Rightarrow \cdots \Rightarrow x_k = y_k \Rightarrow R(x_1, \ldots, x_k) \Rightarrow R(y_1, \ldots, y_k) \]

In any structure \( S \) satisfying these, \( \models_S \) is a strong congruence relation. The quotient structure \( S/\equiv \) is a structure where \( \equiv \) is equality. Thus, theorem 12 can be strengthened to require \( \models \) to be equality if \( T \) contains the axioms of equality. One can check that if \( = \) is considered a logical symbol, the axioms of equality are added to the axioms, and \( \models \) is required to be equality, then the usual theorems, for example corollary 13, hold. Also, closure under substructure or product is not affected, since \( \models \) will still be equality.

**3. Basic definitions.** If \( T \) is a set of sentences then \( S \) is a model of \( T \), written \( \models_S T \), if \( \models_S F \) for each \( F \in T \). In this case \( S \) is a model of any logical consequence of \( T \). A set of sentences which is closed under logical consequence is called a theory. If \( T \) is a theory a subset \( A \subseteq T \) is called a set of axioms for \( T \) if \( T \) is the set of logical consequences of \( A \). Examples of theories include the following.

- Given a structure \( S \), the sentences (of the base language) true in \( S \) form a theory, called the theory of the structure \( S \).
- The set of sentences of the language expanded with constants for the elements of \( S \), which are true in (the expanded) \( S \), is also a theory, called the elementary diagram.
- If \( A \) is any set of sentences, the set \( T \) of logical consequences of \( A \) is a theory, which has \( A \) as a set of axioms. By abuse of language, we may call \( A \) the theory.
Note that a theory \( T \) is consistent if for any sentence \( F \) (in the language under consideration), at most one of \( F, \neg F \) is in \( T \). An arbitrary set \( A \) of sentences is consistent iff the set of its logical consequences is. A theory \( T \) is called complete if for any sentence \( F \) exactly one of \( \models S F \) or \( \models S \neg F \) holds, the theory of a structure is consistent and complete. A theory given by a consistent set of axioms may or may not be complete. It is readily seen for example that the theory of groups (or rings or fields) is not complete, but examples do occur; see chapter 12.

An embedding \( j : S \hookrightarrow T \) of structure \( S \) in structure \( T \) is called elementary if

\[
[F(a_1, \ldots, a_k)]_S = [F(j(a_1), \ldots, j(a_k))]_T
\]

for any formula \( F(x_1, \ldots, x_k) \) and elements \( a_1, \ldots, a_k \in S \). If the inclusion map \( S \subseteq T \) is elementary then \( S \) is called an elementary substructure of \( T \). It is easy to see that not all embeddings are elementary; any isomorphism is, however. Two structures are said to be elementarily equivalent if they have the same theories, i.e., satisfy the same sentences in the base language. Isomorphic structures are elementarily equivalent, and it is easy to see that the converse is not true.

**Theorem 14.** The embedding \( j : S \hookrightarrow T \) is elementary iff for any formula \( F(y, x_1, \ldots, x_k) \) and \( a_1, \ldots, a_k \in S \), if \( \models_T \exists y F(y, j(a_1), \ldots, j(a_k)) \) then \( \models_S \exists y F(y, a_1, \ldots, a_k) \).

**Proof:** One direction is obvious; for the other we show by induction that (1) holds. For atomic formulas it is readily verified that (1) holds for any homomorphism \( j \) (indeed this is so for open formulas). For propositional combinations the induction step is straightforward. For the existential quantifier, using the induction hypothesis, if an existentially quantified formula is true in \( S \) it is certainly true in \( T \); and if it is true in \( T \) it is true in \( S \) by hypothesis. Since the universal quantifier may be defined we are done (although a direct argument for it is straightforward).

A relation \( R \) on a structure \( S \) is said to be definable if there is a formula \( F(x_1, \ldots, x_n) \) such that \( R(a_1, \ldots, a_n) \Leftrightarrow [F(a_1, \ldots, a_n)]_S \). In some contexts, \( F \) may be allowed to contain names for elements of the structure, which are called parameters.

Given a first order language \( L \), the “term algebra” or “Herbrand universe” over \( L \) is defined to be the set of closed terms, that is, terms containing no variables. If \( L \) contains no constant symbol this is empty; and in some contexts the language is expanded with a constant symbol in this case. A “Herbrand structure” is a structure whose domain is the Herbrand universe, where the functions are interpreted in the obvious way, namely, \( [f](t_1, \ldots, t_n) = f(t_1, \ldots, t_n) \). A Herbrand structure is exactly the same as a truth assignment to the closed atomic formulas.

Given a formula \( F(x_1, \ldots, x_n) \), an instance \( F(t_1, \ldots, t_n) \) where the \( t_i \) are closed terms is called a ground instance. Given a set \( S \) of open formulas, let \( S' \) be the set of ground instances of formulas of \( S \). If \( S \) is satisfiable then \( S' \) is PC-satisfiable; simply assign each closed formula its truth value in the model. It is not difficult to see that the converse is also true; let \( S'' \) be a maximal PC-consistent extension of \( S' \), and consider the Herbrand structure determined by it. Call a set of formulas unsatisfiable if it has no model, and similarly for PC-unsatisfiable. By the preceding, and compactness for PC, we have proved Herbrand’s theorem, that if a set of open formulas is unsatisfiable then some finite set of ground instances is PC-unsatisfiable.

Let OFOL be the derivation system for open formulas, consisting of PC for open formulas (where the atoms are the atomic formulas) and the “substitution rule” \( F(x_1, \ldots, x_n) \rightarrow F(t_1, \ldots, t_n) \). We claim that if \( S \) is a set of open formulas, \( F \) is an open formula, and \( S \models F \), then \( S \vdash_{\text{OFOL}} F \). Indeed, let \( F' \) be \( F \) with its variables replaced by new constants; then \( S \cup \{ \neg F' \} \) is unsatisfiable, so for some finite set \( G_1, \ldots, G_k \) of ground instances (in the language expanded with the new constants) of formulas of \( S \), \( G_1, \ldots, G_k \vdash_{\text{PC}} F' \). Replacing the new constants by variables, we obtain an OFOL derivation of \( F \) from \( S \), indeed a PC derivation from instances of formulas of \( S \). As a corollary, an open formula is valid iff it is a tautology.
There is a classification of formulas by “quantifier complexity” which is useful in a variety of contexts. An open formula is both $\Sigma_0$ and $\Pi_0$. Inductively, a $\Pi_n$ formula preceded by existential quantifiers is $\Sigma_{n+1}$; and a $\Sigma_n$ formula preceded by universal quantifiers is $\Pi_{n+1}$. It is convenient to allow 0 quantifiers as the number in the inductive formation, so that a $\Sigma_n$ or $\Pi_n$ formula is $\Sigma_m$ or $\Pi_m$ for any $m > n$.

A formula is said to be in “prenex normal form” if it is in one of these classes. A procedure for transforming any formula to a logically equivalent formula in prenex normal form is easily given. A useful skill in mathematical logic is to do this so that the resulting formula has as low a quantifier complexity as possible. One pretty good general method is based on the observation that the conjunction or disjunction of $\Sigma_n$ ($\Pi_n$) formulas can be brought to $\Sigma_n$ ($\Pi_n$) form; and the negation of a $\Sigma_n$ ($\Pi_n$) formula can be brought to $\Pi_n$ ($\Sigma_n$) form.

4. Ideals and filters. In the next section, the notion of a filter in a Boolean algebra will be required. This is a special case of more general notions, which are given here. In a poset $P$, a subset $S$ such that $x \in S$ and $y \leq x$ imply $y \in S$ is called by various names; we use $\leq$-closed. The notion of a $\geq$-closed set is defined dually. In a preorder, let $x^\leq$ denote $\{w : w \leq x\}$, and dually for $x^\geq$. The following are readily verified.

- The $\leq$-closed sets form an algebraic closure system, as do the $\geq$-closed sets.
  - $S$ is $\leq$-closed iff $S^c$ is $\geq$-closed.
  - $x^\leq$ is $\leq$-closed, and $x^\geq$ is $\geq$-closed.
  - If $P$ has a least element 0 and $S$ is $\leq$-closed then $0 \in S$; dually 1 $\in S$ for a greatest element 1 and a $\geq$-closed set $S$.

In a poset $P$ an ideal is a directed $\leq$-closed nonempty set; dually a filter is a filtered $\geq$-closed nonempty set. It is easily seen that in a $\sqcup$-semilattice, an ideal is a $\leq$-closed, $\sqcup$-closed nonempty subset. Dually in a $\sqcap$-semilattice, a filter is a $\geq$-closed, $\sqcap$-closed nonempty subset. In a lattice, both characterizations hold. An ideal in a $\sqcup$-semilattice $P$ may also be characterized as a $\sqcup$-closed nonempty subset such that for $x \in S$ and $y \in P$, $x \sqcap y$ exists and is in $S$; in particular an ideal is a sub-$\sqcup$-semilattice. The dual characterization holds in a $\sqcap$-semilattice, and in a lattice both characterizations hold.

In a $\sqcup$-semilattice $P$ the following facts hold.

- If $0$ exists the ideals form an algebraic closure system; in general the ideals together with $\emptyset$ do. If $P$ is filtered the ideals are closed under finite intersection.
- $x^\leq$ is an ideal. Such an ideal is called principal; if $P$ is finite every ideal is principal.
- If $I$ is an ideal then $x \sqcup y \in I$ iff $x \in I$ and $y \in I$.
- If $I$ is the ideal generated by $S \subseteq P$ then $x \in I$ iff $x \leq s_1 \sqcup \cdots \sqcup s_k$ for some $s_1, \ldots, s_k \in S$. That is, $I$ is the $\leq$-closure of the $\sqcup$-closure of $S$.
- If $I$ is the join of a family $\{J_i\}$ of ideals in $P$ then $x \in I$ iff $x \leq j_1 \sqcup \cdots \sqcup j_k$ for some $j_1, \ldots, j_k$ where $j_i \in J_i$; the $i$ may be required to be distinct.
- $(x \sqcup y)^\leq = x^\leq \sqcup y^\leq$, where the join on the right is that of the ideals.

The dual facts hold in a $\sqcap$-semilattice, with filters rather than ideals; the last fact becomes

- $(x \sqcap y)^\geq = x^\geq \sqcap y^\geq$.

Suppose $S$ is a subset of a $\sqcup$-semilattice. Since the ideals and $\emptyset$ form an algebraic closure system there is a smallest ideal containing $S$. This can be described more explicitly as follows. Let $S_1$ be the join closure of $S$, that is, $\{\sqcup T : T \subseteq S, T$ finite$\}$. Let $F$ be the downward closure of $S_1$, that is, $\{y \in B : y \leq x, \text{ some } x \in S_1\}$. $F$ is readily verified to be the smallest ideal containing $S$. The smallest filter containing $S$ in a $\sqcap$-semilattice may be described dually.

Suppose the poset is a lattice $L$; then the following hold.

- $(x \sqcup y)^\leq = x^\leq \sqcup y^\leq$.
- $(x \sqcap y)^\geq = x^\geq \sqcap y^\geq$.
- The map \( x \mapsto x^\leq \) is a lattice embedding in the lattice of ideals, and is surjective if \( L \) is finite.
- The map \( x \mapsto x^\geq \) is a lattice embedding in the opposite lattice of filters, and is surjective if \( L \) is finite.
- If \( L \) is a 0-lattice (1-lattice) and \( \equiv \) is a congruence relation then \([0] ([1])\) is an ideal (filter).

In a Boolean algebra \( B \) the converse to the last fact holds. Recall that in \( B, x - y \) is defined to be \( x \cap y^c\), and \( x \oplus y \) is \((x - y) \cup (y - x)\). Dually, \( x \Rightarrow y \) is defined to be \( x^c \cap y \), and \( x \Leftarrow y \) is \((x \Rightarrow y) \cup (y \Rightarrow x)\).

In chapter 3 some identities are given for \( \oplus \); the dual identities hold for \( \leftrightarrow \).

**Lemma 15.** Suppose \( B \) is a Boolean algebra. If \( I \) is an ideal then the relation \( x \equiv y \), which holds iff \( x \oplus y \in I \), is a congruence relation. Dually if \( F \) is a filter then the relation \( x \equiv y \), which holds iff \( x \oplus y \in F \), is a congruence relation.

**Proof:** The relation \( \equiv \) is reflexive since \( x \oplus x = 0 \in I \); symmetric since \( \oplus \) is commutative; and transitive since \( x \oplus z = (x \oplus z) \oplus (x \oplus z) \) and an ideal is closed under \( \oplus \). Suppose \( x \oplus x', y \oplus y' \in I \). The lemma follows from
\[
(x \uplus y) \oplus (x' \uplus y'), (x \cap y) \oplus (x' \cap y') \leq (x \oplus x') \cup (y \oplus y'),
\]
which clearly follows from
\[
(x \uplus y) - (x' \uplus y'), (x \cap y) - (x' \cap y') \leq (x - x') \cup (y - y');
\]
we leave this to the reader.

As usual, the kernel of a homomorphism \( h : B \rightarrow B' \) of Boolean algebras may be defined to be the inverse image of 0; it is an ideal. On the other hand given an ideal \( I \) there is a quotient Boolean algebra \( B/I \), and a canonical epimorphism from \( B \) to \( B/I \) with kernel \( I \). For \( X \subseteq B \) let \( \text{co}-X \) denotes \( \{x^c : x \in X\} \); then \( X \) is an ideal iff \( \text{co}-X \) is a filter. The inverse image of 1 under a homomorphism is a filter; and a filter \( F \) is the inverse image of 1 under the canonical epimorphism to \( B/\text{co}-F \).

In a \( \sqcap \)-semilattice \( L \), an ideal \( I \) is called proper if \( I \subseteq L \); and maximal if it is proper, and if \( J \) is an ideal with \( I \subseteq J \subseteq L \) then then \( J = I \) or \( J = L \). In a \( 1\sqcup \)-semilattice, an ideal \( I \) is proper iff \( 1 \notin I \). The dual notions of a proper and maximal filter are defined in a \( \sqcap \)-semilattice; and a filter \( F \) is proper iff \( 0 \notin F \).

In a lattice, an ideal \( I \) is called prime if it is proper, and whenever \( x \sqcap y \in H, x \in H \) or \( y \in H \). Note that this is so iff \( I^c \) is a filter. A prime filter \( F \) is defined dually. In a Boolean algebra, \( I \) is a proper (maximal, prime) ideal iff \( \text{co}-I \) is a proper (maximal, prime) filter.

**Lemma 16.** Suppose \( L \) is a lattice, \( I \subseteq L \) an ideal, \( F \subseteq L \) a filter, and \( I \cap F = \emptyset \). Then there is an ideal (filter) \( H \), maximal among those with \( I \subseteq H \) and \( H \cap F = \emptyset \) \((F \subseteq H \) and \( H \cap I = \emptyset \)). If \( L \) is distributive any such \( H \) is prime.

**Proof:** As already observed the collection of ideals (filters) is an algebraic closure system, hence inductive, and the first claim is immediate. For the second, suppose \( H \) is not prime, and choose \( a, b \notin H \) with \( a \cap b \in H \).

Since \( H \) is maximal, there are \( p, q \in I \) with \( p \sqcup a, q \sqcap b \in F \). But then
\[
(p \sqcup a) \cap (q \sqcap b) = (p \sqcap q) \sqcup (p \sqcap b) \sqcup (a \sqcap q) \sqcup (a \cap b),
\]
a contradiction. The proof for the filter is dual.

**Corollary 17.** In a distributive lattice \( L \), if \( a \not\leq b \) there is a prime ideal (filter) \( H \) with \( b \in H \) and \( a \notin H \) \((a \in H \) and \( b \notin H \)).

**Proof:** Apply the lemma with \( I = b^\leq \) and \( F = a^\geq \).
Corollary 18. In a distributive 1-lattice (0-lattice), every maximal ideal (filter) is prime.

Proof: For ideals, apply (a slight generalization of) the lemma with $F = \{1\}$ and $I = \emptyset$; and dually for filters.

Theorem 19. For an ideal $H$ is a Boolean algebra $B$ the following are equivalent.

a. $H$ is prime.

b. If $x \in B$ then either $x \in H$ or $x^c \in H$.

c. $H$ is maximal.

The dual statement hold for filters.

Proof: We give the proof for filters. Suppose $H$ is prime, and $x \in B$. Since $x \cup x^c = 1$ and $1 \in H$, $x \in H$ or $x^c \in H$. This proves a$\Rightarrow$b. Suppose $b$ holds, and $H \subseteq F$ for a filter $F$. If $x \in F - H$ then $x^c \in H$, so $x^c \in F$, so $1 \in F$ and $F = B$. This proves b$\Rightarrow$c. Finally, c$\Rightarrow$a follows by corollary 18.

A maximal filter in a Boolean algebra is also called an ultrafilter. Note that any ideal (filter) is contained in a maximal ideal (filter). Exercise 6 shows that (under suitable hypotheses on $L$) if prime ideals in $L$ are maximal then $L$ is a Boolean algebra.

Consider the Boolean algebra $P(\omega)$ of subsets of the natural numbers $\mathcal{N}$, which is also denoted $\omega$. The principal filter generated by a singleton set $\{i\}$ is an ultrafilter, since given any set $S$ either $S$ or $S^c$ contains $i$. The collection of finite subsets of $\omega$ is readily seen to be an ideal; the collection of their complements, the cofinite sets, is a filter $F$. If a filter contains $F$ and a finite set then it is clearly all of $P(\omega)$. Thus, any ultrafilter containing $F$ is nonprincipal; in particular nonprincipal ultrafilters exist.

5. Ultraproducts. Suppose $S_i : i \in I$ is an indexed family of structures, and $F$ is a proper filter in $Pow(I)$, the collection of subsets of $I$. Let $P = \times_i S_i$ be the Cartesian product of the domains. If $a \in P$ let $a_i$ denote the $i$th component. Say that a sentence of the expanded language holds on $J \subseteq I$ if $\models_{S_i} G$ for all $i \in J$. Say that $a \equiv b$ if $a = b$ holds on some $J \in F$. This relation is reflexive since $a = a$ on $I$; symmetric since if $a = b$ holds on $J$ then $b = a$ holds on $J$; and transitive since if $a = b$ holds on $J$ and $b = c$ holds on $K$ then $a = c$ holds on $J \cap K$. Let $\hat{a}$ denote the equivalence class of $a$.

If $R$ is a relation, $R(a_1, \ldots, a_k)$ holds on $K$, and $a_i = a'_i$ on $J_i$, then $R(a'_1, \ldots, a'_k)$ holds on $K \cap J_1 \cap \cdots \cap J_k$. Thus, in the set $P/I$ we may define $R(\hat{a}_1, \ldots, \hat{a}_k)$ to hold if $\exists_i R(a_1, \ldots, a_k)$ holds on some $K \in F$. Similarly for a function $f$ we may define $f(\hat{a}_1, \ldots, \hat{a}_k)$ to equal $\hat{b}$ where $f(a_1, \ldots, a_k) = b$ (in the product structure).

We have thus defined a structure on $P/I$, called the reduced product, which we denote $\times_i S_i/F$. If $F$ is an ultrafilter the reduced product is called an ultraproduct.

Lemma 20. Suppose $P = \times_i S_i/F$ is an ultraproduct, and $G(\hat{a}_1, \ldots, \hat{a}_k)$ is a sentence of the expanded language. Then $\models_P G(\hat{a}_1, \ldots, \hat{a}_k)$ if $G(a_1, \ldots, a_k)$ holds on some $K \in F$.

Proof: First, for a term $t(x_1, \ldots, x_k)$, $t(\hat{a}_1, \ldots, \hat{a}_k) = \hat{b}$ where $t(a_1, \ldots, a_k) = b$; this follows by an easy induction. The lemma follows readily for atomic formulas. Suppose $G$ is $G_1 \land G_2$; if $G_1 \land G_2$ holds on $K \in F$ then so do $G_1, G_2$, so by induction $G_1, G_2$ hold in $P$, so $G_1 \land G_2$ does. Conversely if $G_1 \land G_2$ holds in $P$ then $G_1, G_2$ do, so $G_1$ holds on some $K_1 \in F$, $i = 1, 2$, so $G_1 \land G_2$ holds on $K_1 \cap K_2$. Suppose $G$ is $\neg G_1$; then $G$ holds in $P$ iff $G_1$ does not, which is so iff $\{i : \models_{S_i} G_1\} \notin F$ since $F$ is an ultrafilter; this is so iff $\neg G_1$ holds on some $K \in F$. Suppose $G$ is $\exists y G_1(y, a_1, \ldots, a_k)$. If $G$ holds in $P$ then $G_1(\hat{b}, \hat{a}_1, \ldots, \hat{a}_k)$ does for some $b \in P$, so $G_1(b, a_1, \ldots, a_k)$ holds on some $K \in F$, so $\exists y G_1(y, a_1, \ldots, a_k)$ holds on $K$. Conversely if $\exists y G_1(y, a_1, \ldots, a_k)$ holds on some $K \in F$, then for each $i \in K$ choose $b_i \in S_i$ so that $\models_{S_i} G_1(b_i, a_{i1}, \ldots, a_{ik})$, and for $i \notin K$ choose $b_i$ arbitrarily; then $G_1(b, a_1, \ldots, a_k)$ holds on $K$, so $G_1(\hat{b}, \hat{a}_1, \ldots, \hat{a}_k)$ holds in $P$, so $G$ does.
LEMMA 21. Suppose $S$ is a set of sentences in some first order language. Let $I$ be the set of finite subsets of $S$. For each $s \in I$ let $M_s$ be a model of $s$. Then there is an ultrafilter $Y$ in $P(I)$ such that $\times_s M_s/Y$ is a model of $S$.

PROOF: For each sentence $F \in S$ let $J_F$ be the set of $s \in I$ such that $F \in s$. Since $\{F_1, \ldots, F_k\}$ is an element of $J_{F_1} \cap \cdots \cap J_{F_k}$, the set $X = \{J_F : F \in S\}$ has the finite intersection property. Hence $X$ generates a proper filter, which can be extended to an ultrafilter $Y$ in $P(I)$. For $F \in S$, if $s \in J_F$ then $F \in s$ so $\models_M F$. That is, $F$ holds on $J_F \in Y$, whence it holds in $\times_s M_s/Y$.

This provides another proof of compactness, and as we shall see it sometimes yields further information. If $T$ is a theory we may consider the collection $\text{Mdl}_T$ of structures (for the language of $T$) which are models of $T$. To do so requires attending to a metamathematical detail. $\text{Mdl}_T$ is not a set; rather it is an example of a proper class. As noted in appendix 1, we can consider the set of sentences which are true in every structure in $\text{Mdl}_T$.

THEOREM 22. A class $C$ of structures equals $\text{Mdl}_T$ for some theory $T$ iff

a. whenever $A \in C$ and $B$ is elementarily equivalent to $A$ then $B \in C$; and
b. whenever $A_i \in C$ for $i \in I$ for some set $I$, and $B = \times_i A_i/U$ is an ultraproduct of the $A_i$, then $B \in C$.

PROOF: Clearly a holds in $\text{Mdl}_T$, and by lemma 20 b does also, since every sentence of $T$ holds in every $A_i$. Conversely suppose a and b hold, and let $T$ be the sentences true in every structure in $C$. We claim that any structure $S$ which is a model of $T$ is in $C$. To show this we will construct an ultraproduct $U$ of structures in $C$ which is a model of the theory of $S$, and hence elementarily equivalent to $S$. For this it suffices by lemma 21 to show that for any finite subset $X$ of the theory of $S$ there is a model of $T \cup X$ in $C$. But if this were not so then the conjunction $F$ of the sentences of $X$ would be false in every structure of $C$, whence $\neg F \in T$, a contradiction since $\models_S T$.

Exercises.

1. Show that the recursive definition of the formulas of first order logic has the unique readability property. Hint: First show that there are at least as many left parentheses as right in any prefix of a formula.

2. Prove lemma 1.

3. A set $T$ of PC formulas is said to be finitely PC-satisfiable if every finite subset of $T$ is PC-satisfiable. Show that every finitely PC-satisfiable set is subset of a maximal finitely PC-satisfiable set. Conclude that a finitely PC-satisfiable set is PC-satisfiable.

4. Show the following.
   a. $F \vdash_{\text{PC}} \neg \neg F$ (without using completeness).
   b. A literal is an atom or the negation of an atom. If $\alpha$ is a truth assignment let $[H]_\alpha$ denote the truth value of formula $H$. Show that, if $L_i$ is the literal $F_i$ if $[F_i]_\alpha = 1$ else $\neg F_i$, then $L_1, \ldots, L_k \vdash H$ if $[H]_\alpha = 1$ and $L_1, \ldots, L_k \vdash \neg H$ if $[H]_\alpha = 0$.
   c. Conclude that if $\models_{\text{PC}} F$ then $\vdash_{\text{PC}} F$.

5. Prove compactness without using a derivation system. Proceed as in exercise 3, using the method of lemma 11.

6. Show that if $L$ is a distributive 0-1 lattice and every prime ideal is maximal then $L$ is a Boolean algebra. Hint: Suppose $a \in L$. Let $F$ be the filter $\{x : a \sqcup x = 1\}$. Let $G$ be the filter generated by $F$ and $a^\ge$. Obtain a contradiction from the assumption that $G$ is proper.
12. Computability.

1. Introduction. Formal models of computation provide a precise definition of what functions \( f : \mathbb{Z} \rightarrow \mathbb{Z} \) can be “computed mechanically”, or specified by an algorithm. Mathematicians were aware that such functions were of interest well before the first formal models were given. For example, Hilbert’s 10th problem asked whether there was a mechanical method for determining whether a polynomial with integer coefficients had an integer solution. This question was settled in the negative in 1970 (a proof is given in section 8). Clearly a negative solution requires a formal notion of mechanically computable functions.

There are several equivalent formal models of computation; one simple one is a programming language with two kinds of statements.

- **assignment statement** = variable ← term
- **conditional goto statement** = if atomic formula goto positive integer

A program is a sequence of statements, which are numbered consecutively from 1. Such programs will be called “goto programs”. Allowing “true” as an atomic formula obviates the need for an unconditional goto. A halt statement is also redundant, its function being served by a goto statement with a nonexistent target. (However, if a more “self-documenting” language is desired these statements can be added, and do not unduly increase the complexity of the formal theory.)

Note that a program involves only finitely many variables. A computation is defined to be a sequence of pairs \( (N_i, A_i) \) where \( N_i \) is a positive integer and \( A_i \) an assignment to the variables of the program, such that the following is true.

- \( N_1 = 1 \).
- If statement \( N_i \) is an assignment \( x \leftarrow t \) then \( N_{i+1} = N_i + 1 \) and \( A_{i+1} \) differs from \( A_i \) in that \( x \) is assigned the value \( [t]_{A_i} \).
- If statement \( N_i \) is a conditional goto statement “if \( F \) goto \( n \)” then if \( [F]_{A_i} \) is true then \( N_{i+1} = n \), else \( N_{i+1} = N_i + 1 \); and \( A_{i+1} = A_i \).
- If the sequence is finite the last pair has \( N_i \) greater than the number of any statement.

If the computation is finite it is said to have halted. Conventions may be adopted (from among a small variety of possibilities) to specify a partial function computed by a program. We shall consider that the variables have some “alphabetic” order imposed on them; the input is in the lowest variable; the output is also; and remaining variables are initialized to some distinguished value such as 0. An input value is in the domain of the partial function iff the computation for that input halts. The program can similarly be considered to compute a partial function from \( n \) inputs to \( m \) outputs, provided \( n \) and \( m \) are less than the number of variables.

The foregoing gives a programming language for any structure with a distinguished constant. This topic has been studied, but the most important example is of course \( \mathbb{Z} \), with, say, 0, 1, +, −, ≤ as the language. One can see that the language 0, \( x+1 \), \( x-1 \), \( x > 0 \) gives the same computable functions. However the number of steps in a computation may increase by a significant amount; this point is discussed further below.

A recursive style programming language can also be given.

```plaintext
statement =
      variable ← term;
  {statement ··· statement}
  if atomic formula then statement
  while atomic formula do statement
```

A program is a sequence of statements. Such programs will be called “while programs”.

In the next section it will be shown that these two models of computation are equivalent, and also to a third model, namely “Turing machines”. Beginning in the late 1920’s many formal models of computation
were defined, and by the 1930’s all had been shown to be equivalent. Church’s thesis states that any algorithm on the integers can be implemented by a program in one (hence any) of these models. The fact that the same class of partial functions is computed by a wide variety of models is one piece of evidence that the correct formal model of an algorithm has been given, which is the philosophical content of Church’s thesis. It also has a practical aspect; an algorithm can be given informally, and it is certain that there is a formal one, just as is the case for proofs.

Many models of computation consider the values to be nonnegative integers, and others consider the values to be “strings” (finite sequences from a finite alphabet). Strings can be coded as integers in various ways; this realization was novel in the 1930’s, and led to great breakthroughs in mathematical logic, due to the fact that the formal languages of mathematics are readily formalized in terms of strings.

Given a model of computation the notion of a computation is defined, typically as a sequence of strings. The length of the computation is the number of strings in the sequence. A realization of complexity theory has been that this should be the same, up to a polynomial, for models of computation to be used for measuring the computational complexity of functions; and indeed there is a large family of models for which this is the case.

The “big oh” notation was invented by number theorists in the late 19th century, and has found a new use in the theory of computation. If a function \( f : \mathbb{R} \rightarrow \mathbb{R} \) is complicated, we might wish to consider an estimate of its growth. The statement that \( f \) is \( O(g) \) means that there is a positive constant \( c \) such that for sufficiently large \( x \), \( |f(x)| < c|g(x)| \). Often \( g \) is positive; and often a simple function such as \( n^2 \) or \( 1/n \). For an example, the \( n \)th prime number is \( O(n \log n) \) (much stronger statements are possible; this fact was proved by Chebyshev).

2. Equivalence of three models. A computation of a goto program was already defined. Note that if there are \( n \) “steps” in the computation then there are \( n + 1 \) variable assignments, since the sequence of assignments begins with the initial one.

For the while programs, the computation of an assignment statement consists of two variable assignments, “before” and “after”; this may be denoted as \( A_0SA_1 \), where \( S \) denotes the assignment statement. The computation of \( \{S_1 \ldots S_n\} \) may be denoted as \( A_0S_1A_1 \ldots A_{n-1}S_nA_n \), where \( A_{i-1}S_iA_i \) is the computation of \( S_i \). A proviso must be made for infinite computations; if any of the computations of the \( S_i \) is infinite than the overall computation ends with the first such. The computation of “if \( F \) \( S \)” may be denoted \( A_0FA_0SA_1 \) if \( F \) is true, else \( A_0FA_0 \). The computation of “while \( F \) \( S \)” may be denoted \( A_0FA_0SA_1FA_1S \ldots \), where the sequence terminates with \( A_i \) when \( F \) is false at \( A_i \), with the proviso that it becomes infinite if the computation of \( S \) does.

The while and goto programs will be slightly modified in this section, requiring atomic formulas (terms) to be of the form \( R(t_1, \ldots, t_n) \) or “true” \( (f(t_1, \ldots, t_n) \) or a constant), where \( t_i \) is a constant or variable. This is convenient for simulations, and increases the length of a computation by only a constant factor (the constant depending on the program).

**Theorem 1.** If \( P \) is a while program there is a goto program \( P' \) such that on any input, each step in the computation of \( P \) is simulated by at most 2 steps in the computation of \( P' \).

**Proof:** An assignment statement is translated as itself. \( \{S_1 \ldots S_n\} \) is translated as \( S'_1 \ldots S'_n \), where \( S'_i \) is \( S'_i \) with line numbers in goto’s adjusted. “If \( F \) \( S \)” is translated as

- if \( F \) goto 3
- goto \( N \)
- \( S'' \)

where \( N \) is the last line number plus 1. “While \( F \) \( S \)” is translated as

- if \( F \) goto 3
A Turing machine has a tape alphabet \( A \), a set of states \( Q \), and a "head state" alphabet \( Q' \). One of the tape symbols is distinguished as a blank symbol, denoted \( b \). It also has a set of rules, of the form \( qsa \rightarrow q'b's \) where \( q, q' \in Q \), \( s, s' \in A \), and \( d \) is \( L \) (left) or \( R \) (right). A Turing machine is called deterministic if no pair \( qs \) occurs more than once on the left side of a rule; Turing machines will be required to be deterministic.

The state of the machine is represented by a string over \( Q \cup A \), which contains exactly one symbol from \( Q \). If the state is of the form \( aq_0 \), the rule with left side \( qa_0 \) (if any) is applied to update the state, resulting in a step of the computation. If the right side is \( q'sR \) the new state is \( aq's' \beta \), which may be described as, “if the tape head is reading \( s \) in state \( q \) then it overwrites \( s \) with \( s' \), switches to state \( q' \), and moves one square to the right”.

The remaining types of transitions are as follows. If the state is \( aq \) and there is a rule with left side \( qb \) and right side \( q's'R \) the new state is \( aq's' \) (the tape grows a blank square). If the rule is \( qa \rightarrow q's'L \) then \( atqsa \) becomes \( aq't's' \beta \) and \( qsa \beta \) becomes \( q'bs' \beta \); and if \( s = b \) then \( atq \) becomes \( q'a'ts' \beta \) and \( q \) becomes \( q'bs' \).

A computation is a sequence of states, where each follows from the previous according to a rule. A state may be reached which has no successor, in which case the computation has halted. A computation may also go on forever. Note that blanks may be added to the beginning or end of the initial state, without affecting the computation.
To view a Turing machine as computing a function from \( \mathbb{Z} \) to \( \mathbb{Z} \) a convention for coding the input as the initial state must be adopted. A simple choice is to represent a nonnegative integer \( n \) by its 2-adic notation, a negative integer as its absolute value preceded by a minus sign, and the initial state as the input preceded by the initial head state. If the computation halts in state \( \alpha q \beta \), the output value is the longest valid signed integer following \( q \) (either the longest string of 1’s and 2’s, or a minus sign followed by at least one 1 or 2, and as many subsequent 1’s or 2’s as possible). Similarly a Turing machine may be viewed as computing a function \( f(x_1, \ldots, x_n) \), by adding comma to the tape alphabet.

For the following theorem an important definition will be given. Say that a function \( f(x) \) is computable in polynomial time by a program if there is a polynomial \( p \) such that the number of steps required on input \( x \) is at most \( p(\ell(x)) \); equivalently, there is a constant \( c \) such that the number of steps is \( O(\ell(x)^c) \).

**Theorem 2.** A step of a goto program can be simulated in polynomial time by a Turing machine.

**Proof:** The state \( \langle i, A_i \rangle \) of the goto program is represented by the state of the Turing machine with the head state \( i \) to the left, and the assignment following as a string over “12-,” (possibly with additional leading or trailing blanks). To simulate a step, first either 0, 1, \( x + y \), or \( x - y \) is written to the left of the assignment. Second, for a test the result is recorded in the state; and for an assignment statement the result is moved to the correct place in the new variable assignment. Third, the tape is cleaned up. Notes are made on the tape by using a note track; further details are given in section 10. The extra blanks do not affect the time, since the head does not range into them.

As a corollary, given a goto program there is a Turing machine whose computation simulates that of the goto program, with only a polynomial increase in the number of steps. That is, if the goto program halts after \( t \) steps then the Turing machine halts after \( p(t + \max(\ell(x_1), \ldots, \ell(x_n))) \) steps for some polynomial \( p \). The intermediate states for simulating a step can be “tagged” with the line number, in the simulation of the entire program.

**Theorem 3.** A step of a Turing machine can be simulated in polynomial time by a while program.

**Proof:** The state of the Turing machine is held in 4 variables, \( \alpha \), \( q \), \( s \), and \( \beta \). The string operations necessary to update these are not difficult to implement as while programs; see section 11 for details.

As a corollary, given a Turing machine there is a while program whose computation simulates that of the Turing machine, with only a polynomial increase in the number of steps. An outer while loop has a loop body which is a sequence of “if” statements, one for each head state; the body of each of these is a sequence of “if” statements, one for each tape symbol; and the body of each of these simulates the Turing machine step. The result of this will be the final Turing machine state, from which the output value must be extracted; this is easily done using the methods of the theorem.

By the results above, the functions computable in polynomial time are the same in all three models of computation, illustrating that Church’s thesis seems to hold for polynomial time for “suitable” models of computation, where suitability is something that computation theorists are by now well familiar with.

To conclude this section a few further comments will be given on the models of this section. It is not difficult to see that \( x + y \) cannot be computed in polynomial time by a while program with only \( x + 1 \) and \( x - 1 \) as the basic functions (although it can be computed, indeed in “exponential time”). On the other hand, \( x \cdot y \) can be computed in polynomial time using \( x + y \) and \( x - y \) (exercise). Thus “times” may be omitted from the atomic function; but it is frequently added, especially in measuring the performance of algorithms.

The while programs can be expanded with various other programming language features, making it even more obvious that they provide a complete model of computation, often with the simulation by basic while programs involving only polynomial increase in number of steps. Two examples are arrays and assigned goto’s. Arrays may have special names, distinct from the ordinary variables. An array reference is of the
form $A[t]$ where $A$ is an array name and $t$ a term; array references are terms. For assigned goto's, there is a third set of variables, which may have statement numbers assigned to them; these may appear as the target location of a goto statement. The simulations are left to the reader.

The main advantage to Turing machines is that they are easy to “code”, so that they may be input to other Turing machines. Indeed, there is a “universal” Turing machine $U$ which, on input $\tau; \alpha$, where $\tau$ codes a Turing machine and $\alpha$ its input, simulates the computation of $\tau$ step by step. Details are given in section 12.

Turing machines can be viewed as computing functions directly on strings. This is useful when the inputs are naturally strings already, such as a language of mathematical logic. Note that if the string is considered as an integer via $n$-adic notation, and this in turn as a string via $m$-adic notation, where both $m$ and $n$ are greater than or equal to 2, the length changes by only a constant factor.

3. Basic definitions. Having a formal model of an algorithm, and the partial function computed by it, precise definitions can be given for various notions related to computation. Let $f$ denote a function from $\mathcal{N}$ to $\mathcal{N}$, $\phi$ a partial function, and $S$ a subset of $\mathcal{N}$.

- $\phi$ is said to be partial recursive if it is the partial function computed by some algorithm.
- $f$ is said to be recursive if it is the function computed by some algorithm, which halts on any input.
- $S$ is said to be recursive, or decidable, if there is an algorithm which halts on any input, and outputs, say, 1 or 0 according to whether $x \in S$ or $x \notin S$.
- $S$ is said to be recursively enumerable if it is the set of inputs on which some algorithm halts.

Some remarks on the origin of the terminology “recursive” for a function computable by an algorithm may be found in H. Enderton’s article in [HML]; it seems to be due to Gödel.

The notion of partial recursive function from $\mathcal{N}^k$ to $\mathcal{N}$ is defined similarly, along with recursive functions, recursive predicates, and recursively enumerable predicates of several variables. Needless to say there are many equivalent formulations. A particularly important one is that $R(x)$ is recursively enumerable iff there is a recursive predicate $Q(w, x)$ such that $R(x)$ iff $\exists w Q(w, x)$. Given $R$, $Q$ is given by the predicate “the computation on input $x$ halts in $w$ steps”; given $Q$, $R$ may be computed by an algorithm which evaluates $Q(w, x)$ for successive values of $w$.

We leave it as an exercise to show that a set is recursively enumerable iff it is empty or the range of some recursive function; or if there is some algorithm which outputs only members of $S$, and eventually outputs any given member (this runs forever if $S$ is infinite).

Fixing a model of computation, a program can be coded as an integer, say by first writing it as a string in a finite alphabet. Let $\phi_e$ denote the partial function computed by the program with code $e$; this can be defined for any integer $e$ by letting $\phi_e$ be some fixed partial function if $e$ is not the code of a well-formed program. A value $e$ such that $\phi = \phi_e$ is called an index for $\phi$. Note that $\phi_e(x) = U(E, x)$ where $U$ is the “universal” program. $W_e$ is used to denote the domain of $\phi_e$.

An important fact about indices is that a transformation of programs is given by recursive function on the indices. This is often expressed by saying that the transformation is “effective”; more generally a effective procedure procedure is called effective if after suitable coding it is a recursive function on the integers (or strings over a finite alphabet). For example, the transformation from $R$ to $Q$ given above is effective. For another example, the enumeration of the partial recursive functions $\phi_e$, and the enumeration $W_e$ of the recursively enumerable sets, are effective. The recursive functions are not effectively enumerable; given any enumeration $f_e$ of total recursive functions, the total recursive function $f_e(e) + 1$ is not one of them.

The universal program is useful in another respect; the predicate “$U$ halts within $t$ steps with input machine $e$ and input value $x$” is a (quite easy to compute) recursive predicate. This predicate (denoted $T(t, e, x)$) is called a “Kleene T predicate”. Machine $e$ halts on input $x$ iff $\exists t T(t, e, x)$.
Theorem 4. A set $S$ is recursive iff both $S$ and $S^c$ are recursively enumerable.

Proof: If the characteristic function of $S$ is computed by program $P$, programs $P_1$ and $P_2$ are easily (effectively) derived from $P$, whose domains are $S$ and $S^c$. If $S$ is the domain of $P_1$ and $S^c$ is the domain of $P_2$, a program $P$ computing the characteristic function of $S$ alternately executes a step of $P_1$ and a step of $P_2$, until one halts.

4. Undecidability. There are only countably many Turing machines, and uncountably many sets of integers, so there are sets of integers which are not recursively enumerable. One achievement of computability theory has been to demonstrate that specific sets are not. An example can be given readily.

Theorem 5. Let $K = \{e : e \notin W_e\}$.

a. $K^c$ is not recursively enumerable.

b. $K$ is recursively enumerable but not recursive.

c. If $W$ is recursively enumerable then there is a recursive function $f$ such that $x \in W$ iff $f(x) \in K$.

Proof: For part a, $d \in W_d$ iff $d \notin K^c$, so $K^c \neq W_d$, for any $d$. For part b, a machine whose domain is $K$ consists of running a universal machine $U$ on $(e, c)$. That $K$ is not recursive follows from part a and theorem 4. For part c, given $W_e$, $x \in W_e$ iff $d \in K$ where $d$ is the machine which runs $W_e$ on $x$ (so that $W_d$ is $\mathcal{N}$ or $\emptyset$). It is easy to see that $d$ is effectively computable from $x$, by a recursive function $f$.

A set $S$ is said to be many-one reducible to a set $T$ if there is a recursive function $f$ such that for all $x$, $x \in S$ iff $f(x) \in T$. The theorem shows that $K$ is recursively enumerable, and any recursively enumerable set is many-one reducible to $K$. A set with these properties is called many-one complete. Sets which are many-one complete are not recursive. A predicate is called undecidable if it is not recursive.

Many predicates are shown to be undecidable by showing that they are many-one complete. If a set $S$ is recursively enumerable and a many-one complete set is many-one reducible to $S$ then $S$ is many-one complete, as is easily seen. This provides a method for showing many-one completeness. Although it will not be shown in this brief introduction, there are recursively enumerable sets which are neither recursive nor many-one complete.

Given a set of “generators” $A$, the free semigroup $F$ generated by $A$ is just the finite strings over $A$, with concatenation as the multiplication operation; this semigroup is a monoid, with identity the null string. Given a set of “relations”, equations of the form $u = v$ where $u$ and $v$ are strings, by universal algebra there is a least congruence relation $\equiv$ containing them. As is readily verified, $x \equiv y$ iff there is a sequence $w_1, \ldots, w_n$ of strings with $x = w_1, y = w_n$, and for $1 \leq i < n$ there are strings $s$ and $t$ and an equation $u = v$ (or $v = u$) with $w_i = stu$ and $w_{i+1} = st$. The semigroup specified by the relations is the quotient $F/\equiv$.

A finite presentation of a semigroup consists of a finite set of generators and a finite set of relations. The word problem for semigroups is to decide, given a finite presentation and two strings, whether they are equivalent.

Theorem 6. The word problem for semigroups is many-one complete.

Proof: To see that it is recursively enumerable, consider the procedure which generates the list of strings reachable from $x$ in some order, and halts if $y$ is generated. Suppose $M$ is a Turing machine whose initial state is never re-entered. A semigroup presentation $P$ is defined whose alphabet is $Q \cup A \cup \{E, H\}$. The relations are of three types. The first type simulates the Turing machine rules in a direct manner, as follows.

- $qs \rightarrow q's'\mathcal{R}: qs = s'q'$, and if $s = b qE = s'q'E$,
- $qs \rightarrow q's'\mathcal{L}: tqs = q'ts'$ for each $t \in A$ and $Eqs = Eq'bs'$, and if $s = b tqE = q'ts'E$ for each $t \in A$ and $EqE = Eq'bs'E$. 

106
We claim if that $x$ and $y$ are states of the Turing machine, with $y$ halting, $x \Rightarrow_M y$ iff $ExE \Rightarrow_P EyE$, where $\Rightarrow_M (\Rightarrow_P)$ means reaches by a Turing machine computation (semigroup derivation). For the forward implication the simulation is obvious. For the converse the proof is by induction on the length of the semigroup derivation, the basis of length 0 being trivial. If $ExE \Rightarrow_P EyE$ then by induction $w \Rightarrow_M y$. If $x \rightarrow_M w$ the conclusion is immediate; otherwise $w \Rightarrow_M x$, and this is the first step of the computation $w \Rightarrow_M y$, so $x \Rightarrow_M y$. The second type of semigroup rule is $qs = H$ for each halting $qs$, and if $qE = HE$ if $gb$ is halting. The third type is $sH = H$ and $Hs = H$ for each $s \in A$. Clearly, $ExE \Rightarrow_P EyE$ iff $M$ halts on $x$. It is now easy to see that the theorem is true.

**Theorem 7.** The valid sentences of the predicate calculus are many-one complete.

**Proof:** Since there is a recursive system of axioms for the valid sentences, they are recursively enumerable. Given a semigroup presentation, introduce a function symbol $s_a$ for each $a \in A$. The string $x = a_1 \cdots a_n$ will be coded as the term $\hat{x} = s_a(\cdots(s_a(0))\cdots)$. The rule $u = v$ is coded as the equation $\hat{u} = \hat{v}$. Let $\hat{S}$ be the codes of the relations, and $E$ the axioms of equality. We claim that $x \equiv y$ iff $\hat{S} \cup E \models \hat{x} = \hat{y}$. If $x \equiv y$ a derivation of $y$ from $x$ can be simulated by a proof in open first order logic using the axioms of equality. If the logical implication holds consider the Herbrand structure where the congruence relation is $\equiv$; the premises hold in the structure, so the conclusion does. The theorem has been proved, by reducing the semigroup word problem to validity.

Various refinements of these results can be given; we omit a discussion of this, except to mention that validity of $\Sigma_1$ sentences is many-one complete, and indeed the language can be restricted to have a constant, two unary function symbols, and one binary predicate. This follows fairly readily from the foregoing and is left to the reader. A $\Pi_1$ sentence is valid if it is a tautology (when the quantifiers are removed), so validity of these is decidable.

**5. Subrecursion.** Although the recursive functions are not effectively enumerable, there are subclasses of them that are, which have various uses. These are typically defined as multivariate functions on $N$, and given by an “inductive definition”, i.e., as the smallest class containing certain initial functions and closed under certain operations. The operation of “substitution” is usually among the operations; if $t$ is a term involving functions already defined then the function defined by $t$ is considered to be defined also.

The successor function $\text{Suc}(n)$ is has the value $n + 1$ on argument $n$. A function $f : N^{k+1} \rightarrow N$ is said to be obtained by primitive recursion from functions $g : N^k \rightarrow N$ and $h : N^{k+2} \rightarrow N$ if $f(0, y) = g(y)$ and $f(\text{Suc}(x), y) = h(x, y, f(x, y))$. The primitive recursive functions are defined to be the smallest family containing 0 and $\text{Suc}$, and closed under substitution and primitive recursion. They are recursive by Church’s thesis. Indeed it is easy to give a recursive definition of a while program computing each primitive recursive function; see also theorem 13.

The primitive recursive functions are an easily defined effectively enumerable class of total recursive functions, and have found repeated use in computation theory and mathematical logic. It was realized early on that they were a large class, in particular that functions relevant to applications were in much smaller classes. Kalmar defined one such class in 1943, and Grzegorczyk defined a hierarchy of them in 1953. The main object of this section is to show that the $T$ predicate described in section 3 is primitive recursive; with a bit more work stronger results can be obtained.

Define the functions $A_n(x, y)$ for $n \geq 1$ as follows.

\[
A_1(x, 0) = x, \quad A_1(x, \text{Suc}(y)) = \text{Suc}(A_1(x, y));
\]
\[
A_2(x, 0) = 0, \quad A_2(x, \text{Suc}(y)) = A_1(x, A_2(x, y));
\]
\[
A_n(x, 0) = 1, \quad A_n(x, \text{Suc}(y)) = A_{n-1}(x, A_n(x, y)) \text{ for } n \geq 3.
\]

107
These functions are primitive recursive; $A_1$ is addition, $A_2$ is multiplication, $A_3$ is exponentiation, and $A_4$ is a stack of exponentials of height $y$. Clearly functions growing as fast as $A_4$ have entered realms of theoretical rather than computational mathematics.

A function $f : \mathbb{N}^{k+1} \rightarrow \mathbb{N}$ is said to be obtained by bounded (or limited) recursion from functions $g : \mathbb{N}^k \rightarrow \mathbb{N}$, $h : \mathbb{N}^{k+2} \rightarrow \mathbb{N}$, and $b : \mathbb{N}^{k+1} \rightarrow \mathbb{N}$ if $f(0,y) = g(y)$ and $f(Suc(x), y) = \min(b(x,y), h(x, y, f(x, y)))$.

The class of function $E_n$ for $n \geq 1$ is defined to be the smallest class containing $0$, $\text{Suc}$, and $A_n$, and closed under substitution and bounded recursion. These are the classes of the Grzegorczyk hierarchy; $E_3$ is also known as the (Kalmar) elementary functions. The class $E_0$, where the initial functions are just $0$ and $\text{Suc}$, has been studied; but we omit this. In the following let $\max(x)$ denote $\max(x_1, \ldots, x_k)$ for the $k$-tuple $x$.

**Lemma 8.** Suppose $n \geq 1$ and $x, y \geq 2$.

a. $A_n(x, y) > y$.

b. $A_n(x, \text{Suc}(y)) > A_n(x, y)$.

c. $A_n(\text{Suc}(x), y) > A_n(x, y)$.

d. $A_{n+1}(x, y) > A_n(x, y)$ unless $x = y = 2$.

**Proof:** First, by induction on $n$. $A_n(x, 1) = x$ for $n \geq 2$, and $A_n(2, 2) = 4$ for $n \geq 1$. Parts a-c are verified directly for $n = 1$; suppose $n \geq 2$. For part a, the basis $y = 2$ has been verified; and $A_n(x, y) > y \geq 2$, so $A_{n-1}(x, A_n(x, y)) > A_n(x, y)$ (even in the case $n = 2$) and part a follows; part b follows also. Part c follows from $A_{n-1}(\text{Suc}(x), A_n(\text{Suc}(x), y)) > A_{n-1}(\text{Suc}(x), A_n(x, y))$ for $x \geq 2$, $y \geq 1$, which follows by results already proved. Part d follows from $A_n(x, A_{n+1}(x, y)) > A_{n-1}(x, A_n(x, y))$ for $x \geq 2$, $y \geq 1$, unless $x = 2$ and $y = 1$, which follows by results already proved.

**Lemma 9.** Suppose $n \geq 1$ and $x, y, z \geq 2$.

a. $A_n(\text{Suc}(x), y, z) \leq A_n(x, A_n(y, z))$.

b. $A_n(A_{n+1}(x, y), A_{n+1}(x, z)) \leq A_{n+1}(x, y + z)$.

c. For $n \geq 2$ $A_n(x, A_n(y, z)) \leq A_n(x, A_{n-1}(y, z))$.

d. $A_n(A_{n+1}(x, y), x) \leq A_{n+1}(x, \text{Suc}(y))$.

**Proof:** Parts a and b are proved for successive $n$; equality holds everywhere for $n = 1, 2$ for part a, and $n = 1$ for part b. Part a follows by induction on $z$ using part b for $n - 1$ and the induction hypothesis. Part b follows by induction on $y$ using part a and the induction hypothesis. Part c follows by the same argument as part a (equality holds for $n = 3$). Part d follows by induction on $y$, rewriting the right side as $A_n(x, A_{n+1}(x, y))$ and using part a (equality holds for $n = 1, 2$).

**Theorem 10.** Suppose $n \geq 1$.

a. For each $c$ $A_{n+1}(x, c) \in E_n$.

b. If $f \in E_n$ then for some $c$, for all $x$ $f(x) \leq A_{n+1}(\max(2, x), c)$.

c. $E_n \subseteq E_{n+1}$.

d. For $n \geq 2$ if $f$ is defined by primitive recursion from functions in $E_n$ then $f \in E_{n+1}$.

**Proof:** First, by lemma 8d $E_n \subseteq E_{n+1}$. Part a follows by induction on $c$ (for the basis, $A_n(x, 0) \in E_0$ for any $n$). Part b follows by induction on the definition of $f$. It follows (except for some special cases) for $A_n(x, y)$ since $A_{n+1}(m, 2) = A_n(m, m)$ where $m = \max(x, y)$. To prove the induction step for substitution, suppose $h_i(x) \leq A_{n+1}(\max(2, x), c_i)$, and $g(y) \leq A_{n+1}(\max(2, y), d)$. Apply lemma 9a with $y$ the max of the $c_i$ and $z = d$. For bounded recursion the induction step is immediate. For part c, we have already observed that $E_n \subseteq E_{n+1}$. Also $A_{n+1}(x, x) \in E_{n+1}$; however this function cannot be in $E_n$ by part b and lemma 8c.

108
For part d, define $b(0, w) = w$, $b(Suc(x), w) = A_{n+1}(b(x, w), c)$ where $h(x, y, w) \leq A_{n+1}(\text{max}(2, x, y, w), c)$ and $h$ is the recursion step function of the definition of $f$; then $f(x, y) \leq b(x, A_{n+1}(\text{max}(2, y, d)))$ where $g(y) \leq A_{n+1}(\text{max}(2, y), d)$ and $g$ is the basis function of the definition of $f$. It suffices to show that $b \in E^{n+1}$. First, $b(x, w) \leq A_{n+1}(w, A_{n+1}(c, x))$; this follows by induction on $x$ using lemma 9. From this $b(x, w) \leq A_{n+2}(\text{max}(2, c, x, w), 3)$.

In particular the union of the $E_n$ equals the primitive recursive functions. The function $A_n(x, y)$ of three variables is not primitive recursive, since otherwise it would be in some $E_n$, contradicting part c of the theorem.

For this paragraph it is convenient to modify the max and $\ell$ functions by first max’ing the argument with 2; let max’ and $\ell'$ denote these functions. Alternatively the $O$ notation could be used. For $f$ in $E_1$ (resp. $E_2$), for some $c f(x) \leq \text{max'}(x) \cdot c$ (resp. $f(x) \leq \text{max'}(x)^c$), whence for some $c \ell'(f(x)) \leq \ell'(x) + c$ (resp. $\ell'(f(x)) \leq \ell'(x) \cdot c$). Clearly it would be desirable to have a class where $\ell'(f(x)) \leq \ell'(x)^c$ for some $c$, and the functions in $E_3$ grow faster than this. Such a class, known as $L$, will be defined shortly.

For the following suppose $n \geq 2$; the following functions are in $E_1$.

- Predecessor, $\text{Pred}(0) = 0$, $\text{Pred}(\text{Suc}(x)) = x$.
- Limited difference, $\text{Ldif}(x, 0) = x$, $\text{Ldif}(x, \text{Suc}(y)) = \text{Pred}(\text{Ldif}(x, y))$.
- Conditional, $\text{Cond}(0, y, z) = y$, $\text{Cond}(\text{Suc}(x), y, z) = z$.
- $n$-ary conditional, $\text{Cond}_n(x, y_0, \ldots, y_{n-1}) = \text{Cond}(x, y_0, \text{Cond}(\text{Pred}(x), y_1, \ldots))$.
- Right digit of $n$-adic notation, $\text{Rdig}_n(0, 0) = 0$, $\text{Rdig}_n(\text{Suc}(x)) = \text{Cond}_{n+1}(\text{Rdig}_n(x), 1, \text{Rdig}_n(x) + 2, \ldots, \text{Rdig}_n(x) + 1)$.
- Trim right digit of $n$-adic notation, $\text{Trdig}_n(0, 0) = 0$, $\text{Trdig}_n(\text{Suc}(x)) = \text{Cond}_{n+1}(\text{Rdig}_n(x), 0, \text{Trdig}_n(x), \ldots, \text{Suc}(\text{Trdig}_n(x)))$.
- Trim digits from right of $n$-adic notation, $\text{Trdigs}_n(0, y, y) = y$,
  $\text{Trdigs}_n(\text{Suc}(x), y) = \text{Trdig}_{n+1}(\text{Trdigs}_n(x, y))$.
- Length of $n$-adic notation up to limit, $\text{Len}_n(0, x, 0) = 0$, $\text{Len}_n(\text{Suc}(w), x) = \text{Cond}_n(\text{Trdigs}_n(w, x), \text{Len}_n(w, x), \text{Suc}(\text{Len}_n(w, x)))$.
- Length of $n$-adic notation, $\text{Len}_n(x) = \text{Len}_n(x, x)$.

In all cases a bounding function is clear and left to the reader.

Let $\text{App}_n^b$ be the function $x + \cdots + x + i$ where $x$ is repeated $n$ times. A function $f$ is said to be defined from functions $g$ and $h_i$ for $1 \leq i \leq n$ by recursion on $n$-adic notation if $f(0, y) = g(y)$ and $f(\text{App}_n^b(x), y) = h_i(x, y, f(x, y))$. This operation is convenient for defining string manipulation functions. The recursion is said to be bounded if the right sides of the $h_i$ equations are replaced by $\min(b(x, y), h_i(x, y, f(x, y)))$ for a function $b$.

**Lemma 11.** For $n \geq 2$ the primitive recursive functions are closed under recursion on $n$-adic notation. For $k \geq 1$ the class $E_k$ is closed under bounded recursion on $n$-adic notation.

**PROOF:** Define by primitive recursion the function $k(w, x, y)$ which is the value of $f(u, y)$ where $u$ is the leftmost $w$ digits of $x$. Thus (omitting the subscript $n$), let $\text{Ldif}(x, y) = \text{Trdigs}(\text{Ldif}(\text{Len}(y), x), y)$; and let $k(0, x, y) = g(y)$, and

$k(w + 1, x, y) = \text{cond}_{n+1}(\text{Rdig}(\text{Ldif}(w + 1, x))), 0$, $h_1(\text{Ldif}(w, x), y, k(w, x, y)), \ldots, h_n(\text{Ldif}(w, x), y, k(w, x, y)))$.

Then $f(x, y) = k(x, x, y)$. If the recursion on notation is bounded by $b$, the primitive recursion may be bounded by $b(\text{Ldif}(w, x), y)$.

**Theorem 12.** For any Turing machine $M$ the step function $\text{Step}_M(x)$ is in $E_1$. The $T$ predicate is in $E_1$, provided we consider it to be “$U$ halts within $\ell(t)$ steps with input machine $e$ and input value $x$”, and code machines appropriately.
Theorem 13.

For any Turing machine $M$ the step function $\text{Step}_M(x)$ is in $L$. Also, $f \in L$ iff $f(x)$ is computed within time $t(\ell(x))$ by a Turing machine where $t$ is a polynomial.

Proof: An outline is given in section 13, with details left to the exercises.

It is also easily seen that $L \subseteq \mathcal{E}_3$; and that there is a $T$ predicate in $L$. For many purposes $L$ is more convenient than either $E_2$ or $E_3$; this class was first considered by Cobham in 1964.

6. Decidability of theories. The formulas of a first order language can be specified as strings over a finite alphabet, by using devices such as representing variables as $xi_1 \ldots i_n$ where $i_j = 1, 2$. Functions and predicates on the formulas are easily defined by recursion on notation, and are in fact in $L$. In any case, a formula can be viewed as an integer; in honor of the fact that Kurt Gödel was the first person to realize this, the number is called the Gödel number of the formula. The specific numbering used is of little interest, except that typical functions and predicates on formulas should be at least primitive recursive.

Theories over a first order language provide examples of languages whose decidability is of interest (theories in higher order languages are also of interest, but we omit this). Recall that a theory in a first order language is a set of formulas in the language, which is closed under logical consequence. If $S$ is a recursively enumerable set of formulas and $T$ is the set of logical consequences then $T$ is recursively enumerable (exercise using Church’s thesis). If there is a finite $S$ then $T$ is said to be finitely axiomatizable. A recursively enumerable theory may or may not be recursive (decidable).

Almost by definition a theory $T$ is recursively enumerable iff the predicate $\text{Prf}_T(x)$, which holds iff $x$ is a proof in $T,$ is recursively enumerable. For natural theories $\text{Prf}_T$ is recursive, indeed in $L$; the axioms are easily recognized formulas, and it is easy to check that a sequence of formulas is a proof. Any recursively enumerable theory has a recursive set of axioms; if $F_i$ is the recursive enumeration let $A_i$ be the conjunction of the $F_j$ for $j \leq i$.

The formulas true in a class of structures (i.e., in each structure in the class) form a theory; this need not be recursively enumerable. Recall that a theory $T$ is complete if, for each sentence $F$ in the language, either $F$ or $\neg F$ is in $T$; the theory of a structure is complete, for example. A theory is consistent iff it is not the entire set of formulas in the language; this is so iff it does not contain both $F$ and $\neg F$ for some sentence $F.$ The theory of a class of structures is consistent.

Theorem 14.

a. A recursively enumerable consistent complete theory is recursive.

b. Suppose $T$ is recursively enumerable and there is a recursively enumerable $\bar{T}(i,x)$ such that, with $T_i = \{x : \bar{T}(i,x)\}, T_i$ is complete and $T = \cap_i T_i.$ Then $T$ is recursive.
PROOF: For part a, suppose \( T \) is recursively enumerable and \( F \) is a sentence. To decide if \( F \in T \), enumerate \( T \) until either \( F \) or \( \neg F \) is produced; accept \( F \) in the former case and reject it in the latter. Since \( T \) is consistent and complete exactly one of these sentences will eventually be enumerated. If \( F \) is a formula first replace it by its universal closure. For part b, to decide \( F \) simultaneously enumerate \( T \) and \( \bar{T} \). If \( F \) is enumerated in \( T \) report success, and if \( \neg F \) is enumerated in \( \bar{T} \) report failure. To see that this works, note that \( F \notin T \) iff \( \exists i(F \notin T_i) \) iff \( \exists i(\neg F \in T_i) \) iff \( \neg F \) is enumerated in \( \bar{T} \).

Part a is attributed to Janicak in [ELTT]. Variations of part b have been given by various authors. We will see below that there are recursive theories which are not complete.

Theories in the language of arithmetic were important in the history of mathematical logic, and remain indispensable. Let \( T \) be a theory in a language with a constant 0 and a unary function \( \text{Suc} \). The numeral for a nonnegative integer \( n \) is the closed term \( \text{Suc}(\cdots \text{Suc}(0) \cdots) \) where there are \( n \) occurrences of \( \text{Suc} \). Say that a predicate \( P(x) \) on \( \mathcal{N} \) is numeralwise reducible to \( T \) if there is a formula \( F(x) \) such that \( P(x) \iff F(x_1 \ldots x_k) \in T \) where \( x_i \) is the numeral for \( x_i \). The function mapping \( x \) to the formula is recursive. From this and the definitions, it is immediate that if \( T \) is recursively enumerable and every recursively enumerable predicate is numeralwise reducible to \( T \) then \( T \) is many-one complete.

For theories of interest stronger facts than the hypothesis above can be shown. Again supposing \( T \) to be a theory in a language with a constant 0 and a unary function \( \text{Suc} \), say that a function \( f(x) \) is representable in \( T \) if there is a formula \( F(x, y) \) such that

- if \( f(n) = m \) then \( \vdash_T \forall y(F(N, y) = M \iff y = M) \), and
- if \( f(n) \neq m \) then \( \vdash_T \neg F(N, M) \).

If \( T \) is recursively enumerable and consistent and \( f(x) \) is representable in \( T \) then \( f \) is recursive; if \( F \) is the formula, to compute \( f(n) \) enumerate \( T \) looking for a formula \( F(N, M) \), and if one is found output \( m \).

Theories where every recursive function is representable have interesting properties from the point of view of computability theory. Note that by representability of \( \text{Suc} \), if \( m \neq n \) then \( \vdash_T M \neq N \); also, if these formulas are provable the second requirement of representability is redundant.

**Lemma 15.** Suppose \( T \) is a theory where every recursive function is representable. Suppose \( F \) is a formula with one free variable. Then there is a closed formula \( G \) such that \( \vdash_T G \iff F(N_G) \) where \( N_G \) is the numeral for the Gödel number of \( G \).

PROOF: Let \( \text{Sub}(x, y) \) be the function which substitutes the term \( y \) (i.e., whose Gödel number is \( y \)) in the formula \( x \). The formula \( x \) is assumed to have a single free variable; if this is not the case, or if \( x \) or \( y \) is not well-formed, \( \text{Sub}(x, y) = 0 \). Let \( \text{Num}(x) \) be the function giving the (Gödel number of the) numeral for \( x \). Let \( H(x) \) be \( F(\text{Sub}(x, \text{Num}(x))) \), let \( N_H \) be the numeral for its Gödel number, and let \( G \) be \( H(N_H) \) (standard abbreviations for formulas involving defined functions are used). By representability \( \vdash \text{Sub}(N_H, \text{Num}(N_H)) = N_G \) and the lemma follows.

Call \( G \) the Gödel statement for \( F \); \( G \) states that it has the property \( F \), showing that self-reference is possible in such theories. In recursion theory it can be shown by a not dissimilar argument that there are programs which produce themselves as output. The lemma is a kind of proof-theoretic version of this.

Let \( \text{Thm}(x) \) be the theorem proved by \( x \) if \( \text{Prf}(x) \), else 0. Let \( \text{Prvs}(w, x) \) abbreviate \( \text{Prf}(w) \land \text{Thm}(w) = x \), and \( \text{Prvbl}(x) \) abbreviate \( \exists w \text{Prvs}(w, x) \). Let \( \text{Neg}(x) \) be the negation of the formula \( x \), with a resulting leading double negation removed. A theory \( T \) is said to be essentially undecidable if it is consistent and for any consistent theory \( T' \supseteq T \) (possibly in an expanded language; but see exercise 2), \( T' \) is undecidable. A theory \( T \) is \( \omega \)-consistent if there is no formula \( F(x) \) such that \( \vdash_T \exists x F(x) \) and for each numeral \( N \vdash \neg F(N) \). We will say that a theory \( T \) is a \( G_1 \)-theory if every recursive function is representable, \( \text{Prf}_T(x) \) is recursive, and \( T \) is consistent.

111
Theorem 16. Suppose $T$ is a $G_1$-theory. Let $F_G$ be the Gödel statement for $\neg \text{Prvbl}(x)$. Let $F_R$ be the Gödel statement for $\neg \exists w (\text{Prv}(w, x) \land \forall v < w \neg \text{Prv}(v, \text{Neg}(x)))$.

a. If $T$ is $\omega$-consistent then $T$ is many-one complete.
b. $T$ is essentially undecidable.
c. $F_G$ is not provable.
d. If $T$ is $\omega$-consistent then $\neg F_G$ is not provable.
e. $F_R$ is not provable.
f. If $T$ satisfies some mild further requirements then $\neg F_R$ is not provable.

Proof: To prove part a we show that every recursively enumerable predicate $W$ is numeralwise reducible to $T$. Indeed, $W(x)$ may be expressed as $\exists w R(w, x)$ where $R$ is recursive; if $R$ is represented by $F$ then $W$ is numeralwise reducible to $\exists w F$. Part b follows because every recursively enumerable predicate is reducible to any consistent $T' \supseteq T$, by the same argument. By lemma 15 $\vdash F_G \iff \neg \text{Prvbl}(N_{F_G})$. If $\vdash F_G$ then $\vdash \text{Prvbl}(N_{F_G})$ by representability; on the other hand the negation is also provable, contradicting consistency, which proves part c. If on the other hand $\vdash F_G$ then $\vdash \text{Prvbl}(N_{F_G})$. But by consistency there is no proof of $F_G$, whence by representability, for any numeral $M \vdash \neg \text{Prv}(M, N_{F_G})$. This contradicts $\omega$-consistency, proving part d. Let $\text{Prv}(w, x)$ denote $\text{Prv}(w, x) \land \forall v < w \neg \text{Prv}(v, \text{Neg}(x))$, and $\text{Prvbl}(x)$ denote $\exists w \text{Prv}(w, x)$. If $\vdash F_R$ then $\vdash \text{Prvbl}(N_{F_R})$, by consistency and representability. Part e follows as part c. If $\vdash \neg F_R$ then $\vdash \text{Prvbl}(\neg F_R)$; also $\neg \text{Prvbl}(N_{F_R})$. With mild requirements (provability of trichotomy and single-valuedness of Thm) the provability of both statements contradicts consistency. This proves part f.

See [Rosser] for some background on this theorem. Parts c and d were proved first, by Gödel, for the theory Peano arithmetic (defined below), and show that Peano arithmetic is incomplete. The Gödel statement is “I am not provable”. Part a was proved slightly later, and implies incompleteness, without giving a specific formula. Parts e and f were also proved slightly later, by Rosser; Peano arithmetic satisfies the additional restrictions. Mostowski and Tarski found a simpler theory satisfying the hypotheses, and drew some conclusions in algebra.

The simpler theory, usually called Q, is in the language with 0, $\text{Suc}$, $+$, $\cdot$, and $\equiv$. The axioms are those of equality, and

1. $\text{Suc}(x) = \text{Suc}(y) \Rightarrow x = y$
2. $\text{Suc}(x) \neq 0$
3. $x \neq 0 \Rightarrow \exists y (x = \text{Suc}(y))$
4. $x + 0 = x$
5. $x + \text{Suc}(y) = \text{Suc}(x + y)$
6. $x \cdot 0 = 0$
7. $x \cdot \text{Suc}(y) = (x \cdot y) + x$

If $h(w, x)$ is a function its minimization is defined to be the partial function $\phi(x)$ whose value is the smallest $w$ such that $h(w, x) = 0$ if there is such; else undefined. If $\phi$ is total the minimization is said to be total. The Gödel pairing function (also called Cantor’s) $\text{Pr}(x, y)$ is defined to be $((x + y)(x + y + 1))/2 + y$. This is a convenient bijection between $\mathcal{N}$ and $\mathcal{N} \times \mathcal{N}$. We use $\text{Pr}_1(x)$ and $\text{Pr}_2(x)$ for the first and second components of the pair corresponding to $x$.

Lemma 17. The recursive functions are the smallest class containing 0, $\text{Suc}$, $+$, $\cdot$; and closed under substitution and total minimization.

Proof: Using the Kleene T predicate, every recursive function is obtained by total minimization of a primitive recursive function, and composition with further primitive recursive functions. Thus it suffices
to show that the class is closed under primitive recursion. Let \( \beta(c, d, i) \) be \( c \mod 1 + (i + 1)d \). Given a sequence \( a_0, \ldots, a_{k-1} \) of nonnegative integers there are \( c \) and \( d \) such that \( \beta(c, d, i) = a_i \) for \( 0 \leq i < k \). Indeed, let \( d = (k - 1)! \); then the numbers \( 1 + (i + 1)d \) are pairwise relatively prime. This follows because a prime divisor of \( 1 + (i + 1)d \) must be greater than \( k - 1 \); and a prime divisor of two such values must divide their difference. Now by the Chinese remainder theorem \( c \) may be chosen (this trick is due to Gödel).

Let \( \gamma(x, i) = \beta(\Pr_1(x), \Pr_2(x), i) \), so that \( c \) and \( d \) are coded by a single value. Suppose \( f \) is defined by \( f(0, y) = g(y) \) and \( f(\text{Suc}(x), y) = h(x, y, f(x, y)) \). Using \( \mu \) to denote minimization, let

\[
\text{Vseq}_f(x, y) = \mu s(\gamma(s, 0) = g(y) \land \mu i(\gamma(s, i) \neq h(x, y, \gamma(s, i))) \geq x).
\]

Then \( f(x, y) = \gamma(\text{Vseq}_f(x, y), x) \).

This lemma does not provide an effective enumeration of the recursive functions; indeed because it cannot follow it that it is not decidable if a minimization is total.

**Theorem 18.** A function is recursive iff it is representable in \( Q \).

**Proof:** Note that \( Q \) is consistent since the structure \( N \) is a model. It is also recursively enumerable (in fact finitely axiomatizable). Hence if \( f \) is representable it is recursive, by an earlier observation. The other direction proceeds by induction using lemma 17: \( 0 \) is represented by \( y = 0 \), \( \text{Suc} \) by \( y = \text{Suc}(x) \), \( + \) by \( y = x_1 + x_2 \), and \( \cdot \) by \( y = x_1 \cdot x_2 \). These claims are verified by proving various equalities and inequalities between closed terms, by induction. If \( f \) is represented by \( F \) and \( t_i \) by \( G_i \), for \( 1 \leq i \leq k \) then \( (t_1, \ldots, t_k) \) is represented by (showing only the “dependent” variable for \( G_i \))

\[
\exists z_1 \ldots \exists z_k(G_1(z_1) \land \cdots \land G_k(z_k) \land F(z, y)).
\]

This follows by predicate logic from the induction hypothesis. Let \( x \leq y \) be an abbreviation for \( \exists u(y = x + u) \).

If \( h(w, x) \) is represented by \( H(w, x, y) \) then \( \mu w h(w, x) = 0 \) is represented by

\[
H(w, x, 0) \land \forall v(H(v, x, 0) \Rightarrow w \leq v).
\]

This follows by predicate logic from the induction hypothesis and some facts about \( \leq \), namely \( x \leq N \Rightarrow x = 0 \land \cdots \land x = N \) and \( \neg N \leq x \Rightarrow x = 0 \land \cdots \land x = N - 1 \). The first follows by induction on \( N \); the second follows from the first and \( x \leq N \land N \leq x \), which follows by induction on \( N \).

Peano arithmetic, or \( PA \), has as axioms those of \( Q \) except number 3; and the axioms

\[
F(0) \land \forall x(F(x) \Rightarrow F(\text{Suc}(x))) \Rightarrow \forall x F(x)
\]

for each formula \( F \) in at least one free variable (denoted \( x \) in the axiom). A system of axioms of this kind is called an axiom scheme, in this particular case the first order induction scheme. An informal version allows any set \( F \); this can be formalized within set theory (or weaker systems). It is not difficult to show that the informal system has only \( N \) for a model. Axiom 3 of \( Q \) is readily provable by induction. From this and the fact that \( N \) is a model of \( PA \) it readily follows that both \( Q \) and \( PA \) are \( G_1 \)-theories.

Another system of interest has a symbol for each primitive recursive function. In order for the system of axioms to be recursive the functions symbols must be numbered so that the definition of each is readily computed. One approach is to use “combinators” for substitution and primitive recursion, with the symbol for a function being the term involving the combinators; details are left to the reader.

The axiom system for the primitive recursive functions is called primitive recursive arithmetic, or \( PRA \). In one formulation the axioms are those of equality; the function definition axioms (in obvious notation) \( f_t(x) = t \), \( f_g(0, y) = g(y) \), \( f_{gh}(\text{Suc}(x), y) = h(x, y, f_{gh}(x, y)) \); the induction scheme for open formulas; and \( \neg \cdot 0 = 1 \). \( PRA \) is often given in other formulations; but these are equivalent. The defining equations for \( \text{Pred} \) may be used to show that axioms 1 and 2 of \( Q \) hold, and it follows that \( PRA \) is a \( G_1 \)-theory.

Theorem 16.c is the “first incompleteness theorem”, that for any sufficiently well behaved axiom system \( T \) for arithmetic there is a true statement of arithmetic which is not provable \( T \). The second incompleteness theorem states that indeed, the consistency of \( T \) is such a statement. The article of C. Smorynski in [HML]
gives a discussion of ramifications that this had for the foundations of mathematics in 1931, when it was stated without a complete proof by Gödel.

Additional requirements are placed on \( T \) for the second incompleteness theorem, called the “derivability conditions” in the above cited article. Let \( \Box F \) abbreviate \( \text{Prvbl}(F) \) where \( F \) denotes a formula, in a way to be described shortly. We will say that a \( G_1 \)-theory is a \( G_2 \)-theory if it satisfies the additional requirements

1. \( \vdash_T \Box(F \Rightarrow G) \Rightarrow \Box F \Rightarrow \Box G. \)
2. \( \vdash_T \Box F \Rightarrow \Box \Box F. \)

In a \( G_1 \) theory, if \( \vdash_T F \) then \( \vdash_T \Box F, \) where \( F \) denotes a numeral in the second expression. In condition 1, \( F \) and \( G \) are variables and \( F \Rightarrow G \) is an abbreviation for a function giving this formula from its subformulas. In condition 2, \( F \) is a variable and the inner \( \Box F \) is an abbreviation for a function giving this formula from \( F. \)

PA is a \( G_2 \)-theory. The proof is lengthy and is omitted. One strategy is to prove that PRA can be simulated by PA; and that the derivability conditions for PA (or PRA) are provable in PRA. This in turn is facilitated by proving that recursion and induction on notation can be added to PRA. Finally, a formalized representability theorem can be used to prove the second derivability condition. PRA is also a \( G_2 \)-theory.

**Theorem 19.** (Löb’s theorem.) Suppose \( T \) is a \( G_2 \)-theory. If \( \vdash_T \Box F \Rightarrow F \) then \( \vdash_T F. \)

**Proof:** Using lemma 15 let \( G \) be such that \( \vdash G \Leftrightarrow (\Box G \Rightarrow F); \) then \( \vdash \Box G \Rightarrow \Box G \Rightarrow F), \) whence \( \vdash \Box G \Rightarrow \Box G \Rightarrow F. \) But \( \vdash \Box G \Rightarrow \Box G, \) so \( \vdash \Box G \Rightarrow \Box F, \) and so \( \vdash \Box G \Rightarrow F. \) Thus \( \vdash_G, \) so \( \vdash G, \) so \( \vdash F. \)

**Corollary 20.** If \( T \) is a \( G_2 \)-theory then \( \neg \Box \perp \) is not provable in \( T. \)

**Proof:** Apply Löb’s theorem with \( F \) being \( \perp. \)

The article of J. Paris and L. Harrington in [HML] describes a result from 1977, that a combinatorial principle which is a modification of Ramsey’s theorem implies the consistency of PA. It follows by second incompleteness that this statement is not provable in PA.

**Theorem 21.** The theory of \( \mathcal{N} \) (as usual in the language \( 0, \text{Suc}, +, \cdot, = \)) is not definable in \( \mathcal{N}. \)

**Proof:** If \( F \) were a formula defining the set of true formulas, the Gödel statement of \( \neg F \) would be a statement which could be neither true nor false.

From the representability of the recursive functions in \( Q \) it follows that the recursively enumerable predicates are definable. In section 8 it will be seen that they are definable using \( \Sigma_1 \) formulas. This fact can be proved much more easily if the \( \Sigma_0 \) formulas are expanded to define a robust subrecursive class, for example with the primitive recursive functions, or with “bounded quantifiers” \( \forall x \leq y \) and \( \exists x \leq y. \) In any case, the theory of \( \mathcal{N} \) is not recursively enumerable.

If \( \cdot \) is omitted from the language the theory is decidable; this is proved in chapter 28.

Theorem 16 can be appealed to to deduce that other theories are undecidable, by reduction. For example, by lemma 22 below validity is undecidable for the language \( 0, \text{Suc}, +, \cdot, =. \) Development of these methods was initiated by Tarski and others in the late 1940’s [TRM]. For later use, say that a theory \( T \) is hereditarily undecidable if for any theory \( T' \subseteq T \) in the given language, \( T' \) is undecidable.

**Lemma 22.** Suppose \( T_1 \) is a theory in a language \( L, \) \( F \) is a sentence in \( L, \) and \( T_2 \) is the logical consequences of \( T_1 \cup \{F\}. \) If \( T_2 \) is undecidable then \( T_1 \) is.

**Proof:** Using the deduction theorem, \( G \in T_2 \iff F \Rightarrow G \in T_1 \), giving a many-one reduction from \( T_2 \) to \( T_1. \)

An interpretation of a first order language \( L_1 \) in a language \( L_2 \) consists of the following:

1. For each constant \( c \) of \( L_1 \) a formula \( F_c(x) \) of \( L_2. \)

114
2. For each function \( f(x) \) of \( L_1 \) a formula \( F_f(x, y) \) of \( L_2 \).
3. For each relation \( R(x) \) of \( L_1 \) a formula \( F_R(x) \) of \( L_2 \).

The interpretation requirements \( R_I \) are defined to be the formulas \( y = c \iff F_c(y) \), \( y = f(x) \iff F_f(x, y) \), and \( R(x) \iff F_R(x) \).

For each formula \( G \) of \( L_1 \) an interpretation \( G_I \) may be defined; it suffices to define this for atomic formulas, with more general formulas being translated by translating their atomic subformulas. If an atomic formula contains untranslated occurrences of function symbols, there is one where the arguments are all variables; writing the formula as \( R(f(x)) \), translate it to \( \exists y(F_f(x, y)) \). This step (and a similar one for constants) is applied until all occurrences of constant and functions are translated. Finally, \( R(x) \) is translated to \( F_R(x) \). Throughout, bound variables are renamed so that free variables do not become bound. Clearly, the translation from \( G \) to \( G_I \) is recursive (indeed primitive recursive).

**Theorem 23.** Suppose \( T_1 \) is a theory in the language \( L_1 \) which is finitely axiomatizable and essentially undecidable. Suppose \( T_2 \) is a theory in a language \( L_2 \) disjoint from \( L_1 \), \( I \) is an interpretation of \( L_1 \) in \( L_2 \), and \( T_1 \cup T_2 \cup R_I \) is consistent. Then \( T_2 \) is hereditarily undecidable.

**Proof:** Let \( T_3 \) be the theory generated by \( T_1 \cup T_2 \cup R_I \); being a consistent extension of \( T_1 \) it is undecidable. Let \( T_4 \) be the formulas of \( T_3 \) in \( L_2 \). The formula \( F_I \) may be defined for any formula in \( L_1 \cup L_2 \) (by ignoring symbols in \( L_2 \)). Further \( F \in T_3 \) iff \( F_I \in T_4 \), because \( R_I \) can be used to translate in either direction. Since this is a many-one reduction of \( T_3 \) to \( T_4 \), \( T_4 \) is undecidable. Let \( \exists x F(X) \) be an abbreviation for \( \exists x F(x) \land \forall y (F(y) \Rightarrow y = x) \). Let \( U_I \) be the sentences \( \exists y F_c(y) \land \forall x \exists y F_f(x, y) \), for the constants and functions of \( L_1 \). Let \( T_5 \) be the theory generated by \( T_2 \cup U_I \). A \( T \) is the conjunction of a finite set of axioms for \( T_1 \). We claim that for \( F \) in \( L_2 \), \( F \in T_4 \) iff \( A_1 \models F \in T_5 \). Now, \( U_I \subseteq T_3 \), so \( T_5 \subseteq T_3 \); also \( A_1 \in T_3 \). Thus if \( A_1 \models F \in T_5 \) then \( F \in T_3 \), so \( F \in T_4 \). The forward direction uses a “model-theoretic” argument. Suppose \( F \in T_4 \) and \( M \) is a model of \( T_5 \) and \( A_1 \). Since \( M \) is a model of \( U_I \) it may be expanded to a structure \( M' \) for \( L_1 \cup L_2 \) which satisfies \( R_I \). In fact \( M' \) is a model of \( T_3 \) since it is a model of \( A_1 \) and hence of \( A \). Thus \( M' \) is a model of \( F \), and so \( M \) is a model of \( F \). By the claim just proved, \( T_4 \) is many-one reducible to \( T_5 \) and so \( T_5 \) is undecidable. \( T_5 \) is obtained from \( T_2 \) by adding a sentence, so by lemma \( 22 \) \( T_2 \) is undecidable. If \( T_2 \) is replaced by a subtheory the hypotheses still hold.

The requirement that \( L_2 \) be disjoint from \( L_1 \) can be relaxed by renaming. For example, as a corollary, under the hypotheses on \( L_1 \), if \( L_2 \supseteq L_1 \) and \( T_1 \cup T_2 \) is consistent then \( T_2 \) is hereditarily undecidable. Note that \( Q \) can serve as \( T_1 \); but to obtain interesting reductions a further step is necessary, namely to restrict variables in the interpretation by some predicate. A neat way to do this is as follows.

Let \( L \) be a first order language and \( R \) a unary predicate symbol not in \( L \). The formula \( F^{(R)} \) in \( L \cup \{ R \} \) is obtained by first recursively replacing subformulas of the form \( \forall x G(x) \) (resp. \( \exists x G(x) \)) by \( \forall x (R(x) \Rightarrow G(x)) \) (resp. \( \exists x (R(x) \land G(x)) \)). Finally \( F^{(R)} \) must be prefixed with “\( R(x_1) \land \ldots \land R(x_n) \Rightarrow \)” where \( x_1, \ldots, x_n \) are the free variables. If \( S \) is a set of formulas let \( S^{(R)} = \{ F^{(R)} : F \in S \} \).

**Lemma 24.** Let \( L \) be a first order language and \( R \) be a unary predicate symbol in \( L \). Given a theory \( T \) in \( L \) let \( T' \) be the theory generated by \( T^{(R)} \). Suppose \( A \) is a system of axioms for \( T \); let \( A' \) consist of \( A^{(R)}, \exists x R(x) \), and the formulas \( R(c) \) for constants \( c \) and \( R(x_1) \land \ldots \land R(x_n) \Rightarrow R(f(x_1, \ldots, x_n)) \) for functions \( f \).

a. \( T \) is consistent iff \( T' \) is.
b. If \( T \) is essentially undecidable then \( T' \) is.
c. \( A' \) is a system of axioms for \( T' \).

**Proof:** Suppose \( T \) is consistent, and let \( M \) be a model. Expand \( M \) by interpreting \( R \) as the everywhere true predicate; the result is a model of \( T^{(R)} \) and hence of \( T' \). Thus \( T' \) is consistent. Suppose \( T' \) is consistent, and let \( M \) be a model. A structure for \( L \) may be obtained by restricting to the interpretation of \( R \); the
result is a model of $T$. Thus $T$ is consistent. For part b, suppose $T$ is essentially undecidable; we have already shown that $T'$ is consistent; suppose $U$ is a consistent extension. Let $U_1$ be the sentences $F$ in $L$ such that $F^{(R)} \in U$. Suppose $F_1, \ldots, F_k, G$ are in $U_1$ and $F_1, \ldots, F_k \vdash G$ in $L$; then $F_1 \land \cdots \land F_k \Rightarrow G$ is valid, so $F_1^{(R)} \land \cdots \land F_k^{(R)} \Rightarrow G^{(R)}$ is valid and hence in $U$, whence, since $F_1^{(R)}$ is in $U$, $G^{(R)}$ is in $U$; hence $G$ is in $U_1$. The theory $U_2$ generated by $U_1$ has $U_1$ as its sentences. From this we see that $U_2^{(R)} \subseteq U$, whence $U_2$ is consistent. Also, $T \subseteq U_2$, whence $U_2$ is undecidable. The map $F \mapsto F^{(R)}$ is a many-one reduction of $U_2$ the $U$, whence $U$ is undecidable. For part c, clearly $A^{(R)}$ is in $T'$; $\exists x \top$ is valid, so $\exists x R(x)$ is in $T'$; and similarly there are valid formulas from which it follows that the remaining axioms of $A'$ are in $T'$. It remains to show that if $F$ is provable from $A$ then $F^{(R)}$ is provable from $A'$.

Indeed if $F_1, \ldots, F_l$ is a proof from $A$ then $F_1^{(R)}, \ldots, F_l^{(R)}$ can be filled out to a proof from $A'$; this is an exercise in predicate logic left to the reader.

Consider the theory $\text{Th}(Z)$ of $Z$, in the language 0, 1, $+, \cdot$ (although $+, \cdot$ can be shown to suffice). The language of $Q$ has an obvious interpretation. Lagrange’s “four square” theorem states that an integer is nonnegative iff $\exists a \exists b \exists c \exists d (x = a^2 + b^2 + c^2 + d^2)$ (this is proved in chapter 20). It follows that $\text{Th}(Z)$ is hereditarily undecidable. Since $Z$ is a ring, the theory of rings is undecidable; the same applies to the theory of commutative rings.

In [Robinson] it is shown that the predicate “$x$ is an integer” is definable in the structure $Q$, with the language 0, 1, $+, \cdot$. Since the proof involves some nontrivial facts from number theory we will not give it and refer the reader to the paper, or [FW]. It follows that $\text{Th}(Q)$ is hereditarily undecidable, whence the theory of fields is undecidable.

Let $A_n$ denote the sentence

$$\forall c_0 \ldots \forall c_n \exists x (c_0 + \cdots + c_n x^n = 0).$$

The theory of algebraically closed fields (ALG) is axiomatized by the field axioms, and the axioms $A_n$ for $n \geq 1$. Let $\text{ALG}_p$, for a prime number $p$ have the axiom $1 + \cdots + 1 = 0$ added; this is the theory of algebraically closed fields of characteristic $p$. The theory $\text{ALG}_0$ of algebraically closed fields of characteristic 0 is obtained by adding $1 + \cdots + 1 \neq 0$ for all $n \geq 2$.

The following discussion uses infinite cardinal numbers. We allow a language to be infinite. By set theory, if the language has infinite cardinality $\kappa$ so does the set of formulas. We also need a definition. Recall from chapter 11 that if $Q \subseteq R$ are structures in some language, $R$ is an elementary extension of $Q$ (or $Q$ an elementary substructure of $R$) if for a sentence $F$ in the language expanded with constants for the elements of $Q$, $\models_Q F$ iff $\models_R F$. The names given in the following lemma have various variations.

**Lemma 25.** Suppose $L$ is a first order language and $T$ a theory in $L$.

a. (Lowenheim-Skolem theorem) $T$ has a model of cardinality $\kappa$, where $\kappa = \omega$ if $L$ is finite, else $|L|$.

b. If $T$ has infinite models, and $\kappa$ is an infinite cardinal with $\kappa \geq |L|$, then $T$ has a model of cardinality $\kappa$.

c. (Downward Lowenheim-Skolem theorem) Suppose $S$ is a structure in a language $L$, $Q \subseteq S$ is a subset, and $\kappa$ is an infinite cardinal with $|Q| \leq \kappa \leq |S|$. Then $S$ has an elementary substructure $R$ with $Q \subseteq R$ and $|R| = \kappa$.

d. (Upward Lowenheim-Skolem theorem) Suppose $Q$ is an infinite structure for $L$ and $\kappa$ is an infinite cardinal with $\kappa \geq |Q|$ and $\kappa \geq |L|$; then $Q$ has an elementary extension of cardinality $\kappa$.

e. (Los-Vaught test) Suppose $T$ is a theory with no finite models, and there is a infinite cardinal $\kappa$ such that all models of $T$ of cardinality $\kappa$ are isomorphic. Then $T$ is complete.

**Proof:** Part a follows by the construction in the proof of theorem 11.12; this holds for infinite languages, and the cardinality of the constructed model is as stated. For part b, add to $T$ the sentences $c_\alpha \neq c_\beta$ for
\( \alpha < \beta < \kappa \) where the \( c_\alpha \) are new constants. Every finite subset of the resulting theory \( T' \) has a model, because any infinite model of \( T \) can be expanded to one. Thus, \( T' \) is consistent and by part a has a model of cardinality \( \kappa \). For part c, let \( S_0 \) be a subset of \( S \) of cardinality \( \kappa \) which contains \( Q \). Having \( S_1 \), for each sentence \( \exists x F \) in the language expanded with constants for \( S_1 \) which is true in \( S \) choose an element of \( S \) which can serve as \( x \); let \( S_{i+1} \) be \( S_i \) with these elements added. Let \( R \) be the union of the \( S_i \) for \( i \in \mathcal{N} \). \(| R | = \kappa \) because there are only \( \kappa \) many formulas at each stage so only \( \kappa \) many constants added at each stage; and so only \( \kappa \) many in total. \( R \) is closed under the functions of \( S \), and is a substructure when equipped with the restrictions of the relations. By theorem 11.14 it is an elementary substructure. For part d, let \( T_1 \) be the theory of \( Q \) in the language with constants for \( Q \) added. Form \( T_2 \) by adding \( c_\alpha \neq c_\beta \) for \( \alpha < \beta < \kappa \) where the \( c_\alpha \) are new constants. As above, \( T_2 \) is consistent; if \( M \) is a model \( Q \) may be considered as a substructure by identifying \( q \in Q \) with the interpretation of \( c_q \). \( M \) is clearly an elementary extension of \( Q \), of cardinality at least \( \kappa \). By part c \( M \) has an elementary substructure \( R \) of cardinality \( \kappa \) containing \( Q \), and \( R \) is readily seen to be an elementary extension of \( Q \). For part e, suppose to the contrary that \( F \) is a sentence such that both \( T \cup \{ F \} \) and \( T \cup \{ \neg F \} \) are consistent. By the Lowenheim-Skolem theorem and the hypothesis that \( T \) has no finite models, countable models \( M_1 \) and \( M_2 \) may be chosen for the two sets. These have elementary extensions \( M_1' \) and \( M_2' \) of cardinality \( \kappa \); by hypothesis these are isomorphic, a contradiction.

**Lemma 26.** Any two algebraically closed fields of the same characteristic and the same uncountable cardinality are isomorphic.

**Proof:** We appeal to facts from chapter 20. Given an algebraically closed field \( F \), let \( X \) be a transcendence base for \( F \) over the prime subfield \( P \). By hypothesis \( P \) is determined; the extension by \( X \) depends only on \(|X|\), and so its algebraic closure \( F \) does also. If \( F \) is uncountable then \( X \) must be, and \(|X| = |F|\).

**Theorem 27.** \( \text{ALG} \) is decidable.

**Proof:** By lemma 26 and lemma 25e \( \text{ALG}_0 \) and \( \text{ALG}_p \) for prime \( p \) are complete. The theorem follows using theorem 14b.

\( \text{ALG} \) is clearly not complete. Also, the theory of fields is not essentially undecidable. \( \text{ALG}_0 \) and \( \text{ALG}_p \) have the property known as model completeness; completeness may be shown from this and from the fact that there is a minimal model (the algebraic closure of the prime field); see [CK]. In chapter 28 an explicit decision procedure will be given for another decidable extension of the theory of fields, real closed fields.

**7. \( \mathcal{NP} \) completeness of satisfiability.** The class \( \mathcal{P} \) is defined to be the predicates computable in polynomial time by a Turing machine. These may be effectively enumerated as pairs \( \langle M, p \rangle \) where \( M \) is a Turing machine and \( p \) a polynomial. The Turing machine must have some convention for when it reports success or failure; and it is considered to fail if it runs for \( p(\ell(x)) \) steps without reporting success. Also, \( M \) has an indication of its number of arguments built in to its coding.

As indicated in earlier sections there is a “polynomial time” version of Church’s thesis, that a predicate may be demonstrated to be in \( \mathcal{P} \) by giving an informally presented algorithm and a satisfactory argument that it runs in polynomial time. As an example, let \( \text{PformVal}(F, A) \) be the predicate which is true iff \( A \) is a truth assignment to the atoms of the propositional formula \( F \) which satisfies \( F \). To simplify the argument, we suppose that \( F \) is presented in “prefix” form, that is, a formula \( F \) is one of:

- an atom, given as an integer in dyadic notation;
- \( \neg F_1 \) where \( F_1 \) is a subformula; or
- \( CF_1 F_2 \) where \( C \) is a binary connective and \( F_1 \) and \( F_2 \) are subformulas.

\( A \) is a string of 0’s and 1’s, with the \( i \)th from the right being the value assigned to atom \( i \); atoms whose number is greater than or equal to \( \ell(A) \) are assigned 0.
LEMMA 28. \(P\text{formVal}\) is in \(\mathcal{P}\).

PROOF: We give an informal argument using Turing machines. In the first step, the value of each atom is placed on a track, at the start of the atom. This involves copying the atom, counting down and advancing through \(A\), and cleaning up, using additional tracks as necessary. In the second step, each atom is replaced by its value. The third step is an iteration of the following step. The leftmost subformula of the form \(Cuv\) where \(C\) is a connective and \(u, v\) truth values is found and replaced by its value.

A circuit is a sequence of equations \(p \equiv \neg q\) or \(p \equiv qCr\) where \(C\) is a binary propositional connective. The atoms are either input atoms or defined atoms; atoms on the left of an equation must be defined atoms. Atoms on the right are either input atoms or defined atoms which occur earlier; an atom may only be defined once. The value problem for circuits is also in \(\mathcal{P}\), but we shall not require this.

Given a predicate \(P(x_1, \ldots, x_k)\) and an integer \(n\), a Boolean function \(P_n\) on \(nk\) arguments may be defined. For \(0 \leq j < n\) and \(0 \leq i < k\) argument \(p_{ni+j}\) represents bit \(j\) of \(x_i\); binary notation is more convenient for defining these functions than dyadic.

LEMMA 29. Suppose \(P(x_1, \ldots, x_k)\) is a predicate in \(\mathcal{P}\) and \(P_n\) are the Boolean functions of the preceding paragraph. There is a function \(\text{Rep}_{P_n}(w)\) in \(L\) which, on any input \(w\) of length \(n\) produces (the Gödel number of) a circuit which computes the function \(P_n\).

PROOF: Suppose \(P\) is computed by Turing machine \(M\) in time \(p(n)\) where \(n = \max(\ell(x_1), \ldots, \ell(x_k))\) and \(p\) is a suitable polynomial. The head is required to remain within a range of \(p(n)\) squares. The circuit computing \(p_n\) consists of \(p(n)\) identical layers; layer \(i\) simulates step \(i\). A layer consists of \(p(n)\) blocks, one for each square. A block has \(3r\) inputs and \(r\) outputs, where \(r\) equals \([q + 1 + s]\) where \(q\) is the number of states (with an extra letter for the head not being on the square) and \(s\) the number of tape symbols. The circuit can clearly be written down in polynomial time by a Turing machine.

A predicate \(P(x)\) is said to be in \(NP\) if there is a predicate \(Q(w, x)\) in \(\mathcal{P}\) and a polynomial \(p\), such that \(P(x)\) holds iff \(Q(w, x)\) holds for some \(w\) with \(\ell(w) \leq p(\ell(x))\). The value \(w\) is a “witness” that \(P(x)\); given a witness (which is from a set of exponential size) it can be verified in polynomial time if the predicate holds.

A predicate \(P(x)\) is polynomial time many-one reducible to a predicate \(Q(y)\) iff there is a function \(f \in L\) such that \(P(x)\) iff \(Q(f(x))\). A predicate is \(NP\) complete if it is in \(NP\), and any other predicate in \(NP\) is polynomial time many-one reducible to it. Let Sat be the predicate which is true if the input is a satisfiable formula of the propositional calculus.

THEOREM 30. \(\text{Sat}\) is \(NP\) complete.

PROOF: That \(\text{Sat}\) is in \(NP\) follows by lemma 28; the witness is simply the satisfying truth assignment. If \(P\) is in \(NP\) and \(x\) is an input value, the circuit \(C_n\) may be computed where \(n = \ell(x)\), in polynomial time. This turn can be transformed to a formula which is satisfiable iff \(P(x)\). Namely, replace the input atoms by the bits of \(x\), and let the formula be the conjunction of the equations.

It is open whether \(\text{Sat}\) is in \(\mathcal{P}\). A wide variety of problems, including “real world” optimization problems, are known to be \(NP\) complete; [GJ] is a standard reference on the subject, and an example is given in chapter 28. On the other hand various problems have been shown after some effort to be in \(\mathcal{P}\). Two important examples are linear programming [Wright] (see chapter 22 for a description of the problem) and primality [Granville]. Some other examples are given in chapter 28.

8. **Hilbert’s tenth problem.** A predicate \(R(x)\) in \(N^k\) is said to be Diophantine if there is a multivariate polynomial \(p(x, y)\) with coefficients in \(Z\) such that \(R = \{x \in N^k : p(x, y) = 0\text{ for some }y \in N^l\}\). In this section it will be shown that every recursively enumerable predicate is Diophantine (the converse is
Proof: and thus, to show that a predicate is Diophantine it suffices to give a monotone Boolean combination of predicates already known to be Diophantine which defines the predicate. Inequality, congruence, and divisibility are readily seen to be Diophantine.

The crucial step in proving the claim, and the last accomplished historically (by Y. Matiyasevich in 1970), is to show that the graph of the exponential function is Diophantine. A trick of L. Adelman can then be used to simulate Turing machines, although the original methods were more involved. To begin with, the solutions of a particular Diophantine equation (Pell’s equation) are shown to be Diophantine; again, this is a simpler method than the original, due to G. Chudnovsky. Pell’s equation is the equation \( x^2 - dy^2 = 1 \); the behavior is particularly simple when \( d = a^2 - 1 \) for some \( a > 1 \), which will be assumed from hereon.

**Lemma 31.** The solutions \((x, y) \in \mathbb{N}^2\) of Pell’s equation form a commutative monoid under the multiplication law \((x_1 + y_1 \sqrt{d})(x_2 + y_2 \sqrt{d}) = (x_3 + y_3 \sqrt{d})\). The solution \((a, 1)\) is a generator.

**Proof:** First, \((x_1 + y_1 \sqrt{d})(x_2 + y_2 \sqrt{d}) = (x_3 + y_3 \sqrt{d})\) if \((x_1 - y_1 \sqrt{d})(x_2 - y_2 \sqrt{d}) = (x_3 - y_3 \sqrt{d})\). Second, \((x, y)\) is a solution iff \((x + y \sqrt{d})(x - y \sqrt{d}) = 1\). The first assertion follows. There is no solution \((x, y) \in \mathbb{Z}^2\) with \(1 < x + y \sqrt{d} < a + \sqrt{d}\), else \(1 > x - y \sqrt{d} > a - \sqrt{d}\), and \(0 < y \sqrt{d} < \sqrt{d}\), a contradiction. For any solution \((x, y) \in \mathbb{N}^2\), there is some \(n\) such that \((a + \sqrt{d})^n \leq x + y \sqrt{d} < (a + \sqrt{d})^{n+1}\). Writing \(x_n + y_n \sqrt{d}\) for \((a + \sqrt{d})^n\), \(1 \leq (x + y \sqrt{d})(x_n - y_n \sqrt{d}) < (a + \sqrt{d})\), and equality must hold, proving the second assertion.

From hereon \((a + \sqrt{d})^n\) for \(n \in \mathbb{N}\) will be written as \(x_n(a) + y_n(a) \sqrt{d}\), or simply \(x_n + y_n \sqrt{d}\) when \(a\) is clear.

**Lemma 32.** Suppose \(m \leq n\) when \(n - m\) is considered.

\(a.\) \(x_{n \pm m} = x_n x_m \pm dy_n y_m,\) \(y_{n \pm m} = x_m y_n \pm x_n y_m\)

\(b.\) \(x_{n \pm 1} = ax_n \pm dy_n,\) \(y_{n \pm 1} = ay_n \pm x_n\)

\(c.\) \(x_{n+1} = 2ax_n - x_{n-1},\) \(y_{n+1} = 2ay_n - y_{n-1}\)

**Proof:** Part \(a\) follows because \((x_{n \pm m} + y_{n \pm m} \sqrt{d}) = (x_n + y_n \sqrt{d})(x_m \pm y_m \sqrt{d})\).

Part \(b\) is a special case, and part \(c\) follows.

**Lemma 33.** Suppose \(n \in \mathbb{N}\).

\(a.\) \(\gcd(x_n, y_n) = 1\)

\(b.\) \(x_{n+1} > x_n,\) \(y_{n+1} > y_n,\) \(a^n \leq x_n \leq (2a)^n,\) and for \(n > 0\) \((2a - 1)^{n-1} \leq y_n \leq (2a)^{n-1}\)

\(c.\) \(n\) is even iff \(y_n\) is even iff \(x_n\) is odd

\(d.\) \(y_n \mid y_t\) iff \(n \mid t\)

\(e.\) \(y_n^2 \mid y_{nt}\), and if \(y_n^2 \mid y_t\) then \(y_n \mid t\)

\(f.\) \(y_n \equiv n \mod a - 1\)

\(g.\) if \(a \equiv b \mod c\) then \(x_n(a) \equiv x_n(b) \mod c\) and \(y_n(a) \equiv y_n(b) \mod c\)

\(h.\) for \(n > 0\), if \(y_1 \equiv y_j \mod x_n\) then \(j \equiv i \mod 2n\)

**Proof:** Part \(a\) is obvious from Pell’s equation. Part \(b\) follows by induction using lemma 32b and 32c. Part \(c\) follows by induction using part \(a\) and lemma 32c. For part \(d\), by induction on \(k\) using lemma 32a, \(y_n \mid y_{nk}\)

Writing any \(t\) as \(nk + r\) where \(r < n\), \(y_k = x_k y_{nk} + y_r x_{nk}\). If \(y_n \mid y_t\) then \(y_n \mid y_t x_{nk}\), whence by part \(a\) and \(y_n \mid y_{nk}, y_n \mid y_t\) by part \(b\) \(r = 0\). For part \(e\), \(x_n y_n + y_n y_{\sqrt{d}} = (x_n + y_n \sqrt{d})^{y_n}\), and the terms in the binomial expansion for the right side involving \(\sqrt{d}\) (which is irrational) are divisible by \(y_n^2\), so \(y_n^2 \mid y_{nk}\). If \(y_n^2 \mid y_t\) then by part \(d\) \(t = nk\) for some \(k\); considering \((x_n + y_n \sqrt{d})^k, y_n^2 \mid k x_{nk} y_n\), from which \(y_n \mid k\). Parts \(f\) and \(g\) follow by induction using lemma 32c. For part \(h\), using lemma 32a and Pell’s equation

\(y_{2n \pm m} = y_{n \pm (n \pm m)} \equiv x_{n \pm m} y_n \equiv \pm dy_n y_m \equiv \mp y_m \mod x_n\),

\(y_{4n \pm m} = y_{2n \pm (2n \pm m)} \equiv -y_{2n \pm m} \equiv \pm y_m\).
provided \( m < n \) when \( m \) is subtracted. In particular \( y_k \mod x_n \) is periodic with period \( 4n \). Let \( \sigma \) denote the sequence of residues of \( y_k, 0 < j < n \). Then the sequence for \( n < j < 2n \) is \( \sigma_R \), the reverse of \( \sigma \); the sequence for \( 2n < j < 3n \) is \( -\sigma_R \), the negatives mod \( x_n \) of \( \sigma \); and the sequence for \( 3n < j < 4n \) is \( -\sigma_R \). Also, \( \mod x_n, y_0 \equiv 0, y_2n \equiv 0, \) and \( y_{3n} \equiv y_n \). Suppose \( a \geq 3 \); then from Pell's equation \( y_n < x_n/2, \) and part \( h \) follows using part \( b \). In the remaining case \( a = 2, \) by lemma 32b \( y_n < y_{n+1}/3, \) and so \( y_m < x_n/3 \) for \( m < n \); also \( y_n \leq (2/3)x_n \). The only possible exception to the claim is \( y_{3n} \equiv y_n; \) then \( y_n = x_n/2 \) and \( n = 1, \) and the claim is true here also.

**Lemma 34.** The set \( \{(y,n,a) : y = y_n(a), n > 0, a > 1\} \) is Diophantine.

**Proof:** Consider the system

1. \( a > 1, n > 0 \)
2. \( x^2 - (a^2 - 1)y^2 = 1 \)
3. \( x_i^2 - (a^2 - 1)y_i^2 = 1 \)
4. \( x_2^2 - (b^2 - 1)y_2^2 = 1 \)
5. \( b \equiv 1 \mod 2y \)
6. \( y_2 \equiv n \mod 2y \)
7. \( b \equiv a \mod x_1 \)
8. \( y_2 \equiv y \mod x_1 \)
9. \( y_2 \mid y_1 \)
10. \( y \equiv n \)

Let \( (y,n,a) \) be such that there is a solution in the remaining variables; let \( (x,y) = (x_i(a), y_i(a)), (x_1, y_1) = (x_j(a), y_j(a)), (x_2, y_2) = (x_k(b), y_k(b)). \)

11. \( y_2 \equiv k \mod 2y \) by 5 and lemma 33f
12. \( n \equiv k \mod 2y \) by 6 and 11
13. \( y_i(a) \equiv y_k(b) \mod x_1 \) by 7 and lemma 33g
14. \( y_i(a) \equiv y_k(b) \mod x_1(a) \) by 8 and 13
15. \( k = \pm 1 \mod 2j \) by 14 and lemma 33h
16. \( y \mid j \) by 9 and lemma 33e
17. \( y \equiv 2y \mod 2y \) by 12, 15, and 16
18. \( n = i \) by 10, 17, and lemma 33b

Conversely if \( y = y_n(a) \) for \( a > 1 \) let \( x = x_n(a) \), so that 1, 2, and 10 are satisfied. Let \( j = 2ny, x_1 = x_j(a), \) and \( y_1 = y_j(a) \), so that 3 is satisfied; by lemma 33d and \( e \) 9 is satisfied. By lemma 33c \( x_1 \) is odd, so \( \gcd(2y, x_1) = 1 \). Choose \( b \) so that \( b \equiv 1 \mod 2y \) and \( b \equiv a \mod x_1 \); then 5 and 7 are satisfied. Let \( x_2 = x_n(b) \), and \( y_2 = y_n(b) \), satisfying 4; 6 is satisfied by 5 and lemma 33f. Finally 8 is satisfied by 7 and lemma 33g.

**Lemma 35.**

a. \( x_n(b) - (b - a)y_n(b) \equiv a^n \mod 2ba - a^2 - 1 \)

b. for \( n > 0, a > 0 \), if \( b > a^n \) then \( 2ba - a^2 - 1 > a^n \)

**Proof:** Part a follows by induction on \( n \) by lemma 32c. Indeed, using the induction hypothesis and lemma 32c, \( x_{n+1}(b) - (b - a)y_{n+1}(b) \equiv 2ba^n - a^{n-1} \mod 2ba - a^2 - 1; \) the right side is congruent to \( a^{n+1} \). For part b, for \( a \geq 1 a^2 + a - 1 > 0, \) whence \( a^n + 1 > a + 1 > a^2/(2a - 1), \) whence \( 2a(a^n + 1) > a^n + 1 + a^2. \) This proves part b for \( b = a^n + 1, \) and \( 2ba - a^2 - 1 \) is increasing in \( b. \)

**Lemma 36.** The set \( \{(z,n,a) : z = a^n, n > 1, a > 1\} \) is Diophantine.
Proof: Consider the system
1. $a > 1$, $n > 1$
2. $b = x_n(a) + 1$
3. $x_n(b) - (b - a)y_n(b) \equiv z \mod 2ba - a^2 - 1$
4. $z < 2ba - a^2 - 1$

Suppose for some $(z, n, a)$ the system has a solution. By 3 and lemma 35a, $z \equiv a^n \mod 2ba - a^2 - 1$. By 2 and lemma 33b, $b > a^n$, so by lemma 35b, $2ba - a^2 - 1 > a^n$. This, together with 4, shows $z = a^n$. Conversely if $z = a^n$, $b$ is determined by 2, and $b > a^n$, so 3 and 4 are satisfied by lemma 35.

In the following let $B(n, i)$ denote the binomial coefficient $n!/(i!(n - i)!)$.

**Lemma 37.** If $u > 2^n$ and $0 \leq k \leq n$ then $\sum_{i=k}^{n} B(n, i) = \lfloor (u+1)^n/u^k \rfloor$.

**Proof:** By the binomial theorem

$$\frac{(u+1)^n}{u^k} = \sum_{i=0}^{k-1} B(n, i) u^{i-k} + \sum_{i=k}^{n} B(n, i) u^{i-k}.$$  

Under the hypothesis the first sum is less than 1.

**Lemma 38.** The set $\{(z, n, k) : z = B(n, k)\}$ is Diophantine.

**Proof:** Consider the system
1. $u = 2^n + 1$
2. $z \equiv \lfloor (u+1)^n/u^k \rfloor \mod u$
3. $z < u$

Suppose for some $(z, n, k)$ the system has a solution. By 1 and lemma 37 $\lfloor (u+1)^n/u^k \rfloor = \sum_{i=k}^{n} B(n, i)$; since all terms of the sum other than the first are divisible by $u$, by 2 $z \equiv B(n, k) \mod u$. But $B(n, k) \leq 2^n < u$, so by 3 $z = B(n, k)$. Conversely if $z = B(n, k)$ the conditions are satisfied by choosing $u = 2^n + 1$.

Say that two integers $x$ and $y$ are disjoint if in their binary notations no 1 occurs in a common position. That this predicate is Diophantine follows from the next lemma.

**Lemma 39.** Two integers $x$ and $y$ are disjoint iff $B(x+y, x)$ is odd.

**Proof:** Suppose $x$ has $r$ 1’s, $y$ has $s$ 1’s and $x+y$ has $t$ 1’s. One verifies that $t \leq r+s$, with equality holding iff $x$ and $y$ are disjoint. We claim that the highest power of 2 dividing $x!$ is $x-r$, and the lemma follows. To prove the claim, suppose $x = 2^{e_1} \cdots 2^{e_r}$ where $e_1 < \cdots < e_r$; let $x_j$ be the sum of the first $j$ terms, and consider the intervals $I_1 = [1, x_1]$ and $I_j = [x_{j-1}+1, x_j]$ for $j > 1$. Note that in $I_j$ there are $2^k$ numbers, one of which is $2^k$, where $k = e_j$. For $1 \leq i \leq k$ there are $2^{k-i}$ numbers divisible by $2^i$, whence the number where $2^i$ is the highest power is 1 if $i = k$ and $2^{k-i-1}$ if $i < k$. Thus the highest power of 2 dividing the product of the numbers in the interval equals $k + \sum_{i=1}^{k-1} i2^{k-i-1}$, which may be seen by induction to equal $2^k - 1$, that is, $2^{e_j} - 1$. The claim follows by summing over $j$.

**Theorem 40.** Every recursively enumerable set is Diophantine.

**Proof:** Let $S$ be recursively enumerable. Let $M$ be a Turing machine which halts on input $x$ iff $x \in S$. $M$ is assumed to have tape symbols \{0, 1\} with 0 functioning as blank; the tape is one-way infinite to the left and the input is at the right, so with the tape contents coded in binary the code of the initial tape is simply the input. It is also assumed that the initial state is not reentered. The conditions
- $T$ is a power of 2
- $T \cdot A + T/2 - A$ is a power of 2
- $B$ is a power of 2
- $A < B < 2A$

ensure that if $T = 2^t$ then $B = 2^{kt}$ for some $k$; also $A$ marks the left end of each block of length $t$. In the following conditions $C$ records the tape contents of the steps in consecutive blocks from the right, and $I_\sigma$ for each instruction $\sigma$ the fact that instruction $\sigma$ was executed at a given square.

- $C \equiv x \mod T$.
- The $I_\sigma$ are disjoint from $A$.
- For some $\sigma$ executing in the initial state, $I_\sigma \equiv 1 \mod T$ and $I_\sigma \equiv 0 \mod T$ for $\tau \neq \sigma$.
- For each state $q$ let $L = \sum_\sigma I_\sigma$ where $\sigma$ ranges over instructions executing in state $q$. Let $E_L$ ($E_R$) be $\sum_\sigma I_\sigma$ where $\sigma$ ranges over instructions which move left (right) into state $q$. Let $E_0$ be 1 if $q$ is the initial state, else 0. The condition is $L \equiv 2TE_L + (T/2)E_R + E_0$.
- $\sum_\sigma I_\sigma$ and $B - 1 - C (C)$ are disjoint, where $\sigma$ ranges over instructions which execute reading 0 (1).
- For $i \in \{0, 1\}$ let $W_i = \sum_\sigma I_\sigma$ where $\sigma$ reads $1-i$ and write $i$. The condition is $T(C+W_1-W_0) \equiv C-x \mod B$.
- For some $\sigma$ entering a halting state, $I_\sigma \neq 0$.

These conditions can be satisfied iff $M$ halts on $x$.

Using Gödel's pairing function, every recursively enumerable predicate is Diophantine. It also follows that it is undecidable if a polynomial with coefficients in $\mathbb{Z}$ has a solution in $\mathbb{Z}$ (replace each variable by a sum of four squares and use Lagrange's four square theorem). It is open if it is decidable if such a polynomial has a rational solution, even for two variables. Finally, a predicate is Diophantine iff it is $\Sigma_1$ in the language of arithmetic (move negative terms of the polynomial to the other side of the equation).

Say that a predicate $R(x)$ is $p$-Diophantine if there is a multivariate polynomial $p(x, y)$ with coefficients in $\mathbb{Z}$, and a polynomial $b$, such that $R(x) \iff \exists y (\ell(y) \leq b(\ell(x)) \land p(x, y) = 0)$. Clearly every $p$-Diophantine predicate is in $\mathsf{NP}$; the opposite inclusion is open.

**Corollary 41.** If disjointness is $p$-Diophantine then every predicate in $\mathsf{NP}$ is $p$-Diophantine.

**Proof:** The auxiliary variables in the proof of the theorem may have polynomially bounded length if the Turing machine runs in polynomial time; the witness may also, by definition. Finally, the condition that $T$ be a power of 2 may be expressed as “$T$ is positive and disjoint from $T-1$”.

**9. Word problem for groups.** Given a set of generators $A$ let $A^{-1}$ denote the disjoint set of letters \( \{a^{-1} : a \in A\} \). A finite presentation on $A$ of a group is a finite presentation of a semigroup on $A \cup A^{-1}$, where the relations include $aa^{-1} = 1$ and $a^{-1}a = 1$. The semigroup defined by such a presentation is a group; the inverse of (the class of) $l_1 \cdots l_r$ is $l_r^{-1} \cdots l_1^{-1}$, where $l^{-1}1^{-1} = 1$.

In particular the group with just the relations $aa^{-1} = 1$ and $a^{-1}a = 1$ is the free group on $A$ (this is defined more abstractly in the next chapter). Let this be $F_A$; given any other group $G$ generated by $A$ there is a unique group homomorphism from $F_A$ to $G$ which maps each (class of a) generator to itself.

A string over $A \cup A^{-1}$ is called reduced if it has no adjacent $l$ and $l^{-1}$. Two such may be “multiplied” by concatenating and canceling. This operation is associative. This is clear if in the product $\alpha \beta \gamma$ the cancellation does not overlap. If it does then using obvious notation $\alpha = \alpha_1 \beta_1^{-1} \beta_2^{-1}$, $\beta = \beta_1 \beta_2 \beta_3$, and $\gamma = \beta_3^{-1} \beta_2^{-1} \gamma_1$; the claim follows in this case also.

The reduced strings with this multiplication operation form a group; let this be $R_A$. Since this is generated by $A$ there is a group homomorphism from $F_A$ to $R_A$. On the other hand there is a group homomorphism from $R_A$ to $G_A$ taking each string to its class. These maps are inverse to each other; in particular each class contains a unique reduced string.
Theorem 42. The problem of whether, for a group with two generators and a recursively enumerable set of relations two words are equivalent, is many-one complete.

Proof: The equivalences are recursively enumerable as in theorem 6. Let $s_i$ denote $a^iba^{-i}$. If $S$ is a recursively enumerable set consider the relations $s_i = 1$ for $i \in S$; we claim that $s_i = 1$ in the resulting group iff $i \in S$, proving the theorem. Let $F_1$ be the free group. The $s_i$ freely generate a subgroup $F_2$ (that is, the obvious map is an embedding) (exercise 4a). This in turn contains the subgroup $F_3$ generated by \{s_i : i \in S\}; let $N$ be the normal subgroup of $F_1$ generated by $F_3$. If $w \in F_2$ let $w'$ be obtained by dropping the $s_i$ or $s_i^{-1}$ where $i \in S$. Each $w \in F_2$ has a unique representation of the form $nw'$ where $n \in N$ (exercise 4b). This proves the claim.

The problem of whether two words are equivalent for a finitely presented group is many-one complete. The proof is lengthy and will not be given here. One way to proceed is to use theorem 42 and Higman’s theorem. Higman’s theorem states the following. Suppose $H$ is a recursively enumerable subgroup of $G$ and $G$ is recursively enumerable normal subgroup $N$ of $H$. Then there is a finitely presented group $G_1$ and a recursive embedding of $G_1$ in $G_2$. That is, if $G_2$ is the quotient of the finitely generated free group $F_2$ by the normal subgroup $N_2$ (specified by a finite set of relations), then $f : F_1 \rightarrow F_2$ and for $x \in H_1$, $x \in N_1$ iff $f(x) \in N_2$. When $H_1 = F_1$ such an $f$ provides a many-one reduction of the word problem of $G_1$ to that of $G_2$. A proof of Higman’s theorem may be found in [LynSch] or [Manin].

10. Details for goto step simulation. The goto program state $\langle i, x_1, \ldots, x_n \rangle$ is represented by the Turing machine state “$q x_1, \cdots, x_n$”, where in the latter $q \in Q$ and $x_j$ is a signed 2-adic integer. In Turing machine programs it is often convenient to consider the head state as a tuple, with different components recording different information; for example $i$ can be recorded as the first component. Similarly the tape symbols may be considered as tuples; the components may be considered as “tracks” on the tape.

One main subroutine for the simulation is $\pm x_j \pm x_k$ where $j < k$. This can be entered for various reasons, recording the reason in the head state. A “pseudo-code” description of the simulation is as follows.

1. Mark the end of the new string at the left of the input, with $O$.
2. Move right to the beginning of $x_j$; save the sign; move right to the end; mark it with $J$.
3. Move right to the beginning of $x_k$; save the sign; move right to the end; mark it with $K$.
4. Move left to $O$; set the carry to $0$.
5. Repeat 6-11 until done:
6. Move right to $J$; if there is a digit to the left record it in the head state and move $J$ left, otherwise record $0$.
7. Move right to $K$ and proceed as in step 6.
8. Move left to the end (blank).
9. If both digits are $0$, write the carry if any; write the - sign if any; and exit the loop.
10. Write a digit and update the carry.
11. Move right to $O$.

Another subroutine replaces $x_j$ with the new value $w$ at the left of the tape. One strategy is to move $x_{j+1}, \ldots, x_n$ left or right as necessary, and copy $w$; blanks are written at the left and right as necessary to finish.

1. Move left and write $A$ above the left digit of $w$.
2. Move right and write $B$ above the left digit of $x_j$.
3. Move back and forth, moving $A$ and $B$ right, until $A$ reaches $O$.
4. Write $C$ to the right of $x_j$, and determine if $C$ is left or right of $B$.
5. If \( C \) is left (right) of \( B \), shift the string from \( C \) to the right end right (left) 1 until \( C \) reaches \( B \) (shifting right (left) proceeds right-to-left (left-to-right)).

6. Copy \( w \) from left to right to the space made for it, using a marker on the left.

Using these two subroutines and appropriate “clean up” etc., each of the basic steps of a goto program can be simulated. The reader may supply the lengthy detailed programs if he or she desires. It is easy to see that the simulation runs in time \( O(\ell(x)^2) \).

11. Details for Turing machine step simulation. The Turing machine state \( \alpha q s \beta \) is represented by the contents of 4 variables, \( \alpha, q, s, \) and \( \beta \). Clearly a step can be simulated, given subroutines for appending a digit or trimming a digit, from either the left or the right, of an \( n \)-adic value.

The following pseudo-code program computes the leading digit (ld) and the tail.

```plaintext
if x \neq 0 \{ld=0; tail=0;\} else
op=0; o=1; on=n\cdot1+1;
while on \leq x \{op=o; o=on; on=n\cdot o+1;\}
ldo=ld; ldon=ldo+o-op; if ldon>x ld=1; else {ldo=ld; ld=ldo+o-op; if ldon>x ld=n-1; else {ldo=ld; ld=n; } \cdots\}
tail=x-(ldon-op);}
```

With \( n \) fixed this program runs in time \( O(\ell(x)) \). It can be used as a subroutine when computing the trailing digit and the head. This is a loop which peels of the leading digit, and appends it to the head so far. Thus, a step of a Turing machine can be simulated in \( O(\ell(x)^2) \).

12. Details for a universal Turing machine. Turing machines can be readily coded over a fixed alphabet. For example let the tape symbols be 1, \ldots, \( P \) with 1 blank, and the head states \( P+1, \ldots P+Q \). Code these as dyadic strings. Code the Turing machine as \( P;Q; r_1; \cdots; r_k \); where the rule \( r_1 \) is \( q, s, q', s', d \) where \( d \) is 1 (2) for left (right).

The state of the object machine is represented by the string \( p_1, \ldots, p_i, q, p_{i+1}, \ldots, p_l \). This is appended to the code of the machine to give the state of the universal machine. By Church’s thesis there is a Turing machine which updates the latter to simulate a step of the object machine. A more specific description can be given using methods as in section 10, i.e., tape markers and string operation subroutines.

It does not matter what the universal machine does if its state is not well-formed; there need only be an index for each partial function. In outline, the rules are searched from left to right looking for one starting with \( p, q \). The object machine state is updated in two phases, first shifting portions as necessary, and then copying \( q' \) and \( s' \).

13. Proof sketch for theorem 13. To prove theorem 13 we first prove the analog of lemma 11, that \( \mathcal{L} \) is closed under bounded recursion on \( n \)-adic notation for any \( n \geq 2 \). To begin with, some basic functions are defined by recursion on dyadic notation; let \( i = 1, 2, \ldots \).

- Cond(0, y, z) = y, Cond(x_i, y, z) = z.
- Suc(0) = 1, Suc(x1) = x2, Suc(x2) = (Suc(x))1.
- Pred(0) = 0, Pred(x1) = Cond(x, 0, (Pred(x))2), Pred(x2) = x1.
- Cond_n(x, y_0, \ldots y_{n-1}) is as before.
- Rdig_n(0) = 0, Rdig_n(x) = Cond_{n+1}(Rdig_n(x), C_0, \ldots C_n)
  where \( C_{ji} \) is the right \( n \)-adic digit of \( 2j+i \).
- Trdig_n(0) = 0, Trdig_n(x) = Cond_{n+1}(Trdig_n(x), T_0, \ldots T_n)
  where \( T_{ji} \) is \((Trdig_n(x))1 \) with Pred or some number of Suc’s applied.

124
- \( \text{Trdigs}_n(0, y) = y \), \( \text{Trdigs}_n(x, y) = \text{Trdigs}_n(\text{Trdigs}_n(x, y)) \).

- \( \text{Len}_n(0, x) = 0 \),
  \( \text{Len}_n(wi, x) = \text{Cond}_n(\text{Trdigs}_n(w, x), \text{Len}_n(w, x), (\text{Len}_n(w, x))1) \).

- \( \text{Len}_n(x) = \text{Len}_n(T, x) \),
  where \( T \) is \( x \) concatenated with itself sufficiently many times.

Observe that in \( \text{Trdigs} \), \( \text{Trdigs} \) is iterated \( n \) times where \( n \) is the length of the dyadic notation for \( x \); and \( \text{Len} \) is a string of 1’s in dyadic whose length equals that of the \( n \)-adic notation for \( x \). The rest of the proof of the analog of lemma 11 proceeds essentially as before; details are left to the reader.

That the step function of a Turing machine is in \( \mathcal{L} \) follows as before. That a function computable in polynomial time is follows by standard arguments. Namely, the input value can be converted to the initial ID by a function in \( \mathcal{L} \); the step function can be iterated sufficiently often using functions in \( \mathcal{L} \) to bound above the running time and the ID; and the output can be extracted from the halting ID.

That every function in \( \mathcal{L} \) is computable in polynomial time is an exercise in Turing machine programming; we give some remarks. Supposing we have machines for \( t_1, \ldots, t_k \) and \( f \), the machine for \( f(t_1, \ldots, t_k) \) first copies the inputs to a track where they are saved. For each \( i \) in turn, the inputs are copied to a second track; the machine for \( t_i \) is run; and the output is copied to its proper position on a third track. The machine for \( f \) is run on the third track. Finally the output is copied to obey the output convention. Each step is polynomial in the length of the input, using the closure of the polynomials under composition.

For limited recursion on notation, 5 tracks can be used; the first has \( x, y \); the second a string of 1’s keeping track of the stage of the recursion; the third the value of the function at the current stage of the recursion; and the 4th for computing \( g \) or \( h \); and the fifth for computing \( b \) (additional tracks can be used for markers). The entire computation runs in polynomial time because the length of the value is kept bounded by a polynomial.

**Exercises.**

1. Supply details for the proof of theorem 12. The step function may be defined using concatenation and substring extraction functions; to stay inside \( \mathcal{E}_1 \) concatenation must be bounded. The required functions are readily defined using recursion on notation; for example (abbreviating \( \text{App}^i(x) \) as \( x^i \)) concatenation may be defined as \( \text{Concat}(x, 0) = x \), \( \text{Concat}(x, yi) = \text{Concat}(x, y)i \). Useful functions include the leftmost digit in a set \( S \) of digits; and the substrings to the left and right.

2. Suppose \( T \) is a consistent theory, and for any consistent extension \( T' \supseteq T \) in the same language, \( T' \) is undecidable. Show that \( T \) is essentially undecidable. Hint: if \( T' \supseteq T \) is in an expanded language, consider \( T'' \), the formulas of \( T' \) in the base language.


4. Prove the claims a and b made in the proof of theorem 42. Hint: for part a, in a string of \( s_i \) or \( s_i^{-1} \) call the occurrences of \( b \) or \( b^{-1} \) “significant factors” (these are used in the Schreier proof of the Nelson-Schreier theorem that any subgroup of a free group is a free group; see [Hall]). Observe that significant factors cancel iff the \( s_i \) or \( s_i^{-1} \) do. For part b let \( \hat{F}_2 \) be the subgroup of \( F_1/N \) consisting of the cosets of elements of \( F_2 \). Define a multiplication on the representations and show that the result is a group isomorphic to \( \hat{F}_2 \).
13. Category theory.

1. Basic definitions. Category theory provides an axiomatic setting for diagram properties. There is taken to be
- a collection $\text{Obj}$ of objects $a, b, c, \ldots$;
- a collection $\text{Ar}$ of arrows (or morphisms) $f, g, h, \ldots$;
- functions $\text{Dom}, \text{Codom} : \text{Ar} \mapsto \text{Obj}$;
- a partial function $\circ \subseteq \text{Ar} \times \text{Ar} \times \text{Ar}$; and
- a function $\iota : \text{Obj} \mapsto \text{Ar}$ (the value at $a$ will be denoted $\iota_a$).

These must satisfy

- $f \circ g$ is defined iff $\text{Codom}(g) = \text{Dom}(f)$, and in this case $\text{Dom}(f \circ g) = \text{Dom}(g)$ and $\text{Codom}(f \circ g) = \text{Codom}(f)$;
- $f \circ (g \circ h) = (f \circ g) \circ h$ whenever either side is defined; and
- for each $a \in \text{Obj}$, $\text{Dom}(\iota_a) = \text{Codom}(\iota_a) = a$, and $g \circ \iota_a = g$ and $\iota_a \circ g = g$ whenever the left side is defined.

The operation $\circ$ is called composition; when $\text{Dom}(f) = \text{Codom}(g)$ $f$ and $g$ are said to be composable. The operation symbol $\circ$ may be omitted when no confusion arises.

Write $f : a \mapsto b$ if $\text{Dom}(f) = a$ and $\text{Codom}(f) = b$. Let $\text{Hom}(a, b)$ denote $\{ f : a \mapsto b \}$. If $f : a \mapsto b$ a left inverse is a function $g : b \mapsto a$ such that $g \circ f = \iota_a$; a right inverse is a function $g : b \mapsto a$ such that $f \circ g = \iota_b$. If $g$ is both a left and right inverse it is called a two-sided inverse; an arrow with a two sided inverse is called an isomorphism, and an automorphism if it is an isomorphism from $a$ to $a$.

In category theory there is a metamathematical issue which does not arise for most systems of axioms. Namely, Obj and Ar should be allowed to be collections of higher type than a set. In particular they are often proper classes; this can be handled within ordinary set theory. Even higher types are sometimes referred to, for example, the “category of categories”, whose arrows (functors) will be described below. Handling these requires additional considerations. One method is to postulate a “universe” (a set which is closed under all the operations of set theory) and consider only its members as the sets. Other methods, such as adding higher types to set theory, have been considered. In any case, larger collections behave enough like sets that, for example, theorems of universal algebra continue to hold. Issues such as these have led a school of category theorists to explore the use of categories as an alternative approach to the foundations.

If Obj and Ar are proper classes, and the collections $\text{Hom}(a, b)$ are sets for every $a, b$, we call the category a standard category. For these, the metamathematical issues are more tractable than in general, although even here there are some complications. Further remarks will be made at the end of section 5.

A fundamental example of a category is Set, the category whose objects are the sets; this is a standard category. The arrows $f : a \mapsto b$ are often taken to be the functions from $a$ to $b$; however this does not yield a category as we have defined it since there is no codomain function. We adopt the solution of considering an arrow in Set from $a$ to $b$ to be a pair $\langle f, b \rangle$ where $f$ is a function with domain $a$ and range contained in $b$.

This is rarely a complication; indeed the set-theoretic representation of a category is usually a technicality.

Another solution is to change the definition of a category. It is useful to have the axioms in a form where universal algebra applies. Also, they should be “self-dual” (see below). Dom and Codom can be replaced by the predicate “$f : a \mapsto b$”. The composition predicate has 5 arguments; the axioms are left to the reader. We prefer the first solution because the axioms are simpler and the problem rarely requires consideration.

Another solution is to define a category by giving the Hom sets. Universal algebra is less readily applied with these axioms. Authors who do use them often require the Hom sets to be disjoint; in this case functions in Set must be tagged with the codomain also.

Unless otherwise specified, in the remainder of this text a category will be standard. A standard
category is called concrete if the objects are sets and Hom(a, b) consists of functions from a to b (tagged with
the codomain); and composition is composition of functions. Not all common categories are concrete; for
example there is an important category in topology where the arrows are equivalence classes of functions.
Several important concrete categories are as follows.

Set, with the empty set and all arrows from or to it deleted. ∅ has unusual properties, and it is
sometimes convenient to omit it. In particular, for any set S the only function from S to ∅ is ∅. The only
function into ∅ is ∅. ∅ is always an injection, and is a surjection if S = ∅. ∅ has a left inverse if S = ∅.

Struct_L, where L is a first order language. The objects are the structures of type corresponding to L,
and the arrows are the homomorphisms (strictly speaking tagged with the codomain). ∅ is included among
the objects if L contains no constant symbols. As mentioned in chapter 2, this results in simplifications in
some contexts. We denote the category where ∅ is omitted as Struct_L^T.

Mdl_T, for T a first order theory. The objects are the models of T and the arrows the homomorphisms
between structures. ∅ is not included among the objects. It may be allowed if L contains no constant; we
denote this category as Mdl_T^L. Since ∃x(x = x) is valid ∅ is not strictly speaking a model of T. If T is open
and OFOL rather than FOL is used to define validity, ∅ may be taken as a model. In some contexts this is
more suitable to the development; for example ∅ is often considered to be a partial order.

Mon, the monoids, with the monoid homomorphisms; this is Mdl_T where T is the monoid axioms.
Grp, the groups, with the group homomorphisms.
Rng, the rings, with the ring homomorphisms; and CRng, the commutative rings.
Mod_R where R is a ring, the left R-modules with the R-module homomorphisms. The category of
commutative groups is identical to the category of L-modules; it is denoted Ab. To view this category as a
case of Mdl_T, L must be allowed to be infinite (having a unary function for each r ∈ R); T is infinite also.
A standard category where Obj and Ar are sets is called small. Some examples of small categories are
as follows.

- A discrete category is one where all arrows are identity arrows. The sets are in obvious one-to-one
correspondence with the small discrete categories.
- A preorder may be viewed as a category where there is a single arrow from a to b if a ≤ b, and no arrows
otherwise. Indeed, it is easily seen that the small categories where each Hom set contains at most one
arrow are the same as the preorders.
- A category with one object (and a set of arrows) is essentially the same as a monoid. A category with
one object where the arrows all isomorphisms is essentially the same as a group.
- For a commutative ring R, the R-modules R^n for n ∈ N, with the R-linear maps, form a category. The
arrows are the matrices over R, and we use Mat_R to denote this category.

Basic definitions concerning categories can be arrived at by first considering general definitions of uni-
versal algebra, as applied to small categories. These are multisorted structures, with partial functions. Since
we are only interested in this approach as a guideline, a detailed treatment will not be given. For some
comments, partial functions can be treated as relations, or given a separate treatment. We will presume the
former approach, which is more economical; but treating a partial function as a relation requires an axiom
stating that it is a partial function. The axioms for categories are thus not universally quantified. The
generalization of the basic theorems of universal algebra to homomorphisms between multisorted structures
is left to the reader.

We mention that another important example of a category of multisorted structures is provided by the
modules. An object consists of a ring R and an R-module M. An arrow from (R, M) to (S, N) is a pair
⟨ρ, f⟩ where ρ is a ring homomorphism, f a commutative group homomorphism, and f(rm) = ρ(r)f(m).
Universal algebra (considering only small categories to begin with) yields the definition of a morphism
between categories, which is called a functor. A functor from C to D is a pair of maps F : Obj_C → Obj_D
and $F : \text{Ar}_C \rightarrow \text{Ar}_D$ (the same symbol is used for both maps), such that the following hold.

- $\text{Dom}(F(f)) = F(\text{Dom}(f))$ and $\text{Codom}(F(f)) = F(\text{Codom}(f))$;
- $F(f \circ g) = F(f) \circ F(g)$ whenever $f \circ g$ is defined; and
- $F(\iota_a) = \iota_{F(a)}$.

It is sometimes helpful to think of $F$ as giving a “picture” of $C$ in $D$.

A simple example of a functor is a “forgetful” functor, say mapping a ring or module to its additive group, or an object in any concrete category to its underlying set. For a slightly less trivial example, a group may be mapped to its left regular representation, considered as an algebra over some field. In these examples, as in many, it is clear what the image of an arrow should be.

By the usual abuse of notation we may write $C$ for $\text{Obj}_C$, when no confusion arises. In particular $x \in C$ is an abbreviation for $x \in \text{Obj}_C$; and the notation $F : C \rightarrow D$ may be used for a functor from $C$ to $D$. The notation $\text{Hom}_C(a,b)$ is used to specify that the category is $C$. $D$ is called a subcategory of $C$ if $D \subseteq C$; for $a,b \in D$, $\text{Hom}_D(a,b) \subseteq \text{Hom}_C(a,b)$; and composition in $D$ is the restriction of composition in $C$. That the identity arrows of $D$ are those of $C$ follows. For example a concrete category is just a subcategory of $\text{Set}$; and $\text{Mdl}_T$ is a subcategory of $\text{Struct}_L$ where $L$ is the language of $T$.

**Lemma 1.**

a. A subcollection of a category determines a substructure iff the inclusion map is a functor.

b. The composition of functors is a functor.

c. The inverse to a bijective functor is a functor.

d. The inverse image of a subcategory under a functor is a subcategory.

**Proof:** The first three parts follow by universal algebra, and the last is readily verified directly.

Note that the small categories form a standard category, with the functors as arrows, which we denote by $\text{Cat}$. The empty set is considered to be a small category.

For another example of a functor, the map $R \mapsto \text{Mat}_R$ is a functor from $\text{CRng}$ to $\text{Cat}$. The image of a ring homomorphism $h : R \rightarrow S$ is a functor, which takes the object $R^n$ to $S^n$, and an arrow (matrix) $A$ to the matrix with components $h(A_{ij})$.

A functor $F : C \rightarrow D$ determines maps from $\text{Hom}_C(a,b)$ to $\text{Hom}_D(a,b)$ for each $a,b$. If these maps are injective the functor is called faithful, and if they are surjective it is called full. Note that a functor can be faithful and full without being bijective. A subcategory $D$ of $C$ is called a full subcategory if the inclusion functor is full, that is, if for $a,b \in D$ $\text{Hom}_D(a,b) = \text{Hom}_C(a,b)$. For example, $\text{Mdl}_T$ is a full subcategory of $\text{Struct}_L$, and $\text{Grp}$ is a full subcategory of $\text{Mon}$. For another example, the preorders with the order preserving maps form a full subcategory of $\text{Cat}$ (whether or not the empty set is allowed).

The product of categories can be defined, and has several uses. For example, a category of multisorted structures can be considered as a subcategory of $\text{Set}^k$, generalizing the notion of a concrete category. It is worth mentioning that there is no metamathematical problem arising from considering the Cartesian product of proper classes; it is simply the class of tuples.

Multisorted structures can be treated as ordinary structures by introducing predicates for the sorts, requiring the domain to be the disjoint union, etc. However this introduces additional axioms, which are not universal, and in this case we prefer to generalize the notion of a structure. The product of categories is also very useful for defining functors of several arguments.

The product of categories $C_1, \ldots, C_k$ has objects $\langle a_1, \ldots, a_k \rangle$ where $a_i$ is an object of $C_i$, and arrows $\langle f_1, \ldots, f_k \rangle$ where $f_i$ is an arrow of $C_i$. The definitions of $\text{Dom}$, $\text{Codom}$, $\circ$, and $\iota$ follow by universal algebra, and it is readily verified that the product is a category. Note that $\text{Hom}(\langle a_1, \ldots, a_k \rangle, \langle b_1, \ldots, b_k \rangle)$ equals $\text{Hom}(a_1, b_1) \times \cdots \times \text{Hom}(a_1, b_1)$; in particular the product of standard categories is standard.
It is simplest to consider functors of several arguments as ordinary functors on a product category. Given a functor \( F : C \times D \to E \), for each \( c \in C \) there is a functor from \( D \) to \( E \) mapping \( d \in D \) to \( F(c,d) \); this functor may be denoted \( F(c,-) \). It maps \( g : d \to d' \) to \( F(g,c) \). Similarly for each \( d \in D \) there is a functor \( F(-,d) \) from \( C \) to \( E \), which maps \( f : c \to c' \) to \( F(f,d) \). Given functors \( F(c,-) : D \to E \) for each \( c \in C \), and \( F(-,d) : C \to E \) for each \( d \in D \), it is readily verified for these to determine a binary functor \( F(-,-) \) it is necessary and sufficient that the following diagram always commute.

\[
\begin{array}{ccc}
F(c,d) & \xrightarrow{F(f,d)} & F(c,d') \\
\downarrow & & \downarrow \\
F(f\circ d) & \xrightarrow{F(f\circ d')} & F(f\circ d')
\end{array}
\]

Given a category \( C \) its opposite category \( C^{op} \) is defined by “reversing the arrows”, that is, interchanging \( \text{Dom} \) and \( \text{Codom} \), and the arguments of \( \circ \). One use of \( C^{op} \) is defining functors. A functor \( F \) from \( C^{op} \) to \( D \) induces a map from \( C \) to \( D \), which maps an arrow \( f : a \to b \) of \( C \) to an arrow \( F(f) : F(b) \to F(a) \) of \( D \). Such a map is called a contravariant functor; and an ordinary functor is called covariant by comparison. For a contravariant functor, \( F(fg) = F(g)F(f) \) for composable arrows.

Products and opposites are often used together. For example for any standard category \( C \), the map taking \( a, b \) to \( \text{Hom}(a,b) \) is a functor from \( C^{op} \times C \) to \( \text{Set} \). A pair of arrows \( f : a' \to a, g : b \to b' \) is mapped to the function from \( \text{Hom}(a,b) \) to \( \text{Hom}(a',b') \) which takes \( h \) to \( ghf \). The covariant functor \( \text{Hom}(a,-) : C \to \text{Set} \) maps \( g : b \to b' \) to the function which maps \( h \in \text{Hom}(a,b) \) to \( gh \in \text{Hom}(a,b') \). The contravariant functor \( \text{Hom}(-,b) : C^{op} \to \text{Set} \) maps \( f : a' \to a \) to the function which maps \( h \in \text{Hom}(a,b) \) to \( hf \in \text{Hom}(a',b) \).

Another use of \( C^{op} \) is dualizing definitions and theorems of category theory. Given some definition, the dual is obtained by considering what is defined in the opposite (also called dual) category. For example, an initial object is an object \( a \) such that there is a unique arrow \( f : a \to b \) for any object \( B \). An initial object in the opposite category is called a terminal object; that is, \( a \) is a terminal object if there is a unique arrow \( f : b \to a \) for any object \( b \). Any two initial objects \( a, b \) are isomorphic by a unique isomorphism, since if \( f : a \to b \) and \( g : b \to a \) are the arrows then \( fg = \iota_b \) since this is the only map from \( b \) to \( b \), and similarly \( gf = \iota_a \). It follows immediately that any two terminal objects are isomorphic by a unique isomorphism.

An object which is both initial and terminal is called a zero object. If \( C \) has a zero object \( 0 \) the composite of the unique arrows \( a \to 0 \to b \) is called the zero arrow from \( a \) to \( b \). In a category with a zero object \( 0 \) is often used to denote the zero object or any zero arrow.

Some examples of initial and terminal objects follow; we leave the verification as an exercise.

- In \( \text{Set} \), the empty set is initial and any set containing one element is terminal.
- In \( \text{Struct}_L \), provided \( L \) contains a constant symbol, the Herbrand structure where every relation other than equality is made false everywhere is initial. A structure with a one element domain and the relations made true everywhere is terminal.
- In \( \text{Mon} \) the trivial monoid \( \{1\} \) is a zero object; this is also true in \( \text{Grp} \).
- In \( \text{Mod}_R \) the zero module \( \{0\} \) is a zero object.
- In \( \text{Rng} \) or \( \text{CRng} \) \( Z \) is initial and the trivial ring \( \{0\} \) is terminal.
- If \( L \) contains no constants then \( \emptyset \) is the initial object in \( \text{Struct}_L \).

2. Natural transformations. If \( F, G : C \to D \) are two functors, a natural transformation from \( F \) to
$G$ is defined to be a map $\alpha : \text{Obj}_C \mapsto \text{Ar}_D$ such that for any $f : c \mapsto c'$ in $\text{Ar}_C$ the diagram

\[
\begin{array}{ccc}
F(c) & \xrightarrow{\alpha(c)} & G(c) \\
F(f) & \downarrow & G(f) \\
F(c') & \xrightarrow{\alpha(c')} & G(c')
\end{array}
\]

(1)

commutes. Thus, there is a system of maps between the pairs of objects representing an object of $C$; further this system is “natural”, in the sense that it respects the representations of the arrows of $C$.

A good example of a natural transformation in algebra is the map from an $R$-module $M$ to its double dual $M^{**}$, which recall maps $x$ to the function $\psi_x$ where $\psi_x(f) = f(x)$. The map $M \mapsto M^{**}$ is just the contravariant functor $\text{Hom}(-, R)$; the map $M \mapsto M^{**}$ is a covariant functor from $\text{Mod}_R$ to $\text{Mod}_R$, and we let this be $G$. $F$ is the identity functor, and $\alpha(M)$ is the map $x \mapsto \psi(x)$. That the diagram (1) is commutative follows if for $x \in M$ $f^{**}(\psi_x) = \psi_{f(x)}$. But $f^{**}(\psi_x) = \psi_x \circ f^*$, and $\psi_x(f^*(h)) = \psi_x(h \circ f) = h(f(x)) = \psi_{f(x)}(h)$.

The maps $\alpha(a) : F(a) \mapsto G(a)$ are called the components of the natural transformation $\alpha$. A useful notational device allows writing $\alpha_a$ for $\alpha(a)$. If the components are all isomorphisms $\alpha$ is called a natural equivalence. The above example is such, in the subcategory of $\text{Mod}_R$ of finite dimensional free $R$-modules.

Given functors $F, G : B \times C \mapsto D$, and a natural transformation $\alpha$ from $F$ to $G$, with components $\alpha(b, c)$, there are natural transformations $\alpha(b, -)$ from $F(b, -)$ to $G(b, -)$ and $\alpha(-, c)$ from $F(-, c)$ to $G(-, c)$. That these are natural follows because diagram (1) is certainly commutative for the maps $\langle f, \iota_c \rangle : \langle b, c \rangle \mapsto \langle b', c \rangle$ and $\langle \iota_b, g \rangle : \langle b, c \rangle \mapsto \langle b, c' \rangle$. Conversely if these special cases hold then commutativity of (1) holds in general, as is easily seen. Thus, to show $\alpha$ is “natural in $b, c$”, it suffices to show that it is natural in $b$ for each $c$, and natural in $c$ for each $b$. This fact generalizes to functors of any number of arguments.

For $C$ a standard category and $J$ a small category we define the “functor category” $C^J$ (as usual the definition can be given for larger categories, with appropriate care). The objects of $C^J$ are the functors $F : J \mapsto C$, which may be taken as sets. The arrows $\alpha : F \mapsto G$ are the natural transformations. These may be taken as comprising a set, since there is a set of maps from $F(i)$ to $G(i)$ for each $i \in J$. The commutative diagram

\[
\begin{array}{ccc}
F(i) & \xrightarrow{\alpha(i)} & G(i) \\
\downarrow F(f) & \downarrow G(f) & \downarrow H(f) \\
F(j) & \xrightarrow{\alpha(j)} & G(j) \\
\downarrow H(j)
\end{array}
\]

shows that composition may be defined componentwise, i.e., $(\beta \circ \alpha)(i) = \beta(i) \circ \alpha(i)$. The components of the identity natural transformation are the identity maps $1_F(i)$. Thus, $C^J$ is a standard category.

Note that if $J$ is discrete an object of $C^J$ is essentially an indexed family $\langle c_i : i \in J \rangle$ of objects of $C$ (in general, it is such a family with some arrows given as well). An arrow in $C^J$ from $\langle c_i \rangle$ to $\langle c'_j \rangle$ is any family of arrows $\langle \alpha_i : c_i \mapsto c'_j \rangle$ of $C$. Thus, $C^J$ is the (possibly infinite) product of $J$ copies of $C$. If $n$ is taken as the discrete $n$ element category then $C^n$ is $C \times \cdots \times C$, $n$ times. Another important example for $J$ is a preorder; in a functor $F$ of $C^J$, if $i \leq j$ there is a unique map between $F(i)$ and $F(j)$, which we denote $F_{ij}$.

3. Universal arrows. Suppose $G : D \mapsto C$ is a functor, and $c \in C$. A universal arrow from $c$ to $G$ is defined to be a pair $\langle d, f \rangle$, $d \in D$, $f : c \mapsto G(d)$ such that given any other pair $\langle d', f' \rangle$, $d' \in D$, $f' : c \mapsto G(d')$ there is a unique arrow $h : d \mapsto d'$ such that the diagram

\[
\begin{array}{ccc}
c & \overset{f}{\rightarrow} & G(d) \\
\downarrow f & & \downarrow G(h) \\
f' & \overset{G(h)}{\rightarrow} & G(d')
\end{array}
\]

(2)
is commutative. If \(|d, f|\) is universal then \(|d', f'|\) is also universal iff \(h\) is an isomorphism. One direction follows by composition with the isomorphism and its inverse. For the other, let \(h : d \mapsto d', h' : d' \mapsto d\) be the two maps. Then \(G(h'h)f = f\), and since the map with this property is unique \(h'h = \iota_d\); similarly \(hh' = \iota_{d'}\). In particular the objects of two universal arrows are isomorphic, by a unique isomorphism.

A universal arrow resembles an initial object; indeed it is one in the category \((c \downarrow G)\), which has as objects the pairs \(|d, f|\), \(d \in D\), \(f : c \mapsto G(d)\); an arrow \(h : d \mapsto d'\) is an arrow from \(|d, f|\) to \(|d', f'|\) exactly if diagram (2) commutes.

A universal arrow from \(G\) to \(c\) is defined to be a pair \(|d, f|\), \(d \in D\), \(f : G(d) \mapsto c\) such that given any other pair \(|d', f'|\), \(d' \in D\), \(f' : G(d') \mapsto c\) there is a unique arrow \(h : d' \mapsto d\) making the diagram

\[
\begin{array}{ccc}
G(d) & \xrightarrow{G(h)} & G(d') \\
\downarrow{G(h)} & & \downarrow{G(h')}
\end{array}
\]

commutative. This is a terminal object in the category \((G \downarrow c)\), whose definition we leave to the reader. Note also that it is a universal arrow from \(c\) to \(G\) in \(C^{\text{op}}\), when \(G\) is considered as a functor from \(D^{\text{op}}\) to \(C^{\text{op}}\).

**Theorem 2.** If a universal arrow \(|d, f|\) from \(c\) to \(G\) (resp. \(G\) to \(c\)) exists for each \(c \in C\), choose one and let \(F(c) = d\). Given \(g : c_1 \mapsto c_2\), there is a unique \(h : d_1 \mapsto d_2\) (where \(d_i = F(c_i)\)) such that \(G(h)f_1 = f_2g\) (resp. \(f_2G(h) = gf_1\); let \(F(g) = h\). The resulting map \(F : C \mapsto D\) (resp. \(F : C^{\text{op}} \mapsto D^{\text{op}}\)) is a functor.

**Proof:** The entire diagram

\[
\begin{array}{ccc}
c_1 & \xrightarrow{g} & c_2 \\
\downarrow{f_1} & & \downarrow{f_2} \\
G(d_1) & \xrightarrow{G(h)} & G(d_2) \\
\downarrow{G(h)} & & \downarrow{G(h')} \\
G(d_3) & \xrightarrow{f_3} & G(d_3)
\end{array}
\]

commutes. Since the map from \(c_1\) to \(G(d_3)\) is unique, it follows that \(F(g'g) = F(g')F(g)\). The claim for arrows from \(G\) follows similarly.

Given two functors \(F(c) = d\), \(F'(c) = d'\) as in the theorem let \(\theta_c : d \mapsto d'\) be the unique isomorphism given by universality of \(d, d'\). We claim that the \(\theta_c\) are a natural equivalence from \(F\) to \(F'\). Indeed, given \(g : c_1 \mapsto c_2\) let \(f_i (f'_i)\) be the map from \(c_i\) to \(GF(c_i) (GF'(c_i))\), and let \(h = F(g)\), \(h' = F'(g)\). Then

\[
G(h'\theta_c) f_1 = G(h')G(\theta_c)f_1 = G(h')f'_1 = f'_2g = G(\theta_{c_2})f_2g = G(\theta_{e_2})G(h)f_1.
\]

By uniqueness of the map from \(d_1\) to \(d_2\), \(h'\theta_{c_1} = \theta_{c_2}h\), which shows that \(\theta\) is natural.

If \(G : D \mapsto \text{Set}\) is the forgetful functor for some standard category \(D\), and \(c \in \text{Set}\), if \(|d, f|\) is a universal arrow from \(c\) to \(G\) then \(d\) is often called a free object generated by \(c\). For example if \(D\) is \(\text{Struct}_L\) let \(d\) be the Herbrand structure generated by \(c\), that is, the set of terms over the language expanded with names for the elements of \(c\); relations other than equality are made false everywhere. Let \(f\) map \(a \in c\) to its name. It is readily verified that \(|d, f|\) is a universal arrow from \(c\) to the forgetful functor \(G : \text{Struct}_L \mapsto \text{Set}\); given \(f' : c \mapsto G(d')\), the required function \(h : d \mapsto d'\) of diagram (2) is defined to map \(t(a_1, \ldots, a_n)\) to \([t]_d (f'(a_1), \ldots, f'(a_n))\). If \(c = \emptyset\) then \(d\) is the initial object.

If \(D\) is \(\text{Mdl}_T\) where \(T\) is a set of equations, a free object is readily constructed. It is the quotient of the Herbrand structure \(d\) by the least congruence relation \(\equiv\) on \(d\) determined by \(T\), i.e., where the equations of

131
Suppose $K$ hold. Indeed let $t \equiv u$ iff $t = u$ follows from $T$ by the axioms of equality (i.e., from instances of $T$ and the axioms of equality by propositional calculus). Clearly $\equiv$ is the least congruence relation determined by $T$; the quotient $d/\equiv$ is a model of $T$. If $L$ contains no constants we must consider the category $\text{Mdl}_T^L$ for the free object to exist when $c = \emptyset$.

Note that we have not required that $T$ contain the axioms of equality, but rather that equality is treated as a logical symbol. By this convention, $\text{Struct}_L$ may be considered a special case of $\text{Mdl}_T$, namely where $T$ is empty, even an empty set of equations. Later in the text we will use this convention that $\text{Struct}_L$ as a special case of $\text{Mdl}_T$ where $T$ is a set of equations.

That $d/\equiv$ is the “most general” model of $T$ is exactly the claim that it is the free object generated by $c$. To see that this is so, given any map $f' : c \to G(d')$ define $h$ as for structures. If $t \equiv u$ holds in $d$, in the proof of $t = u$ the expansion names may be replaced by variables, yielding an equation which follows from $T$. It follows that $h(t) = h(u)$ holds in $d'$; thus $h$ induces a map from $d/\equiv$ to $d'$, which is clearly a homomorphism.

In specific cases the free object often has alternative descriptions. For example the free $R$-module $M_c$ generated by a set $c$ is all formal linear combinations $\sum_i r_i a_i$, where the $a_i$ are distinct elements of $c$. (The notion of a “formal linear combination” can readily be given a precise set-theoretic formulation, but it is common to use the less formal notion. Indeed, a formal linear combination is a map from a finite subset of $c$ to $R$. The collection of such is an $R$-module in an obvious manner.) $M_c$ is readily verified to be a free object; the arrow $f : c \to G(M_c)$ takes $a \in c$ to its counterpart in $M_c$. Given $f' : c \to G(N)$, the required function $h : M_c \to N$ of diagram (2) maps $\sum_i r_i f(a_i)$ to $\sum_i r_i f'(a_i)$.

There are important universal arrows to forgetful functors where $C$ is not Set. One example is provided by the free $R$-algebra over a monoid $S$; the forgetful functor takes an algebra to its multiplicative monoid. For another, define a graph to be a structure with sorts Obj and Ar, and functions Dom, Codom : Ar \to Obj (we consider only small graphs). The graphs are the objects of a category Graph. Let $G : \text{Cat} \to \text{Graph}$ be the forgetful functor. We leave it as an exercise to show that for any graph $c$ there is a universal arrow from $c$ to $G$. As a corollary we have a description of the free monoid generated by a set $c$.

Define a groupoid to be a small category where every arrow has a two sided inverse. The groupoids form the objects of a full subcategory $\text{Grpd}$ of $\text{Cat}$. Again we leave it as an exercise to show that there is a universal arrow from a graph $G$ to the forgetful functor from $\text{Grpd}$ to $\text{Graph}$, giving as a corollary a description of the free group generated by a set $c$.

We also leave it as an exercise to give descriptions of the free commutative group, ring, or commutative ring, generated by a set $c$. When $T$ is not a set of equations, a free object often still exists. For example, the free preorder on a set $c$ is that where $x \leq y$ iff $x = y$. This is also the free partial order.

4. Adjoint functors. The following lemma gives a useful characterization of universal arrows, and is known as Yoneda’s lemma.

**Lemma 3.** Suppose $D$ is a standard category, $d \in D$, and $K : D \to \text{Set}$ is a functor.

a. For any $u \in K(d)$ the map $g \mapsto K(g)(u)$ from $\text{Hom}_D(d,d')$ to $K(d')$ is a component of a natural transformation $\alpha_u$ from $\text{Hom}_D(d,-)$ to $K$.

b. Every natural transformation $\alpha$ from $\text{Hom}_D(d,-)$ to $K$ is obtained in this way from exactly one $u \in K(d)$, namely $u = \alpha(d)(u_d)$.

Suppose $K$ is $\text{Hom}_C(c,G(-))$ where $C$ is a standard category, $c \in C$, and $G : D \to C$ is a functor.

c. $u \in \text{Hom}_C(c,G(d))$ is a universal arrow from $c$ to $G$ iff $\alpha_u$ is a natural equivalence from $\text{Hom}_D(d,-)$ to $\text{Hom}_C(c,G(-))$.

**Proof:** That $\alpha_u$ is natural follows because, given $f : d' \to d''$, for $g \in \text{Hom}_D(d,d')$

$$K(f)(K(g)(u)) = K(f \circ g)(u).$$

132
Certainly $\alpha_u(d)(\iota_d) = u$. Suppose $\alpha$ is any natural transformation with $\alpha(d)(\iota_d) = u$; since $\alpha$ is natural, for $f \in \text{Hom}_D(d,d') \alpha(d')(f) = K(f)(\alpha(d)(\iota_d))$, or $\alpha(d')(f) = K(f)(u)$. For part $c$, $\alpha_u(d')(g) = G(g) \circ u$; so $u$ is universal exactly if $\alpha_u(d')$ is bijective for each $d'$.

The dual statements follow immediately by replacing $D$ with $D^{op}$. The case $K = \text{Hom}(c,-)$ in part b is of special interest; the natural transformations from $\text{Hom}(d,-)$ to $\text{Hom}(c,-)$ are in canonical correspondence with the arrows from $c$ to $d$. It follows using part c that the transformation is an equivalence if the arrow is an isomorphism (see section 8). A functor $H(d,-)$ is said to be representable, or represented by $d$. This case of Yoneda’s lemma shows that properties of the functor are related to properties of $d$.

Two functors $F : C \to D$ and $G : D \to C$ between standard categories are said to be adjoint if the functors $\text{Hom}_D(F(-),-)$ and $\text{Hom}_C(-,G(-))$ from $C^{op} \times D$ to Set are naturally equivalent. $F$ is called a left adjoint of $G$, and $G$ a right adjoint of $F$. If we think of $F$ as “raising” $c$, and $G$ as “lowering” $d$, then given $c, d$ we may raise $c$ or lower $d$, and the Hom sets will be naturally in one-to-one correspondence.

For the functors to be naturally equivalent there must be bijections

$$\alpha_{cd} : \text{Hom}_D(F(c), d) \leftrightarrow \text{Hom}_C(c, G(d))$$

such that, given $f : c' \to c$ and $g : d \to d'$, for any $h : F(c) \to d$

$$\alpha_{cd}(ghF(f)) = G(g)\alpha_{cd}(h)f.$$  

It suffices that this hold when $f$ or $g$ is an identity. The triple $\langle F, G, \alpha \rangle$ is called an adjunction, from $C$ to $D$.

Theorem 2 provides one example of adjoint functors. For arrows to $G$, the functor $F$ is a left adjoint to $G$; that $\text{Hom}_D(F(c), d)$ and $\text{Hom}_C(c, G(d))$ are naturally equivalent is just lemma 3.c. In particular a functor taking a set to a free object generated by it is a left adjoint to the forgetful functor. Dually, for arrows from $G$, $F$ is a right adjoint to $G$. This can also be seen by noting that $\text{Hom}_{D^{op}}(F(c), d)$ and $\text{Hom}_{C^{op}}(c, G(d))$ are naturally equivalent, and $\text{Hom}_{C^{op}}(c, c') = \text{Hom}_C(c', c)$.

On the other hand any left adjoint can be seen as arising in the manner of theorem 2. Suppose $F : C \to D$, $G : D \to C$ is an adjoint pair of functors, with natural equivalence $\alpha$. By lemma 3 there is a universal arrow from $c$ to $G$

- $\mu_c = \alpha_{cF(c)}(\iota_{F(c)})$; and $\alpha_{cd}(h) = G(h)\mu_c$

for $h : F(c) \to d$. Further, the map $c \mapsto \mu_c$ is a natural transformation from the identity functor on $C$ to $GF$; this follows since for $f : c \to c'$

$$G(F(f))\mu_c = G(F(f))\alpha_{cF(c)}(\iota_{F(c)}) = \alpha_{cF(c')}(F(f)\iota_{F(c)}) = \alpha_{cF(c')}(F(f))\mu_{c'}f = \mu_{c'}f$$

using the naturality of $\alpha$. Similarly there is a universal arrow from $F$ to $d$

- $\nu_d = \alpha_{G(d)}^{-1}(\iota_{G(d)})$; and $\alpha_{cd}^{-1}(h) = \nu_dF(h)$

for $h : c \to G(d)$. The map $d \mapsto \nu_d$ is a natural transformation from $FG$ to the identity functor on $D$.

The natural transformation $\mu$ is called the unit of the adjunction, and $\nu$ is called the counit. An adjunction is determined by the unit; indeed it is determined by $G$, the object function of $F$, and the universal arrows $\mu_c : c \to G(F(c))$ for each $c$. Similarly it is determined by $F$, the object function of $G$, and the universal arrows $\nu_d : F(G(d)) \to d$ for each $d$. By the remarks following theorem 2 any two left adjoints to a functor are naturally isomorphic.

The following diagram summarizes the universality property of the unit; letting $e = F(c)$, $\mu : c \to G(e)$ is the universal arrow. The bijection $\text{Hom}(e,d) \leftrightarrow \text{Hom}(c,G(d))$ is $h \leftrightarrow G(h)\mu$; the inverse bijection is
$f \mapsto \nu F(f)$. As just noted, a system $\{F(c) = e, \mu_c : c \mapsto G(e)\}$ of universal arrows to $G$ determines a left adjoint to $G$.

\[
\begin{array}{c}
\text{c} \\
\downarrow h \quad \downarrow f \\
G(e) \quad G(d)
\end{array}
\]

Now, $\alpha_{G(d)} = \alpha_{G(d)}d(\nu_d) = G(\nu_d)\mu_{G(d)}$ and $\alpha_{F(c)} = \alpha_{cF(c)}(\mu_c) = \nu_{F(c)}F(\mu_c)$.

Writing $G\mu$ for the natural transformation $d \mapsto G(\mu_d)$ and $\mu_G$ for the natural transformation $d \mapsto \mu_{G(d)}$, these say that the composition $G\nu \circ \mu_G$ of natural transformations is the identity natural transformation from $G$ to $G$; and $\nu F \circ F\mu$ the identity from $F$ to $F$. These are called the “triangular identities”. Two natural transformations satisfying them determine a unique adjunction with the transformations as unit and counit; we leave this as an exercise.

Another example of adjoint functors makes use of the following lemma. Recall that $\alpha_{XYZ}$ is the function $\alpha_{XYZ}(g) = x \mapsto (y \mapsto g(x, y))$.

That is, the function $g$ of two variables is mapped to a function $\bar{g}$ of one variable, where $\bar{g}(x)(y) = g(x, y)$; $\bar{g}(x)$ is the function $y \mapsto g(x, y)$. Clearly $\alpha_{XYZ}$ is a bijection; given $\bar{g}$ let $g(x, y) = \bar{g}(x)(y)$.

**Theorem 4.** $\alpha$ is natural from $\text{Hom}(- \times -, -)$ to $\text{Hom}(-, \text{Hom}(-, -))$.

**Proof:** Suppose $f : X' \mapsto X$; then

\[
\begin{array}{c}
\langle x, y \rangle \mapsto g(x, y) \\
\downarrow \quad \downarrow \alpha_{XYZ} \\
\langle x', y \rangle \mapsto g(f(x'), y)
\end{array}
\]

Now, $\bar{g}$ gets mapped to $\bar{g}f$ by $\text{Hom}(-, \text{Hom}(Y, Z))(f)$, and

\[
\bar{g}f(x') = \bar{g}(x) = y \mapsto g(x, y) = y \mapsto g(f(x'), y),
\]

so $\alpha$ is natural in $X$. Suppose $f : Y' \mapsto Y$; then

\[
\begin{array}{c}
\langle x, y \rangle \mapsto g(x, y) \\
\downarrow \quad \downarrow \alpha_{XYZ} \\
\langle x, y' \rangle \mapsto g(x, f(y'))
\end{array}
\]

Now, $\text{Hom}(-, Z)(f)$ maps $y \mapsto g(x, y)$ to $y' \mapsto g(x, f(y'))$, so

\[
\text{Hom}(X, \text{Hom}(-, Z)) \text{ maps } \bar{g} \text{ to } \bar{g}' = \text{Hom}(-, Z)(f)\bar{g};
\]

134
Given functor $g : C \to D$.

**Proof:**

Suppose $f : Z \to Z'$; then

$$g'(x) \text{ thus equals } \operatorname{Hom}(-, Z)(f)(g(x)), \text{ showing that } \alpha \text{ is natural in } Y. \text{ Suppose } f : Z \to Z' ; \text{ then}

$$

\[
\begin{array}{cccc}
(x, y) & \mapsto g(x, y) & \xrightarrow{\alpha \times_{Y} \beta} x & \mapsto (y \mapsto g(x, y)) = \bar{g}
\end{array}
\]

Now, $\operatorname{Hom}(Y, -)(f)$ maps $y \mapsto g(x, y)$ to $y \mapsto f(g(x, y))$, so

\[
\operatorname{Hom}(X, \operatorname{Hom}(Y, -)) \text{ maps } \bar{g} \text{ to } \bar{g}' = \operatorname{Hom}(Y, -)(f)\bar{g};
\]

$\bar{g}'(x)$ thus equals $\operatorname{Hom}(Y, -)(f)(\bar{g}(x))$, showing that $\alpha$ is natural in $Z$.

In particular, $\operatorname{Hom}(Y, -)$ is a right adjoint to $- \times Y$. By definition, this shows that Set is a “Cartesian closed” category; further discussion is deferred to section 18.2.

**5. Limits.** Another important class of universal arrows are those to some $F$ in a functor category $C^J$ from the “diagonal functor” $\Delta : C \to C^J$. The diagonal functor maps the object $c$ to the functor $\Delta_c$ whose objects $\Delta_c(i)$ are all $c$, and whose maps $\Delta_c(f) : \Delta_c(i) \mapsto \Delta_c(j)$ are all the identity $\iota_c$.

An arrow in $C^J$ to $F$ from $\Delta(c)$ consists of maps $\alpha_i : c \mapsto F(i)$ such that each diagram

\[
\begin{array}{ccc}
F(i) & \xrightarrow{F(f)} & F(j) \\
\downarrow{\alpha_i} & & \downarrow{\alpha_j} \\
c & & c
\end{array}
\]

for $f \in \operatorname{Hom}_J(i, j)$ commutes. Such an arrow is called a cone to $F$ from $c$. A universal arrow is called a limit. That is, a limit is a pair $\langle c, \alpha \rangle$ where $\alpha$ is a cone to $F$ from $c$, such that for any other cone $\alpha'$ to $F$ from $c'$ there is a unique arrow $h : c' \mapsto c$ such that $\alpha_i h = \alpha'_i$; $c$ is called the limit object and $\alpha$ the limit cone.

One may similarly consider universal arrows in $C^J$ to $\Delta$ from some $F \in C^J$. The arrows are called cones to $c$ from $F$, and a universal arrow is called a colimit. That is, a colimit is a pair $\langle c, \alpha \rangle$ where $\alpha$ is a cone from $F$ to $c$, such that for any other cone $\alpha'$ from $F$ to $c'$ there is a unique arrow $h : c \mapsto c'$ such that $h \alpha = \alpha'_i$. If $J$ is fixed, the limit and colimit are often said to be dual.

Suppose $C^J$ is a functor category for some small category $J$, and for every $F \in C^J$ the limit exists. Then there is a functor $\operatorname{Lim}_J$ from $C^J$ to $C$, giving the limit object, which is right adjoint to the diagonal functor from $C$ to $C^J$. Dually if the colimit always exists there is a colimit functor $\operatorname{Colim}_J$.

A category where all limits exist is called complete; and where all colimits exist cocomplete.

**THEOREM 5.** Set is complete.

**Proof:** Given $F \in \operatorname{Set}^J$ call a sequence $\langle x_i \rangle$, $x_i \in F(i)$, $i \in J$, consistent if $F(g)(x_i) = x_j$ for all $g \in \operatorname{Hom}_J(i, j)$, $i, j \in J$. Let $c$ be the set of consistent sequences. Define a cone $\alpha$ from $c$ to $F$, by letting $\alpha_i$ be the projection map (i.e., $\alpha_i((x_i)) = x_j$). That this is a cone follows since the elements of $c$ are consistent. We claim that $\langle c, \alpha \rangle$ is universal. Given a cone $\alpha'$ from $c'$ to $F$, define $h : c' \mapsto c$ by setting $h(w) = (\alpha'_i(w))$; this clearly is the unique map where $\alpha_i h = \alpha'_i$. It must be shown that $h(w)$ is consistent, i.e., that $F(g)(\alpha'_i(w)) = \alpha_j(w)$; but this is just the requirement that $\alpha'$ be a cone.

Limits may be obtained in Struct$_L$ by defining a structure on the limit in Set in an obvious way. Indeed, say that a functor $G : D \to C$ determines limits for an $F \in D^J$ if, for each limit $\langle c, \alpha \rangle$ of $GF$ in $C$ there is a unique pair $\langle d, \beta \rangle$ with $\beta$ a cone from $d$ to $F$, such that $G(d) = c$, $G(\beta_i) = \alpha_i$ for all $i \in J$, and $\langle d, \beta \rangle$ is a limit of $F$. 135
Theorem 6. The forgetful functor $G : \text{Struct}_L \to \text{Set}$ determines limits, for any $J \in \text{Cat}$ and any $F \in \text{Struct}_L^J$.

Proof: First suppose there are no relation symbols. It suffices to consider the limit $\langle c, \alpha \rangle$ of theorem 5. Now, $d$ must have $c$ as its domain and $\beta_i$ equal $\alpha_i$, so we must define a structure on $c$ so that $\alpha_i$ is a homomorphism. In particular, for a function $f$ we must have

$$\alpha_i(f(\langle x_{1i}, \ldots, x_{ni} \rangle)) = f(\alpha_i(\langle x_{1i} \rangle), \ldots, \alpha_i(\langle x_{ni} \rangle)),$$

or

$$f(\langle x_{1i}, \ldots, x_{ni} \rangle) = (f(x_{1i}, \ldots, x_{ni})).$$

so that $f$ must be defined in the usual “pointwise” manner. It must be verified that the sequence on the right is consistent if the $\langle x_{ti} \rangle$ are, i.e., that for $g : i \to j$ if $F(g)(x_{ti}) = x_{tj}$ then $F(g)(f(x_{1i}, \ldots, x_{ni})) = f(x_{1j}, \ldots, x_{nj})$. This follows since $F(g)$ is a homomorphism. Finally $\langle c, \alpha \rangle$ with this structure on $c$ is universal in $\text{Struct}_L$, because the map $h$ of theorem 1 is a homomorphism if the $\alpha'_i$ are, as is readily verified. If there are relation symbols we obtain a limit by making $R$ defined dually to that of determining limits.

Corollary 7. Suppose $T$ is a set of universally quantified Horn formulas. The forgetful functor $G : \text{Mdl}_T^J \to \text{Set}$ determines limits, for any $J \in \text{Cat}$ and any $F \in \text{Mdl}_T^J$.

Proof: The structure $c$ is a substructure of the product of the $F(i)$. Since the axioms are Horn formulas the product is in $\text{Mdl}_T^J$, and so $c$ is.

If $L$ contains no relation symbols $\langle d, \beta \rangle$ is determined by the requirements $G(d) = c$, $G(\beta_i) = \alpha_i$ for all $i \in J$. In this case $G$ is said to create limits.

If $J$ is discrete a limit is called a product. The components of the limit cone are called the projections. Thus, we have another proof that products exist in $\text{Struct}_L$, etc.; indeed they may be defined on the Cartesian product in a unique manner. If a product of $a$ and $b$ exists for all $a, b$ we use $a \times b$ to denote the object of some chosen one; projections are also chosen. By theorem 2 this determines a unique functor. More generally we may write, for example, $\times_i \alpha_i$. The same convention applies to other kinds of limits.

A commonly used property of products states that if $f_i : a_i \to b_i$ is a family of morphisms a unique morphism $f : \times a_i \to \times b_i$ is induced such that $\pi_i f = f_i \pi_i$ for all $i$. This is an immediate consequence of universality; the map $f$ is often denoted $\times_i f_i$.

Some other examples of limits are as follows.

- If $J$ is the empty functor a limit is a terminal object. We assume Cat contains the empty category.
- If $J$ is a filtered preorder a functor from $J$ to $C$ is called an inverse system, and a limit is also called an inverse limit.
- If $F$ consists of objects $a_1, a_2, b$, with arrows $f_1 : a_1 \to b$, $f_2 : a_2 \to b$, a limit is called a pullback.
- If $F$ consists of objects $a, b$ with arrows $f, g : a \to b$ a limit is called an equalizer. This is a map $h : c \to a$ with $fh = gh$, such that for any other $h' : c \to a$ with $fh' = gh'$ there is a unique $k : c' \to c$ such that $hk = h'$. In Set, $\{ x \in a : f(x) = g(x) \}$, with the inclusion map, is an equalizer.
- If $C$ has a zero object an equalizer of $f : a \to b$ with 0, the zero map from $a$ to $b$, is called a kernel.

In $\text{Grp}$ or $\text{Mod}_R$ the usual kernel, with the inclusion map, is a kernel; in $\text{Mod}_R$ the equalizer of two functions is the kernel of their (pointwise) difference.

Given an indexed family $c_i$ of sets, the disjoint union $c$ of the family is defined to be $\bigcup_i c_i \times \{ i \}$. There is a canonical injection from $c_i$ to $c$, taking $x \in c_i$ to the pair $\langle x, i \rangle$. The notion of determining colimits is defined dually to that of determining limits.
THEOREM 8. If $J$ is a directed preorder and $F \in \text{Set}^J$ then a colimit of $F$ exists.

PROOF: Let $c$ be the disjoint union of the $F(j)$. Define the relation $x \equiv y$ on $c$ iff $x \in F(i)$, $y \in F(j)$, and for some $m \geq i, j$ in $J$, $F_{im}(x) = F_{jm}(y)$. This relation is trivially reflexive and symmetric. It is also transitive; if $F_{im}(x) = F_{jm}(y)$ and $F_{jn}(y) = F_{kn}(z)$ choose $p \geq m, n$. Then $F_{ip}(x) = F_{kp}(z)$. Define the map $\alpha_i : F(i) \rightarrow c/\equiv$ by taking $x$ to its equivalence class in the disjoint union. The $\alpha_i$ comprise a cone, since if $y = F_{ij}(x)$ then $x \equiv y$. We claim that $(c/\equiv, \alpha)$ is a colimit. Given a cone $\alpha'$ from $c$ to $c'$, and $x \in c$, define $h(x)$ to be $\alpha_i'(w)$ for any $w$ such that $\alpha_i(w) = x$. If $\alpha_j(u) = x$ then for some $m F_{im}(w) = F_{jm}(u)$, so

$$\alpha_j'(w) = \alpha_m'(F_{im}(w)) = \alpha_m'(F_{jm}(u)) = \alpha_j'(u),$$

so $h$ is well defined. It is clearly the unique map from $c$ to $c'$ such that $h(\alpha_i(w)) = \alpha_i'(w)$.

THEOREM 9. The forgetful functor $G : \text{Struct}_L \rightarrow \text{Set}$ determines colimits, for any directed preorder $J$ and any $F \in \text{Struct}_L^J$.

PROOF: Again we first suppose there are no relation symbols, and construct a unique structure in $c/\equiv$ such that $\alpha_i$ is a homomorphism. Given $x_1, \ldots, x_n$ there is an $i$, and $w_i \in F(i)$, such that for each $t \alpha_i(w_i) = x_i$. Take any such, and let $f(x_1, \ldots, x_n) = f(w_1, \ldots, w_n)$. Clearly $h$ is (if well defined) the unique function such that $\alpha_i(f(x_1, \ldots, x_n)) = f(\alpha_i(w_1), \ldots, \alpha_i(w_n))$. It is well defined, since if we chose $j$ and $w_t$ instead, there would be a $k \geq i, j$ with $F_{im}(w_i) = F_{jm}(u_t)$, and

$$F_{im}(f(w_1, \ldots, w_n)) = f(F_{im}(w_1), \ldots, F_{im}(w_n)) = f(F_{jm}(u_1), \ldots, F_{jm}(u_n)) = F_{jm}(f(u_1, \ldots, u_n)).$$

The map $h$ of theorem 8 is a homomorphism, since the $\alpha_i'$ are. If there are relations, we must define $R(x_1, \ldots, x_n)$ to hold iff there are $i, w_i$ with $\alpha_i(w_i) = x_i$, such that $R(w_1, \ldots, w_n) \in F(i)$.

COROLLARY 10. Suppose $T$ is a set of open formulas. The forgetful functor $G : \text{Mdl}_T^+ \rightarrow \text{Set}$ determines colimits, for any directed preorder $J$ and any $F \in \text{Mdl}_T^J$.

PROOF: We must show that the colimit structure is a model of $T$. This follows since $\alpha_i$ is a homomorphism, and any finite set in the domain of the colimit structure is the image under $\alpha_i$ of some subset of $F(i)$.

If $J$ is a directed preorder a functor from $J$ to $C$ is called a direct system, and its colimit also called a direct limit. It can be seen from the proof of theorem 8 that direct limits exist in $\text{Set}^-$. It follows that they exist in $\text{Struct}_L^-$. Inverse limits do not exist in $\text{Set}^-$; consider the system of open intervals $\{0, 1/n\}$ in $\mathcal{R}$, ordered by inclusion.

The dual to a product, i.e., a colimit when $J$ is discrete, is called a coproduct. The components of the colimit cone are called the injections. We use the symbol $\oplus$ to denote a chosen coproduct. The disjoint union $c$ is a coproduct in $\text{Set}$. Indeed, given maps $f_i : c_i \rightarrow c'$ the map $h$ must take $(x, i) \in c$ to $f_i(x)$. The forgetful functor does not determine coproducts; however there is a general construction.

THEOREM 11. The coproduct of $a_1$ and $a_2$ exists in $\text{Struct}_L$. It also exists in $\text{Mdl}_T$ if $T$ is a set of equations.

PROOF: First suppose $L$ has no relation symbols. Let $c$ be the Herbrand structure generated by $a_1$ and $a_2$, where these are assumed disjoint. Let $E_i$ be the closed equations which hold in $a_i$. Let $\equiv$ be the congruence relation determined by $E_1 \cup E_2$. We claim that $c/\equiv$ is a coproduct, with the injection map from $a_i$ to $c$ mapping $w$ to its equivalence class. Given $f_i : a_i \rightarrow c'$ define a map $h : c \rightarrow c'$ by letting $h(t(w_1, \ldots, w_n, v_1, \ldots, v_m))$ be $t(w_1, \ldots, w_n, v_1, \ldots, v_m))$ when $w_i \in a_1, v_i \in a_2$ equal $\lbrack t \rbrack c'$ of $f_i(w_1), \ldots, f_i(w_n), f_i'(v_1), \ldots, f_i'(v_m))$. If $t \equiv u$ then there is a PC proof, using instances of the axioms of equality, of a formula $H \Rightarrow t = u$, where $H$ is the conjunction of a finite set of hypotheses from $E_1 \cup E_2$. Replace $w_i$ by $f_i(w_i)$ and $v_i$ by $f_2(v_i)$, yielding $H' \Rightarrow t' = u'$, which holds in $c'$. Since $f_i$ is a homomorphism, $H'$ holds in $c'$, so $t' = u'$ does. Thus $h$ is well defined. The induced map is the required homomorphism from $c/\equiv$ to $c'$. The proof is identical for $\text{Mdl}_T$, allowing instances of $T$ in the proof of $t = u$. If there are relation symbols, define $R(x_1, \ldots, x_k)$ to hold iff $x_1, \ldots, x_k$ all have representatives which are expansion constants from the same $a_i$. 137
By exercise 2 these categories have all nonempty finite coproducts. A category which has nonempty finite coproducts and colimits of directed preorders has all nonempty coproducts. Indeed, given \( J \) let \( J' \) be the finite subsets \( S \) of \( J \) ordered by inclusion. Given \( F \in C^J \) let \( F' \in C^{J'} \) have \( F'(S) \) the coproduct of \( S \); and \( F_{ST} \) for \( S \subseteq T \) the inclusion map determined by the injections. The colimit is easily seen to be a coproduct for \( J \). Finally, an empty coproduct is an initial object.

The existence of arbitrary coproducts can be proved directly, without using theorem 9. In theorem 11, consider any set \( \{a_i\} \) of structures. Let the structure \( c \) be that generated by all the constants, and proceed similarly.

For some specific examples, in \( \text{Mod}_R \) the direct sum is a coproduct; this recall is those elements of the Cartesian product which are nonzero only finitely often. The injection map \( f_i \) maps \( x \in c_i \) to the sequence which is \( x \) in the \( i \)th component and 0 elsewhere. Given maps \( f'_i : c_i \to c' \), the map \( h : c \to c' \) must take \( \sum_j f_i(x_i) \) to \( \sum_j f'_i(x_i) \). In \( \text{Grp} \) the coproduct of a family of groups is those words which have no two consecutive letters from the same group. Multiplication of such is obtained by concatenating, multiplying adjacent elements in the same group, and dropping identitites. This is also called the free product of the groups.

The dual of an equalizer is called a coequalizer. Given \( f, g : a \to b \) a coequalizer is a map \( h : b \to c \) such that \( hf = hg \), and for any map \( h' : b \to c' \) with \( h'f = h'g \) there is a unique map \( k : c \to c' \) such that \( kh = h' \). Coequalizers exists in the categories of theorem 11, and \( \text{Set} \). Let \( E \) be the equations in \( b \), of the form \( f(u) = g(u) \) where \( u \in a \). Let \( \equiv \) be the least congruence relation determined by \( E \cup T \). Then the canonical epimorphism from \( b \) to \( b/\equiv \) is a coequalizer. Indeed, if \( v \equiv w \) then \( h'(v) = h'(w) \), by induction on the length of the proof, and \( k \) is then the map induced on \( b/\equiv \).

**Theorem 12.** If nonempty products and equalizers exist in a category \( C \) then nonempty limits exist. Dually if nonempty coproducts and coequalizers exist then nonempty colimits exist.

**Proof:** Suppose \( F \in C^J \) is given. Let \( a \) be the product of the indexed family \( F(i) \), \( i \in J \), with cone \( \alpha \); and let \( b \) be the product of the indexed family \( F(\text{dom}(l)) \), \( l \in \text{Ar}J \), with cone \( \beta \). Let \( f : a \to b \) be the map determined by the cone \( \alpha_{\text{dom}(l)} \) from \( a \) to \( F(\text{dom}(L)) \), and \( g : a \to b \) the map determined by the cone \( F(l)\alpha_{\text{dom}(l)} \). Let \( h : c \to a \) be the equalizer of \( f, g \). We claim that \( \langle c, \gamma \rangle \), where \( \gamma_i = \alpha_i h \), is a limit. Certainly \( \gamma \) is a cone, because \( \alpha \) is. Given any other cone \( \langle c', \gamma' \rangle \) to \( F \), let \( h' : c' \to a \) be the unique map such that \( \gamma'_i = \alpha_i h' \). Now,

\[
\beta_i f h' = F(l)\alpha_{\text{dom}(l)} h' = \alpha_{\text{dom}(l)} h' = \beta_i f h',
\]

and since \( b \) is a product \( gh' = fh' \). Since \( h \) is an equalizer there is a unique map \( k : c' \to c \) such that \( hk = h' \); \( k \) is then the unique map such that \( \gamma'_i = \alpha_i hk = \gamma_i k \).

Thus, colimits exist in \( \text{Set} \), \( \text{Struct}_L \), \( \text{Mdl}_T^+ \) for \( T \) a set of equations; and nonempty colimits exist in \( \text{Set}^- \), \( \text{Struct}_L^- \), and \( \text{Mdl}_T \) for \( T \) a set of equations. As another application, by exercise 2 and the proof of the theorem, if the product of any pair of objects, and equalizers, exist than all finite limits do; and dually for finite colimits.

The dual of a kernel, that is, a coequalizer with a zero map, is called a cokernel. In \( \text{Grp} \) this is \( b/n \) where \( n \) is the smallest normal subgroup containing \( f[a] \). In \( \text{Mod}_R \) it is \( b/f[a] \). The dual of a pullback is called a pushout; that is, given \( f_1 : a \to b_1, f_2 : a \to b_2 \), a pushout is a colimit pair of maps \( g_1 : a \to c, g_2 : b_2 \to c \) such that \( g_1 f_1 = g_2 f_2 \). In \( \text{Grp} \) the pushout of monomorphisms \( f_1, f_2 \) is called the amalgamated product; it may be defined similarly to the free product of \( b_1, b_2 \), using the same letters for members of \( a \).

We have been ignoring some metamathematical issues arising from the fact that a standard category \( C \) is a proper class; some discussion will be given here. There are two methods for handling proper classes in axiomatic set theory. One is to require that any proper class \( X \) be given by a formula of set theory with one
free variable, and prove theorems using the system of axioms known as ZFC (Zermelo-Fraenkel with choice); call these strict proper classes. The other is to add variables for classes, which must always occur as free variables. Theorems are proved in the system of axioms GBC (Gödel-Bernays with choice).

GBC may be shown to be a “conservative extension” of ZFC, meaning that theorems without class variables, which are provable in GBC, are provable in ZFC ([Drake], exercise 5.1.13(2)). This is easy to see if the formula is obtained from a proven GBC formula by replacing the free class variables by formulas (and performing well-known operations to make sure the result is as intended). The class variable proof is a “template” for a proof of the formula applying to any strict specific proper class.

This suffices, for example, to give a precise definition of a category; it might as well be a proper class in the strictest sense, although we can argue with phrases such as “a category C”, which is clearly convenient.

A more complicated problem arises in theorem 2, where a universal arrow is chosen for each c ∈ C, given that such exists. As stated, this requires a form of the axiom of choice stronger than the usual form, which states that a choice function exists for any set; namely, that a choice function exists for a proper class.

The resulting system, which we denote GBCC (CC for class choice), is again a conservative extension ([Jech], section 20, last paragraph). It is convenient to be able to use CC, and there is little objection to doing so. However, mathematical logicians prefer to avoid it when possible. Doing so in category theory is easier in some cases than in others.

One method is to require that every proper class be a strict one. Thus, implicitly in theorem 2 we might assume that the universal object is given by a formula, which is provably single valued. This could also be required for any Lim, such as ×, etc. Indeed, this is what is done in a concrete category when a specific limit is constructed. Alternatively, a × b might stand for any product; this will cause no problems in some arguments, but in others one might have to restrict the argument to a subset of the universe, etc. We will continue to use class choice in an informal manner, since this is convenient and mathematical logic has taught us that there will be some resolution of any difficulties that seem to arise.

6. Monics and Epics. An arrow f : b → c is said to be
- monic if fh = f'h' ⇒ h = h' for any h, h' : a → b;
- split monic if f'f = ιb for some f' : c → b;
- epic if hf = h'f ⇒ h = h' for any h, h' : c → a; and
- split epic if ff' = ιc for some f' : c → b.

Split epics are also called retractions, and split monics coretractions.

In any concrete category, split monics are injective and split epics surjective; also injective arrows are monic and surjective arrows epic. It follows that the two notions of isomorphism (category theoretic and universal algebraic) agree in StructL, or its full subcategory MdlF. In Set all nonempty monics and epics are split. Indeed, if f : S → T is not injective, say f(s) = f(s') where s' ≠ s, let h(0) = s, h'(0) = s' be maps from {0} to S; then fh = f'h' but h ≠ h'. If f : S → T is not surjective, say t ∉ f[S], let h(t) = 0, h'(t) = 1, h, h'(t') = 0 if t' ≠ t, be maps from T to {0, 1}; then hf = h'f but h ≠ h'.

In a concrete category it may or may not be the case that monics are injective, or epics surjective (and even if so they may not be split). In Grp monics are injective; for if f is not injective let h be the inclusion of the kernel and let h' map the kernel to 0. It is also true that epics are surjective in Grp; we leave this as an exercise. Epics are not surjective in Rng; the embedding of an integral domain in its field of fractions is epic, but not surjective in general.

The composition of monics is readily verified to be monic, and the composition of epics epic. The same is true for split monics and split epics. Also, if fg is monic then g is monic; and if fg is epic then f is epic. For another useful fact, it is readily verified that if the fi are monic then ∏i fi is monic.

An arrow which is both split monic and split epic is an isomorphism, by a well known argument. In
fact, an arrow \( f : b \to c \) which is split monic and epic, or dually monic and split epic, is an isomorphism. Indeed, suppose \( f'f = \iota_b \) and \( gf = g'f \Rightarrow g = g' \); then \( f'f = f \) and \( \iota_c f = f \), so \( f f' = \iota_c \). A category is called balanced if an arrow which is both monic and epic is an isomorphism. Grp is balanced; the arrow is bijective, so an isomorphism. Rng is not balanced; the embedding of an integral domain in its field of fractions is certainly monic as well as epic. This applies equally well to CRng.

**Theorem 13.** An equalizer is monic; dually a coequalizer is epic.

**Proof:** Suppose \( f : a \to b \) is the equalizer of \( h_1, h_2 : b \to c \), and \( fg_1 = fg_2 \) where \( g_i : d \to a \). Then \( h_1f g_i = h_2f g_i \) for \( i = 1, 2 \), whence \( f g_i = f \alpha \) for a unique \( \alpha \), and \( g_1 = g_2 = \alpha \).

**Theorem 14.** If \( g_i : a \to b_i, i = 1, 2 \), is a pullback of \( f : b \to c \), and \( f_1 \) is monic, then \( g_2 \) is monic. Dually if \( g_i : b_i \to c, i = 1, 2 \), is a pushout of \( f_i : a \to b_i \), and \( f_1 \) is epic, then \( g_2 \) is epic.

**Proof:** Suppose \( h, h' : d \to a \) satisfy \( g_2h = g_2h' \). Then \( f_2g_2h = f_2g_2h' \), so \( f_1g_1h = f_1g_1h' \). Since \( f_1 \) is monic, \( g_1h = g_1h' \). By uniqueness of the map from \( d \) to \( a \), \( h = h' \).

Under mild conditions the projections of a product are split epic. It suffices that there be arrows between any two \( a_i \) in the product \( \times_i a_i \), which is certainly the case if a 0 object exists. Dually the injections of a coproduct are usually split monic.

7. **Preservation of limits.** Suppose \( G : D \to C, H \in D^J \), and \( \alpha \) is a cone from \( d \) to \( H \) in \( D \). Recalling that \( \alpha \) is a map from \( \text{Obj}_J \) to \( \text{Arr}_D \), \( G\alpha \) is readily seen to be a cone from \( G(d) \) to \( GH \) in \( C \). \( G \) is said to preserve limits if \( (G(d), G\alpha) \) is a limit of \( GH \) in \( C \) whenever \( (d, \alpha) \) is a limit of \( H \) in \( D \).

Use of the term “preserves” is generalized from the above. A functor may preserve limits of some particular small category \( J \), or of some set of \( J \); for example it may preserve products, or kernels, etc. A functor preserves monics if \( F(f) \) is monic whenever \( f \) is; and similarly for epics. Note that a functor always preserves split monics, split epics, and isomorphisms.

A contravariant functor \( F \) on \( C \) is said to have a preservation property iff \( F \), considered as covariant functor on \( C^{op} \), does. For a small category \( J \), a contravariant functor \( F \) preserves limits of \( J \) if it maps a limit in \( D^{op} \) of \( J \), or equivalently a colimit in \( D \) of \( J^{op} \), to a limit of \( J \). For example \( F \) preserves products iff it maps coproducts to products; and monics if it maps epics to monics.

**Theorem 15.** Suppose \( (F, G, \psi) \) is an adjunction from \( C \) to \( D \). Then \( G \) preserves limits and monics, and dually \( F \) preserves colimits and epics.

**Proof:** Suppose \( (d, \alpha) \) is a limit of \( H \) in \( D \), and \( (c, \beta) \) is any cone to \( GH \). Then there is a unique arrow \( f : F(c) \to d \) with \( \alpha f = \psi_c^{-1} H(\beta) \); and \( f' = \psi_{cd}(f) \) is a map from \( c \) to \( G(d) \), and

\[
G(\alpha f) = G(\alpha i) \psi_{cd}(f) = \psi_{c, H(\beta)}(\alpha f) = \beta i.
\]

On the other hand given such an \( f' \), it equals \( \psi_{cd}(f) \) for some \( f \), which satisfies \( \alpha f = \psi_{c, H(\beta)}^{-1}(\beta i) \); thus, \( f' \) is unique. Suppose \( f : d \to d' \) is monic, and for \( g_1, g_2 : c \to G(d) \), \( G(f)g_1 = G(f)g_2 \). Then, since

\[
\psi_{cd}(f \psi_{cd}^{-1}(g_1)) = G(f)g_1, \quad \psi_{cd}(f \psi_{cd}^{-1}(g_1)) = \psi_{cd}(f \psi_{cd}^{-1}(g_2)), \quad \text{so} \quad f \psi_{cd}^{-1}(g_1) = f \psi_{cd}^{-1}(g_2), \quad \text{so} \quad \psi_{cd}^{-1}(g_1) = \psi_{cd}^{-1}(g_2), \quad \text{so} \quad g_1 = g_2.
\]

Recall the covariant functors \( \text{Lim}_J \) and \( \text{Colim}_J \) from \( C^J \) to \( J \) defined in section 5. As a consequence of the foregoing, the \( \text{Lim}_J \) preserves limits and monics, and \( \text{Colim}_J \) preserves colimits and epics.

It is convenient to introduce a notation for the covariant functor \( \text{Hom}(a, -) \) from \( D \) to \( \text{Set} \); we will use \( \text{Hom}_a \), and also \( \text{Hom}_b \) for the contravariant functor \( \text{Hom}(-, b) \).

**Theorem 16.** The functor \( \text{Hom}_a \) from a category \( D \) to \( \text{Set} \) preserves limits and monics. Dually \( \text{Hom}_b \) preserves limits and monics.
Proof: Suppose \( \langle d, \alpha \rangle \) is a limit of \( K \) in \( D \), and \( \langle c, \beta \rangle \) is any cone to \( \text{Hom}^a K \). For each \( x \in c \beta_i(x) \) is an arrow from \( a \) to \( K(i) \); further if \( K(t) : K(i) \rightarrow K(j) \) then \( K(t) \beta_i(x) = \beta_j(x) \). That is, the \( \beta_i(x) \) form a cone from \( a \) to \( K \). There is thus a unique arrow \( f_x \) from \( a \) to \( d \) such that \( \alpha_i f_x = \beta_i(x) \). This determines a function \( f' \) from \( c \) to \( \text{Hom}(a, d) \); further

\[
((\text{Hom}^a(\alpha_i)) \circ f')(x) = \alpha_i f_x = \beta_i(x), \quad \text{so} \quad (\text{Hom}^a(\alpha_i)) \circ f' = \beta_i.
\]

On the other hand given such an \( f' \), writing \( f_x \) for \( f'(x) \), it must satisfy \( \alpha_i f_x = \beta_i(x) \). By definition an arrow \( f : b \rightarrow b' \) is monic iff for all \( a \), \( \text{Hom}^a(f) \) is injective, i.e., monic in \( \text{Set} \). In particular, \( \text{Hom}^a \) preserves monics.

8. Galois adjunctions. An adjunction between preorders is called a Galois adjunction. That is, a Galois adjunction between preorders \( C, D \) is a pair of order preserving maps \( F : C \rightarrow D, G : D \rightarrow C \) such that \( F(c) \leq d \Rightarrow c \leq G(d) \). By properties of adjunctions a pair of order preserving maps forms a Galois adjunction iff both \( c \leq GF(c) \) and \( FG(d) \leq d \); and in a Galois adjunction \( FG = F \) and \( GFG = G \). These facts are also readily proved directly.

The left adjoint \( F \) to a given order preserving map \( G \) is unique if it exists; this follows since any two such are naturally equivalent. Dually the right adjoint is unique. A direct proof is as follows. Recall that in a preorder, \( x \leq c \) denotes \( \{ w : w \leq x \} \), and dually for \( x \geq c \). The defining property for a Galois adjunction is easily seen to be equivalent to \( F^{-1}[c \leq] = G(c) \leq \), and to \( G^{-1}[d \geq] = F(d) \geq \). Thus, if \( G \) is order preserving there is a left adjoint \( F \) iff for all \( d \in D \) \( G^{-1}[d \geq] = c \geq \) for some \( c \), in which case \( F(d) = c \). Dually, if \( F \) is order preserving there is a right adjoint \( G \) iff for all \( c \in C \) \( F^{-1}[c \leq] = d \leq \) for some \( d \), in which case \( G(c) = d \).

In a preorder \( C \) a product of a subset \( S \) is just an infimum; dually a coproduct is a supremum. By theorem 13 the right adjoint \( G \) of a Galois adjunction preserves infs; dually a left adjoint preserves supers. For a direct proof, let \( s \) be the infimum of \( S \). Certainly \( G(s) \) is a lower bound for \( G[S] \). On the other hand given any other lower bound \( t \), for all \( x \in S \) \( t \leq G(x) \), whence \( F(t) \leq x \); so \( F(t) \) is a lower bound for \( S \). But then \( F(t) \leq s \), so \( t \leq G(s) \).

The following are readily verified, for a Galois adjunction \( F : C \rightarrow D, G : D \rightarrow C \).
- For \( c \in C, c \in G[D] \) iff \( c = GF(c) \); and for \( d \in D, d \in F[C] \) iff \( d = FG(d) \).
- \( F \) is injective (strictly monotone) on \( G[D] \); and \( G \) is injective on \( F[C] \).
- \( G \) is surjective iff \( GF = \iota_C \) iff \( F \) is injective.
- \( F \) is surjective iff \( FG = \iota_D \) iff \( G \) is injective.

The map \( c \mapsto GF(c) \) satisfies axioms for a closure operator in a partial order, namely monotonicity, idempotence, and \( c \leq GF(c) \). Elements \( c \) such that \( c = GF(c) \) (that is, elements in \( G[D] \)) are called closed. The map \( d \mapsto FG(d) \) is called a kernel operator; there does not seem to be a standard name for the sets with \( d = FG(d) \). We omit further discussion of this topic, except to note that the adjunction establishes a bijection between \( G[D] \) and \( F[C] \).

The name “Galois adjunction” arises from the example of the automorphism group of a field. More generally, if \( M \) is a monoid acting on a set \( S \), let \( C \) be the subsets of \( S \) ordered by inclusion, and let \( D \) be the submonoids of \( M \) ordered by opposite inclusion. For \( T \subseteq S \) let \( F(T) = \{ \nu \in M : \nu(x) = x, \text{ all } x \in T \} \); and for \( N \subseteq M \) a submonoid let \( G(N) = \{ x \in S : \nu(x) = x, \text{ all } \nu \in N \} \). Then \( F : C \rightarrow D, G : D \rightarrow C \), and

\[
F(T) \supseteq N \text{ iff } \forall x \in T \forall \nu \in N(\nu(x) = x) \text{ iff } T \subseteq G(N).
\]

In the case of the automorphism group of a field, \( C \) may be taken as the collection of subfields.

For another example of a Galois adjunction, suppose \( f : X \rightarrow Y \) is a function between sets. Let \( F \) act on the subsets of \( X \) by \( A \mapsto f[A] \), and \( G \) on the subsets of \( Y \) by \( B \mapsto f^{-1}[B] \). Ordering subsets by inclusion,
this is readily seen to be a Galois adjunction, and familiar facts such as $A \subseteq f^{-1}[f[A]]$ and $f[f^{-1}[B]] \subseteq B$ follow. In this case the right adjoint $B \mapsto f^{-1}[B]$ preserves ifs as well as sups.

9. Normals and conormals. For arrows $h_1, h_2$ with the same codomain define $h_2 \leq h_1$ if $h_2 = h_1 \alpha$ for some $\alpha$. This relation is a preorder; as usual, $h_1 \equiv_c h_2$ iff $h_1 \leq_c h_2$ and $h_2 \leq_c h_1$. If $h_2$ is monic then $\alpha$ is unique. If $h_1$ is also monic and $\alpha h_1 = \alpha h_2$ then $h_1 h_2 = h_2 \alpha h_1 = h_2 \alpha h_2 = h_1 \beta$, whence $\beta = \beta_2$; that is, $\alpha$ is monic. If $h_1 \equiv_e h_2$ where $h_1, h_2$ are monic, say $h_1 = h_2 \alpha$ and $h_2 = h_1 \beta$, then $\alpha, \beta$ are readily seen to be isomorphisms; the monics may be said to be isomorphic.

For arrows $h_1, h_2$ with the same domain say that $h_2 \leq_d h_1$ if $h_1 = \alpha h_2$ for some $\alpha$. This relation is a preorder; if $h_2$ is epic then $\alpha$ is unique; if $h_1$ is also epic then $\alpha$ is epic, and equivalent epics are isomorphic. We will write $m_1 \equiv m_2$ or $e_1 \equiv e_2$ for isomorphism of monics or epics.

For the remainder of the section suppose the category has a zero object and kernels and cokernels. Let Ker$(f)$ denote a chosen kernel map, and Coker$(f)$ a chosen cokernel map. This conflicts somewhat with earlier uses of the notation, but the meaning is usually clear from context and will be stated explicitly when necessary. It is readily verified that

- $f_1 \leq_d f_2$ implies Ker$(f_1) \leq_c$ Ker$(f_2);$
- $f_1 \leq_c f_2$ implies Coker$(f_1) \leq_d$ Coker$(f_2);$ and
- $g \leq_c$ Ker$(f)$ iff $fg = 0$ iff Coker$(g) \leq_d f.$

Thus, Coker and Ker are the left and right adjoints of a Galois adjunction. In particular

- Ker$(\text{Ker}(\text{Ker}(f))) \equiv \text{Ker}(f)$ and Coker$(\text{Ker}(\text{Coker}(f))) \equiv \text{Coker}(f).$

Call a monic normal iff it is a kernel of some arrow, and an epic conormal iff it is a cokernel of some arrow. By remarks in section 8, an arrow $f$ is normal (resp. conormal) iff $f = \text{Ker}$(Ker$(f)$) (resp. $f = \text{Coker}$(Ker$(f)$)). In an arbitrary category, monics need not be normal. For example, in Grp the inclusion of a subgroup is normal iff the subgroup is a normal subgroup. In an important class of categories (to be defined in section 12), monics are always normal; Ab and Mod$_R$ are basic examples. In these categories, also epics are conormal; this is true of Grp as well.

If $m$ is monic then Ker$(mf) \equiv \text{Ker}(f).$ Indeed, $mf /\text{Ker}(mf) = 0,$ so $f /\text{Ker}(mf) = 0.$ Also, if $fh = 0$ then $mh = 0,$ so $h = \text{Ker}(mf) \alpha$ for some $\alpha,$ which is unique since Ker$(mf)$ is monic. It is readily verified that Ker$(m) = 0$ if $m$ is monic; Ker$(0) \equiv \iota$ (the identity of the domain); and any arrow out of a terminal object is monic. Dually if $e$ is epic then Coker$(fe) \equiv \text{Coker}(f)$; Coker$(e) = 0$ if $e$ is epic; Coker$(0) \equiv \iota$ (the identity of the codomain); and any arrow into an initial object is epic.

10. Images and coimages. Given an arrow $f : a \mapsto b$ a monic $m : c \mapsto b$ is said to be an image for $f$ if there is a $g$ with $f = mg,$ and whenever $f = m'g'$ with $m'$ monic there is a $j$ with $m = m'j.$ Since $m'$ is monic $j$ is unique, and $g' = jg$ also. One verifies as usual that if $m'$ is also an image then $j$ is an isomorphism. One also verifies that if $f$ is monic then $f$ is an image for $f.$

The dual of an image is called a coimage. That is, given an arrow $f : a \mapsto b$ an epic $e : a \mapsto c$ is a coimage for $f$ iff there is a $g$ with $ge = f;$ and whenever $f = g'e'$ with $e'$ epic, there is a $j$ with $je' = c.$ If $e'$ is also a coimage then $j$ is an isomorphism. If $f$ is epic then $f$ is a coimage for $f.$

Lemma 17. Suppose the category $C$ has equalizers and coequalizers. If $m$ is an image for $f$ and $f = mg$ then $g$ is epic. Dually if $e$ is a coimage for $f$ and $f = ge$ then $g$ is monic.

Proof: Suppose $h_1 g = h_2 g.$ Let $q$ be an equalizer for $h_1, h_2;$ then there is a (unique) $t$ such that $q = qt.$ Then $f = mqt,$ so (since $q$ is monic) $m = mqj$ for some $j,$ and so $qj = \iota.$ It follows that $h_1 = h_2.$

If $f = me$ where $m$ is an image and $e$ a coimage, we call $me$ a coimage-image factorization. Suppose $f = me$ is a coimage-image factorization. It is readily seen that if $f = m'k'$ where $m'$ is an image then $k'$ is a coimage, and dually if $f = k'e'$ where $e'$ is a coimage then $k'$ is an image. Further if $f = m'e'$ where $m'$ is
monic and \( e' \) is epic then \( m'e' \) is a coimage-image factorization. Indeed, \( m = m'\mu \) and \( e = \eta e' \) for some \( \mu, \eta \), whence \( m'e' = me = m'\mu \eta e' \), whence \( \mu \eta = \iota \). Thus if \( f = m_1k_1 \) where \( m_1 \) is monic then \( m = m_1j \) for some \( j \), and \( m' = m_1j\eta \), showing that \( m' \) is an image and proving the claim.

In Set the inclusion of \( f[a] \) in \( b \) is an image for \( f \), and the projection of \( a \) onto \( a/\equiv \) is a coimage, where \( x \equiv y \iff f(x) = f(y) \). Since \( a/\equiv \) is isomorphic to \( f[a] \) the inclusion of \( f[a] \) in \( b \) yields a coimage-image factorization in Set. The forgetful functor from \( \text{Struct}_T \) determines coimage-image factorizations in the usual sense, and this also holds for \( \text{Mdl}_T \) for \( T \) a set of open axioms. These claims are left to the reader. In such categories, the set \( f[a] \) is also called the image of \( f \). Likewise the set \( f^{-1}(0) \) is often called the kernel in categories where the inclusion of this set is the kernel map.

In a category \( C \) which has has coimage-image factorizations we let \( \text{Im}(f) \) denote a chosen image for \( f \), and \( \text{Coim}(f) \) a chosen coimage. We do not assume that \( \text{Coim}(f)\text{Im}(f) = f \); indeed usually this is not so. In \( \text{Grp} \), for example, \( f = mje \) where \( e \) is the canonical epimorphism, \( j \) the canonical isomorphism, and \( m \) the inclusion of the range in the codomain. The map \( je \) is the corestriction of \( f \), and is a coimage.

Call a category an exact sequence category if
- it has finite limits and colimits, a zero object, and coimage-image factorizations;
- \( f \) is monic when \( \ker(f) = 0 \);
- \( f \) is epic when \( \text{coker}(f) = 0 \).

Examples include \( \text{Grp} \) and \( \text{Mod}_R \). The term “exact category” would be less cumbersome, but this term is already in use.

For the remainder of this section suppose \( C \) is an exact sequence category. We claim that
- \( \ker(\text{coim}(f)) \equiv \ker(f) \), \( \text{coim}(f)\ker(f) = 0 \), \( \text{coim}(f) \geq_d \text{coker}(\ker(f)) \),

and \( \text{coim}(f) \equiv \text{coker}(\ker(f)) \iff \text{coim}(f) \) is conormal.

Indeed, \( f = m\text{coim}(f) \) where \( m \) is monic and the first claim follows. Also, \( m\text{coim}(f)\ker(f) = 0 \) and the second claim follows, and the third claim follows from this. If \( \text{coim}(f) \equiv \text{coker}(\ker(f)) \) then \( \text{coim}(f) \) is conormal by definition; and if \( \text{coim}(f) \) is conormal then \( \text{coim}(f) \equiv \text{coker}(\text{coim}(f)) \equiv \text{coker}(f) \).

Dually
- \( \text{coker}(\text{im}(f)) \equiv \text{coker}(f) \), \( \text{coker}(f)\text{im}(f) = 0 \), \( \text{im}(f) \leq_c \text{ker}(\text{coker}(f)) \),

and \( \text{im}(f) \equiv \text{ker}(\text{coker}(f)) \iff \text{im}(f) \) is normal.

Finally, \( \text{im}(0) = \text{coim}(0) = 0 \).

In common concrete categories, including \( \text{Grp} \), definitions might be such that various equivalences are equalities. Examples include \( \ker(mf) = \ker(f) \) for \( m \) monic, \( \text{im}(f) = \text{coker}(\ker(f)) \) if \( \text{im}(f) \) is conormal, and \( \ker(0) = \iota \); and the dual equivalences.

Say that a composable pair of arrows \( f : a \rightarrow b, g : b \rightarrow c \) is exact at \( b \) if \( \ker(g) \equiv \text{im}(f) \) and \( \text{coker}(f) \equiv \text{coim}(g) \). Again, in \( \text{Grp} \) and similar concrete categories \( f[a] \) may be required to equal \( g^{-1}[0] \), so that “the image of \( f \) equals the kernel of \( g \)”.

The terminology in use is usually clear, but if necessary will be stated.

Suppose \( \ker(g) \equiv \text{im}(f) \); then \( \text{coker}(f) \equiv \text{coker}(\text{im}(f)) \equiv \text{coker}(\ker(g)) \), so \( \text{coker}(f) \equiv \text{coim}(g) \) iff \( \text{coim}(g) \) is conormal. Dually if \( \text{coker}(f) \equiv \text{coim}(g) \) then \( \ker(g) \equiv \text{im}(f) \) iff \( \text{im}(f) \) is normal. In particular, a pair of arrows as above is exact at \( b \) iff \( \ker(g) \equiv \text{im}(f) \), \( \text{coker}(f) \equiv \text{coim}(g) \), or either, in categories where epis are conormal, monics are normal, or both respectively.

In \( \text{Grp} \) \( \text{coim}(g) \equiv \text{coker}(f) \) always holds, but \( \text{im}(f) \equiv \ker(g) \) need not. Let \( b \) be a group with a subgroup \( a \) which is not normal (see chapter 14) and let \( f \) the inclusion.

Suppose \( I \subseteq Z \) is an interval, \( \langle f_i : a_i \rightarrow a_{i+1} : i, i + 1 \in I \rangle \) a composable sequence of arrows, and \( i, i + 1 \in I \). The sequence is said to be exact at \( a_i \) if the pair \( f_i, f_{i+1} \) is; the sequence is said to be exact sequence exact if it is exact at each \( a_i \).
Endpoints of an exact sequence, if any, are frequently 0. For future reference, we consider the following types of exact sequences.

\[0 \rightarrow a \xrightarrow{f} b \xrightarrow{g} c \rightarrow 0\]  \hspace{1cm} (I)

\[0 \rightarrow a \xrightarrow{f} b \xrightarrow{g} c\]  \hspace{1cm} (II)

\[a \xrightarrow{f} b \xrightarrow{g} c \rightarrow 0\]  \hspace{1cm} (III)

\[a \xrightarrow{f} b \xrightarrow{g} c\]  \hspace{1cm} (IV)

One verifies that a sequence of type II is exact iff \(g\) is conormal and \(f \equiv \ker(g)\). Dually a type III sequence is exact iff \(f\) is normal and \(g \equiv \ker(f)\). An exact sequence of type I is called a short exact sequence; it is exact iff \(f \equiv \ker(g)\) and \(g \equiv \coker(f)\). Such arises, for example, when \(f\) is the inclusion of a normal subgroup and \(g\) the canonical epimorphism onto the quotient. More generally if \(f\) is a normal epimorphism, \(g \equiv \coker(f)\), and \(g\) is conormal then the sequence is exact. Given a short exact sequence, \(c\) is called the quotient of \(b\) over \(a\), and denoted \(b/a\).

11. Biproducts. Suppose \(a' = \bigoplus a'_j\), with injections \(m'_j\); \(a = \times_i a_i\), with projections \(p_i\); and \(f_{ij}: a'_j \rightarrow a_i\) are given maps. There is a unique map \(f_j: a'_j \rightarrow a\) with \(p_i f_j = f_{ij}\) for all \(i\). There is a unique map \(f: a' \rightarrow a\) with \(f_m' = f_j\), or \(p_i f m'_j = f_{ij}\) for all \(j\). Dually there is a unique map \(f'_j: a' \rightarrow a_i\) with \(f'_{im} = f_{ij}\) for all \(i\); and it is readily verified that \(p_i f = f_{ij}\) for all \(i\). The given system of maps can be thought of as the entries of a matrix, where \(f_{ij}\) is in row \(i\) and column \(j\) if vectors are written as column vectors. Each such matrix determines a unique map from \(a'\) to \(a\).

If the category \(C\) has a zero object, and \(a'_i = a_i\) for all \(i \in I\), let \(f_{ij} = \delta_{ij}\) where \(\delta_{ii} = \iota_{a_i}\) and \(\delta_{ij} = 0\) for \(i \neq j\). Write \(m_i\) for the induced map from \(a_i\) to \(a\); \(p'_j\) for the induced map from \(a'\) to \(a_j\); and \(n\) for the induced map from \(a'\) to \(a\).

Lemma 18. With notation as above, suppose \(C\) is an exact sequence category, and \(I = \{1, 2\}\). Then

\[m_2 = \ker(p_1)\text{ and } m_1 = \ker(p_2); \text{ dually } p'_2 = \coker(m'_1)\text{ and } p'_1 = \coker(m'_2)\]

Proof: Certainly \(p_1 m_2 = 0\); suppose \(p_1 k = 0\). Then \(p_1 m_2 p_2 k = 0 = p_1 k\), and \(p_2 m_2 p_2 k = p_2 k\), so \(m_2 p_2 k = k\). This shows \(m_2 = \ker(p_1)\), and the rest of the claim follows similarly.

Again in specific categories the equivalences of the lemma are actually equalities. For \(I\) finite, if \(n\) is an isomorphism \(a\) is said to be a biproduct. If \(\times_i a_i\) is a biproduct then it is a coproduct, with injections \(m_i\). Indeed, given \(h_i: a_i \rightarrow c\) there is a \(k: c \rightarrow \bigoplus_a a_i\) such that \(kh_i = m'_i\); then \(nk: c \rightarrow \times_i a_i\) and \(nkh_i = nm'_i = m_i\). Clearly, \(p_i m_j = \delta_{ij}\). We leave it to the reader to show that if all products of pairs of objects exist and are are biproducts, then this is so for nonempty finite sets of objects.

12. Abelian categories. Call a category Abelian if it has finite limits, finite colimits, and a zero object; every monic is normal; and every epic is conormal. Ab, for example, is Abelian; Grp is not. For any ring \(R\), \(\text{Mod}_R\) is Abelian. Abelian categories were invented to provide an axiomatic setting for categories having certain properties in common with \(\text{Mod}_R\). Clearly an exact sequence category with all monics normal and all epis conormal is Abelian; the converse will be shown in lemma 20.

Lemma 19. An Abelian category is balanced.

Proof: If \(f\) is monic then \(f = \ker(\text{Coker}(f))\); if \(f\) is also epic then \(\text{Coker}(f) = 0\), so \(f = \ker(0) = \iota\) and \(f\) is an isomorphism.

Lemma 20. An Abelian category is exact, and \(\text{Im}(f) = \ker(\text{Coker}(f))\) and \(\text{Coim}(f) = \text{Coker}(\ker(f))\).
that is monic it equals Ker(\(f\)) for some \(p\). Now, \(pf = pmqk = 0\), so \(p = wCoker(f)\) for some \(w\). Then \(pm = wCoker(f)m = 0\), so \(m = mqt\) for some \(t\). Since \(m\) is monic \(qt = \iota\), and since \(rq = sq\) \(r = s\). We now claim that \(m\) is an image. Suppose \(f = m\'k\) where \(m'\) is monic. Choose \(r\) with \(m' = Ker(r)\). Then \(rm' = 0\), so \(rme = rm'k' = 0\), so \(rm = 0\), so there is a \(\mu\) with \(m = m'\mu\). Dually, \(Coker(Ker(f))\) is a coimage for \(f\). Suppose now that \(f = mk\) where \(m\) is an image, and \(f = le\) where \(e\) is a coimage. Then \(k\) is epic, whence there is a unique map \(\eta\) such that \(\eta k = e\). Likewise \(l\) is monic and there is a unique map \(\mu\) such that \(l\mu = m\); further \(\mu k = e\). Since \(\eta\) is unique \(\mu = \eta\); since \(\eta\) is epic and \(\mu\) is monic, \(\eta\) is an isomorphism. Thus, \(f\) has a coimage-image factorization. Now, if \(Ker(f) = 0\) then \(Coi(f) = \iota\), so \(f = Im(f)\) and \(f\) is monic. Dually if \(Coker(f) = 0\) then \(f\) is epic.

**Lemma 21.** In an Abelian category, finite products are biproducts, and dually for finite coproducts.

**Proof:** It follows from lemma 18 that \(p_1 \equiv Coker(m_2)\) and so forth if \(C\) is Abelian. Now let \(k = Ker(n)\); then \(p_2k = p_1mk = 0\). Since \(m_2^2 \equiv Ker(p'_1), k = m'_1h\) for some \(h\). We then have \(h = \iota_a h = p'_1m'_1h = p'_1k = 0\), whence \(k = 0\). Thus, \(n\) is monic; dually it is epic. Hence it is an isomorphism.

A partial function \(+ : Ar \times Ar \rightarrow Ar\) is called an addition if
1. \(f + g\) is defined iff \(f, g\) have the same domains and codomains;
2. restricted to \(Hom(a, b)\) \(+\) is the addition of a commutative group; and
3. \(h(f + g) = hf + hg\) and \((f + g)h = fh + gh\) whenever the compositions are defined.

**Theorem 22.** An Abelian category has a unique addition defined on it. The zero map is the additive identity.

**Proof:** Call \(+\) a weak addition if it satisfies condition 1 above and makes each \(Hom(a, b)\) a commutative monoid. First we show that there is a weak addition. Suppose \(f_1, f_2 : a \rightarrow b\). Let \(\delta : a \rightarrow a \times a\) be the "diagonal" map, induced by the projections \(\iota : a \rightarrow a\). Let \([f_1, f_2]\) denote the map from \(a \times a\) to \(b\) determined by the indicated matrix. Let \(f_1 + f_2 = [f_1, f_2]\delta; \) we claim this is the unique addition. Let \(\delta : b \times b \rightarrow b\) be the "coidiagonal" map, induced (since \(b \times b\) is a coproduct) by the injections \(\iota : b \rightarrow b\). Let \([f_1, f_2]^t\) (using the transpose to simplify the notation) denote the map from \(a\) to \(b \times b\) induced by this matrix. Let \(f_1 +^t f_2 = \delta^t[f_1, f_2]^t\). By uniqueness \(g[f_1, f_2] = [gf_1, gf_2]\), so \(g(f_1 + f_2) = gf_1 + gf_2\). Dually \((f_1 +^t f_2)g = f_1g +^t f_2g\). Now,

\[
\begin{bmatrix}
[f_{11} & f_{21} \\
f_{12} & f_{22}
\end{bmatrix} = \begin{bmatrix}
f_{11} & f_{21} \\
f_{12} & f_{22}
\end{bmatrix} = \begin{bmatrix}
f_{11} \\
f_{12}
\end{bmatrix} \begin{bmatrix}
f_{21} \\
f_{22}
\end{bmatrix},
\]

as is readily verified. It follows (using (3)) that

\[
(f_{11} + f_{21}) +^t (f_{12} + f_{22}) = (f_{11} +^t f_{12}) + (f_{21} +^t f_{22}).
\]

Also, \([f, 0] = f p_1\), since \(f p_1 m_1 = f_1, f p_1 m_2 = 0\); hence \(f + 0 = f p_1 \delta = f \iota = f\). Similarly \(0 + f = f\), and dually \(f +^t 0 = f, 0 +^t f = f\). Letting \(f_{12} = f_{21} = 0\) in (4), it follows that \(+\) and \(+^t\) coincide. Letting \(f_{21} = 0\) then yields associativity, and letting \(f_{11} = f_{22} = 0\) yields commutativity. Next we show that, given a weak addition and a finite product \(a = \times a_i, \sum_i m_i p_i = 1\), this follows since the left side composed with \(m_i\) equals \(m_i\), for all \(i\). From this is follows that composition of maps determined by (finite) matrices obeys matrix multiplication. Indeed

\[
(g f)_{ij} = p_i g f m_j = p_i g \left(\sum_k m_k p_k\right) f m_j = \sum_k g_{ik} f_{kj}.
\]
Uniqueness now follows, observing that \( \delta = [ \iota, \iota]^t \). It remains to find an additive inverse; for this it suffices to find one for \( \iota \), since \(-f\iota = f(-\iota)\) by the distributive law. Consider the map \( \theta \) given by the matrix

\[
\begin{bmatrix}
\iota & \iota \\
0 & \iota
\end{bmatrix}.
\]

If \( \alpha \theta = 0 \) then, expressing \( \alpha \) as a matrix, it follows that \( \alpha = 0 \); thus \( \text{Coker}(\theta) = 0 \) and \( \theta \) is epic. Similarly \( \theta \) is monic, hence an isomorphism. Write

\[
\theta^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix};
\]

\( \theta \theta^{-1} \) is the identity matrix, and \( \iota_b = 0 \) follows.

Useful consequences of this theorem include the following.

- The equalizer of \( f_1 \) and \( f_2 \) is the kernel of \( f_1 - f_2 \); dually the coequalizer is the cokernel.
- \( f \) is monic iff \( h = 0 \) whenever \( fh = 0 \); dually \( f \) is epic iff \( h = 0 \) whenever \( hf = 0 \).

Abelian categories are often given together with a “natural” addition. For example, in \( \text{Mod}_R + \) is pointwise addition of functions.

The hypotheses for an Abelian category can be weakened, requiring only 0, kernels, cokernels, products, coproducts, normal monics, and conormal epics. It suffices to show that \( C \) has equalizers and coequalizers, so suppose \( f, g : a \to b \). Note that \( f' = [\iota, f]^t \) is monic, so equal to \( \text{Ker}(q) \) for some \( q \). Let \( g' = [\iota, g]^t \), and let \( m = \text{Ker}(gg') \). Then \( gg'm = 0 \), so \( g'm = f'n \) for some \( n \); since the first components are \( m, n \) we have \( n = m \), whence from the second components \( gm = fm \). If \( gm' = fm' \) for some \( m' \), then \( g'm' = f'n' \), so \( gg'm' = gf'm' = 0 \), so \( m' = mt \) for some \( t \). Thus, \( m \) is an equalizer for \( f, g \); dually \( C \) has coequalizers.

**Lemma 23.** In an Abelian category, suppose \( m_i : a_i \to a, p_i : a \to a_i, p_j m_i = \delta_{ij} \) for \( 1 \leq i, j \leq n \). Then \( a \) is a biproduct of the \( a_i \) (with injections \( m_i \) and projections \( p_i \)) iff \( \sum m_i p_i = \iota_a \). When \( n = 2 \) the conditions \( p_1 m_1 = \iota, m_2 = \text{Ker}(p_1), m_1 p_1 + m_2 p_2 = \iota \) suffice; further when the first two hold \( p_2 \) exists.

**Proof:** \((\sum_i m_i p_i) m_i = \iota m_i \), so if \( a \) is a coproduct then \((\sum_i m_i p_i) = \iota \). Conversely, given \( f_i : a_i \to b \) let \( f = \sum_i f_i p_i \); then \( f m_i = f_i \) for all \( i \), and is the unique such arrow since if \( f' m_i = f_i \) for all \( i \) then \( f' = f' \sum_i m_i p_i = \sum_i f_i p_i = f \). Thus, \( a \) is a coproduct. For the second claim, let \( \theta = m_1 p_1 \). Then \( m_2 = \text{Ker}(\theta), \theta^2 = \theta, \) and \( \theta(\iota - \theta) = 0 \), so \( \iota - \theta = m_2 p_2 \) for some \( p_2 \), which now we assume given. Then \( m_2 p_2 m_2 = (\iota - m_1 p_1) m_2 = m_2 \), so \( p_2 m_2 = \iota \). Also \( m_1 = m_1 p_1 m_1 + m_2 p_2 m_1 = m_1 + m_2 p_2 m_1 \), so \( m_2 p_2 m_1 = 0 \), so \( p_2 m_1 = 0 \).

It is readily seen that a functor between any categories which preserves finite products and finite coproducts acts on finite matrices “componentwise”. It follows that a functor between Abelian categories which preserves finite products (equivalently finite coproducts) preserves addition. A functor with the latter property is called additive.

A functor between Abelian categories is called

- left exact if it preserves kernels;
- right exact if it preserves cokernels; and
- exact if it preserves both kernels and cokernels.

It follows using Lemma 23 that a left exact functor preserves finite products, hence is additive, hence preserves equalizers since \( \text{Eq}(f, g) = \text{Ker}(f - g) \), hence preserves finite limits; it also clearly preserves monics. Dually a right exact functor preserves finite colimits and epics. Finally, an exact functor preserves images and coinages.

Recalling the types of exact sequences defined in section 10, a functor between Abelian categories is clearly left exact iff it preserves exactness of type II sequences. A fortiori if a functor is left exact it takes a type I exact sequence to a type II exact sequence.
Conversely suppose \( F \) takes type I exact sequences to type II exact sequences, and

\[
\begin{align*}
0 \to a & \overset{f}{\to} b \overset{g}{\to} c
\end{align*}
\]

is exact. Let \( g = me \) be the coimage-image factorization, where \( e : c \to d \). The sequence

\[
\begin{align*}
0 \to d & \overset{m}{\to} c
\end{align*}
\]

can be completed to a short exact sequence; applying \( F \) shows that \( F(m) \) is monic. Thus, \( \text{Ker}(F(g)) = \text{Ker}(F(me)) = \text{Ker}(F(m)F(e)) = \text{Ker}(F(e)) \). Applying \( F \) to

\[
\begin{align*}
0 \to a & \overset{f}{\to} b \overset{e}{\to} d \overset{0}{\to}
\end{align*}
\]

shows that \( \text{Ker}(F(e)) = F(f) \). Thus, \( \text{Ker}(F(g)) = F(f) \), showing that \( F \) is left exact.

Dually a functor is right exact iff it preserves exactness of type III sequences, iff it takes type I exact sequences to type III. If a functor is exact it clearly takes type I exact sequences to type I exact sequences; the converse holds also, since as we have just shown the functor is both left and right exact.

A functor is called half exact (or exact in the middle) if it takes type I exact sequences to type IV. A half exact functor \( F \) preserves biproducts. To see this, in the notation of lemma 23 it suffices to show that \( F(p_1)F(m_1) = \iota; \) and \( F(m_2) = \text{Ker}(F(p_1)) \). The first requirement follows immediately. In particular \( F(m_2) \) is monic, so the second requirement follows by half-exactness. As a corollary, a half exact functor is additive.

A contravariant functor between Abelian categories is called left exact if it maps cokernels to kernels, or equivalently maps type III (or I) sequences to type II sequences; a left exact contravariant functor maps finite colimits to finite limits. A contravariant right exact functor is defined dually, and maps finite limits to finite colimits. A contravariant exact functor is one which is both left and right exact.

**Exercises.**

1. Show that there is a universal arrow from
   - a graph \( c \) to the forgetful functor \( G : \text{Cat} \to \text{Graph}; \)
   - a graph \( c \) to the forgetful functor \( G : \text{Grpd} \to \text{Graph}. \)

Give explicit descriptions of a free object with underlying set \( c \) in
   - \( \text{Mon, Grp, Ab, Rng, CRng} \).

Hint: For \( \text{Cat} \), the objects are those of \( c \) and the arrows are the “composable” sequences \( f_1 \ldots f_k \) of arrows of \( c \). If two sequences are composable, their composition is just the concatenation. For \( \text{Grpd} \), add \( f^{-1} \) for each arrow \( f \) of \( c \), and only allow sequences where neither \( ff^{-1} \) nor \( f^{-1}f \) occur. Composition involves canceling after concatenation. For \( \text{Rng} \), the elements are integer linear combinations of distinct monomials; order \( c \) in the commutative case.

2. Suppose \( C \) is a category with a product \( a \times b \) for each pair \( a, b \) of objects. Show the following.
   a. \( a_1 \times (\cdots \times (a_k \times a_k) \cdots) \) is a product object for any \( a_1, \ldots, a_k \).
   b. \( a \times (b \times c) \) is naturally equivalent to \( (a \times b) \times c \).
   c. If \( C \) has a terminal object \( t \) then \( a \times t \) and \( t \times a \) are naturally equivalent to \( a \).
   d. The dual statements hold for coproducts and initial objects. Hint: Consider the opposite category.

3. In \( \text{Cat} \), show that \( a^{op} \times b^{op} \) is naturally equivalent to \( (a \times b)^{op} \).

4. Show that two natural transformations satisfying the triangular identities determine a unique adjunction with the transformations as unit and counit. Hint: Define \( \alpha_{cd}(h) = G(h)\mu_c, \beta_{cd}(h) = \nu_dF(h) \), and show that \( \alpha \) is a natural equivalence with inverse \( \beta \).

5. Show that nonempty colimits exist in the category of preorders, or partial orders. Hint: The coproduct of a pair is the disjoint union. The coequalizer preorder of \( f, g : a \mapsto b \) is the preorder defined on
the coequalizer set $b/\equiv$ as $x \leq y$ iff there exist $x' \in x$, $y' \in y$ such that $x' \leq y'$. For partial orders, take the usual quotient of the coequalizer preorder.

6. Show that epics are surjective in Grp. Hint: Construct two maps to the group of permutations on the elements of $H$, the left regular representation, and a map taking $x$ to a conjugate of left multiplication by $x$, by a permutation which transposes two right cosets of the image. If the image has index 2 use the fact that it is normal.

1. Basic definitions. As mentioned in chapter 5, a group $G$ acts on itself by conjugation. Indeed if $a \in G$ then the map $x \mapsto a^{-1}xa$ is an automorphism of $G$. The orbit of $x$ under conjugation is called the conjugacy class of $x$, and two elements in the same conjugacy class are called conjugate. Thus, $x$ and $y$ are conjugate if $y = a^{-1}xa$ for some $a \in G$. These notions may be carried over to sets of elements; that is, the conjugates of a subset $S \subseteq G$ are those subsets of $G$ which are in the orbit of $S$ under the action induced on the subsets.

The stabilizer of a point $x$ under conjugation, considered as a subgroup of $G$, is called the normalizer of $x$, and denoted $N(x)$, or $N_G(x)$ to specify the group. That is, $N(x)$ consists of those $a \in G$ such that $a^{-1}xa = x$; being the inverse image of a subgroup (the stabilizer in the group of conjugations) under a homomorphism (the map from an element to its conjugation), $N(x)$ is a subgroup of $G$. For finite $G$, by theorem 5.4 the size of the conjugacy class of $x$ equals $|G|/|N(x)|$. Thus, $|G| = \sum_{x \in R} |G|/|N(x)|$, where $R$ is a system of representatives of the conjugacy classes. This equation is called the class equation of the group; it is used in the proofs of several basic facts, as we shall see.

The (setwise) stabilizer of a subset $S \subseteq G$ under conjugation is called the normalizer of $S$, and denoted $N(S)$ or $N_G(S)$. If $S$ is a subgroup $H$, then $H \triangleleft N(H)$; indeed $N(H)$ is the largest subgroup $K$ of $G$ such that $H \triangleleft K$, and $H \triangleleft G$ iff $G = N(H)$. The group of elements which fix $S$ pointwise, that is, for which $a^{-1}xa = x$ for all $x \in G$, is called the centralizer of $S$ and denoted $C(S)$ or $C_G(S)$; for a single element, its normalizer and its centralizer are the same. Note that the centralizer of the whole group is exactly the center of the group. Also, an element $x$ is in the center iff $N(x) = G$; and a subgroup $H \subseteq G$ is normal iff $N(H) = G$.

If $p$ is a prime number, a finite group is called a $p$-group if its order is $p^e$ for some nonnegative integer $e$. Note that any subgroup of a $p$-group is a $p$-group. If $e = 0$ the group is trivial. If $e = 1$ it is easy to see that the group is the cyclic group of order $p$; there must be an element of order other than 0, so its order is $p$ and it generates the entire group. Using theorem 8.11, there are two commutative groups of order $p^2$, namely $\mathbb{Z}_p \times \mathbb{Z}_p$ and $\mathbb{Z}_{p^2}$. We will see below that there are no noncommutative ones. A group $\mathbb{Z}_p^k$ for $p$ prime and $k \geq 0$, the $k$-dimensional vector space over $\mathbb{F}_p$, is called an elementary commutative group.

**Theorem 1.** If $G$ is a $p$-group then the center $C(G)$ is nontrivial.

**Proof:** In any group, $x \in C(G)$ iff the conjugacy class of $x$ has size 1. For a $p$-group, by the class equation, the number of elements whose class size is 1 must be a multiple of $p$, because each term in the sum is either 1 or divisible by $p$, and the sum is divisible by $p$. Since $C(G)$ is nonempty (it always contains the identity), it must be nontrivial.

**Theorem 2.** If $|G| = p^2$ then $G$ is commutative.

**Proof:** By theorem 1, the center has order $p$ or $p^2$. Supposing it has order $p$, choose $a \in G - C(G)$; clearly $C(G) \subseteq N(a)$ and $a \in N(a)$, whence $N(a) = G$. But this is a contradiction.

**Theorem 3.** If a prime $p$ divides $|G|$ for a finite group $G$ then $G$ contains an element of order $p$.

**Proof:** This is vacuously true if $|G| = 1$. If $p$ divides $|H|$ for any proper subgroup $H \subseteq G$, inductively $H$ contains an element of order $p$, which is an element of order $p$ in $G$. In the remaining case, the size of the conjugacy class of an element not in the center must be divisible by $p$, whence by the class equation $p$ divides the order of the center, and so $C(G) = G$ and the group is commutative. In fact, it is easily seen that it is cyclic of order $p$.

2. Double cosets. Double cosets are a useful technical device, which we will use for example in the following section. If $H, K$ are subgroups of a group $G$ the set $HxK$ for $x \in G$ is called a double coset.
Clearly $HxK$ is the union of those left cosets of $K$ containing the members of $Hx$; and the union of those right cosets of $H$ containing the members of $xK$.

**Lemma 4.** The double cosets $HxK$, $x \in G$, form a partition of $G$.

**Proof:** Suppose $h_1xk_1 = h_2xk_2$; multiplying by $H$ on left and $K$ on right on the right $Hx_1K = Hx_2K$.

**Lemma 5.** In a finite group, with $HxK$.

**Proof:** Immediate from lemmas 4 and 5.

**Lemma 6.** If $H$ is a subgroup of a finite group $G$ then

$$|G|/|H| = \sum_{x \in R} |H|/|H \cap x^{-1}Hx|$$

where $R$ is a system of representatives of the double cosets $HxH$.

**Proof:** Immediate from lemmas 4 and 5.

**Lemma 7.** In a finite group, with $H$ a subgroup, the following are equivalent: $x \in N(H)$, $HxH = xH$, $HxH = Hx$, $|HxH| = |H|$, $|H| = |H \cap x^{-1}Hx|$.

**Proof:** $x \in N(H)$ if $HxH = xH$; certainly if this is so then $HxH = xH$, and conversely if $HxH = xH$ then $Hx \subseteq xH$, and since they are the same size they are equal. Similarly $Hx = xH$ if $HxH = Hx$. Clearly if $HxH = Hx$ then $|HxH| = |H|$: the converse follows because $Hx \subseteq HxH$, so if the sizes are equal the sets are. The last claim follows by lemma 5.

**Lemma 8.** $N(H)$ is a disjoint union of left, or right, cosets of $H$.

**Proof:** $x^{-1}Hx = H$ implies $(xh)^{-1}H(xh) = H$ and $(hx)^{-1}H(hx) = H$.

### 3. Sylow subgroups.

**Theorem 9.** If $G$ is a finite group, $p$ is a prime with $\text{ord}_p(|G|) = e$, and $f \leq e$, then $G$ has a subgroup $H$ of order $p^f$. If further $f < e$ then $G$ has a subgroup $J$ with $|J| = p^{f+1}$ and $H \triangleleft J$.

**Proof:** By theorem 3 the theorem is clear if $f = 0$. Inductively, suppose $H$ has order $p^f$, $f < e$. We claim that $|N(H)|/|H|$ is divisible by $p$; for it follows by lemmas 7 and 8 that $|N(H)|/|H|$ equals the number of 1’s in the sum of lemma 6, and the remaining terms of the sum, and the left side, are divisible by $p$. By theorem 3 $N(H)/H$ contains a subgroup of order $p$. The inverse image of this subgroup under the canonical epimorphism from $N(H)$ to $N(H)/H$ is easily seen to have order $p^{f+1}$; further $H$ is normal in it.

A subgroup $H$ of a finite group $G$ is called a $p$-subgroup if it is a $p$-group. The theorem implies that a maximal $p$-subgroup, i.e., a $p$-subgroup not contained in any larger $p$-subgroup, has order $p^e$ where $e = \text{ord}_p(|G|)$. Such $p$-subgroups are called Sylow $p$-subgroups. Note further that every $p$-subgroup $J$ is a subgroup of a Sylow $p$-subgroup $H$, and if $|J| = p^{e-1}$ then there is such an $H$ with $J \triangleleft H$.

As an example of further consequences of these methods, suppose $p$ is a prime, $\text{ord}_p(|G|) = e$, and $|G|/p^e < p$ (for example, suppose $|G| = 20$ and $p = 5$). Then any Sylow $p$-subgroup $H$ is normal; for by lemma 6, $|H|$ must equal $|H \cap x^{-1}Hx|$ for all $x$, so $x \in N(H)$ for all $x$. 

150
Theorem 12. Let $G$ be a finite group and $H$ a Sylow $p$-subgroup.


b. The number of Sylow $p$-subgroups is $|G|/|N(H)|$.

c. $H$ is normal iff it contains every element whose order is a power of $p$, iff there is only one Sylow $p$-subgroup.

d. If $K \subseteq G$ is a $p$-subgroup and $K \subseteq N(H)$ then $K \subseteq H$.

e. The number of Sylow $p$-subgroups is congruent to 1 mod $p$.

f. If $J \subseteq G$ is a subgroup and $N(H) \subseteq J$ then $N(J) = J$.

Proof: Let $H, K$ be two Sylow $p$-subgroups; by lemmas 4 and 5 $|G|/|H| = \sum_{x \in R} |K|/|K \cap x^{-1}Hx|$. The left side is not divisible by $p$, and every term on the right side is either 1 or divisible by $p$, so some term must equal 1. For that term, $K = x^{-1}Hx$; this proves part a. Part b follows by theorem 5.4. For part c, clearly a Sylow $p$-subgroup is normal iff it is the only Sylow $p$-subgroup, in which case by theorem 9 it contains every element of order a power of $p$. Since $H$ is normal in $N(H)$ it is the only Sylow $p$-subgroup of $N(H)$, and part d follows. For part e, let $H, K$ be two distinct Sylow $p$-subgroups. By part d $K$ is not a subgroup of $N(H)$, so $N(H) \cap K$ is a proper subgroup of $K$. The size of the orbit of $H$ under conjugation by elements of $K$ is thus divisible by $p$, since it equals $|K|/|N(H) \cap K|$. Under conjugation by elements of $K$, the orbit of $K$ has size 1, and the other orbits have size divisible by $p$, proving the claim. For part f, if $x^{-1}Jx = J$ then $x^{-1}Hx$ is a Sylow $p$-subgroup of $J$, so $x^{-1}Hx = w^{-1}Hw$ for some $w \in J$; then $wx^{-1} \in N(H) \subseteq J$, so $x \in J$.

4. Primitivity and regular normal subgroups. The next section benefits from some further discussion of permutation groups. Let $G$ be a transitive permutation group on $S$. A subset $B \subseteq S$ is called a block of imprimitivity if either $gB = B$ or $gB \cap B = \emptyset$ for all $g \in G$. For example, suppose $G$ is a group and $H$ a proper subgroup; the left action of $G$ on itself has the cosets as blocks of imprimitivity.

If $B$ is a block of imprimitivity, we claim that the sets in the orbit of $B$ are the parts of a partition of $S$. Indeed, if $x \in gB \cap hB$ then $g^{-1}x \in B \cap g^{-1}hB$, so $B = g^{-1}hB$ and $gB = hB$. Further if $y \in S$ choose $x \in B$ and $g \in G$ so that $y = gx$; then $y \in gB$. Such a partition of $S$ is called a system of imprimitivity for $G$; $G$ induces an action on the parts or “blocks”. If $S$ is finite the blocks have equal size.

The systems of imprimitivity are ordered by the usual refinement order on partitions. In fact, they are a sublattice; to show this it suffices to show that given a family of systems of imprimitivity, the intersection of the corresponding equivalence relations corresponds to a system of imprimitivity. This follows because $g(\cap_a B_a) = \cap_a gB_a$ where $B_a$ is a family of blocks, one from each system. There is a least system, namely the system where each block consists of a single element, and a greatest, namely where there is a single block. Any system other than these is called nontrivial.

A permutation group is called primitive if it is transitive and has no nontrivial blocks of imprimitivity. For example, if the degree is prime then a transitive group is primitive.

Theorem 11. A transitive finite permutation group $G$ is primitive iff the stabilizer $\text{Stab}_x$ of an element is a maximal proper subgroup.

Proof: If $B$ is a nontrivial block of imprimitivity we claim its setwise stabilizer $\text{Stab}_B$ is a subgroup with $\text{Stab}_B \subseteq \text{Stab}_B \subseteq G$. First, $g \in \text{Stab}_B$ implies $gB = B$ since $x \in B \cap gB$; this shows $\text{Stab}_B \subseteq \text{Stab}_B$. Second, choose $x \in B$, $y \in B - \{x\}$, and $g \in G$ with $y = gx$; then $g \in \text{Stab}_B - \text{Stab}_x$. Third, choose $x \in B$, $y \in S - B$, and $g \in G$ with $y = gx$; then $g \in G - \text{Stab}_B$. Conversely, if $\text{Stab}_B \subset H \subseteq G$ we claim that $Hx$ is a nontrivial block of imprimitivity. Certainly $1 < |Hx|$, and $|Hx| = |H|/|\text{Stab}_x| < |G|/|\text{Stab}_x| = |Gx|$. Now, if $y \in gHx \cap Hx$, say $y = gh_1x = h_2x$, then $h_2^{-1}gh_1 \in \text{Stab}_x$, so $g \in H$ and $gHx = Hx$.

Theorem 12. If a finite permutation group $G$ is 2-homogeneous it is primitive.
Proof: Suppose $G$ acts on $S$ and $B \subseteq S$ is a block of imprimitivity with $|B| > 1$, say $x, y \in B$. Given $z \in S$, let $g\{x,y\} = \{x,z\}$; then $gB \cap B \neq \emptyset$, so $gB = B$ and $z \in B$. Since $z$ was arbitrary $B = S$.

Lemma 13. For finite $G$, if $G$ is transitive and $N \triangleleft G$ then the orbits of $N$ are a system of imprimitivity for $G$. In particular $N$ is half transitive, and if $G$ is primitive and $N$ nontrivial then $N$ is transitive.

Proof: It suffices to observe that if $Nx$ is an orbit of $N$ then so is $gNx$, $g \in G$, since $gNx = gN g^{-1} gx = Ngx$.

A group is called simple if it has no nontrivial proper normal subgroups.

Theorem 14. Let $G$ be primitive. If the stabilizer $\text{Stab}_x$ of an element is simple then either $G$ has a regular normal subgroup or $G$ is simple.

Proof: Suppose $N$ is a nontrivial normal subgroup of $G$; by lemma 4 $N$ is transitive. Now, $\text{Stab}_x$ is simple and $N \cap \text{Stab}_x \triangleleft \text{Stab}_x$. Thus, either $N \cap \text{Stab}_x$ is trivial, in which case $N$ is regular; or $N \cap \text{Stab}_x = \text{Stab}_x$, whence $\text{Stab}_x \subseteq N$ and so $N = G$ since $N$ is transitive. Thus, unless $G$ contains a regular normal subgroup $G$ is simple.

The existence of a regular normal subgroup $N$ has strong consequences for a finite permutation group $G$. For one thing, the extension of $N$ by $G$ splits. That is, there is a subgroup $H \subseteq G$ which forms a system of coset representatives; this is easily seen to be so iff $N \cap H$ is trivial. Indeed, we may choose $H = \text{Stab}_x$. For another, we may identify the set $S$ on which $G$ acts with $N$, by the usual correspondence $g \mapsto gx$ for some fixed $x \in S$. This correspondence induces an action of $\text{Stab}_x$ on $N - \{e\}$, namely $h \mapsto ghg^{-1}$, $g \in \text{Stab}_x$, $h \in N - \{e\}$, since $ghg^{-1}x = ghx$; this action is permutation group isomorphic to the action of $\text{Stab}_x$ on $S - \{x\}$.

5. Simple groups. An obvious question is whether there is a subgroup of order $d$ for every $d$ dividing $|G|$. The answer is no; to show this, we first prove a useful fact.

Theorem 15. If $H$ is a subgroup of index 2 in a group $G$ then $H \triangleleft G$.

Proof: If $x \in H$ then $xH = Hx = H$. Otherwise, $x \in G - H$, and $xH = G - H = Hx$, since every left or right coset is either $H$ or $G - H$.

If we could find a simple group of even order greater than 2, it would have no subgroups of index 2. The cyclic groups of prime order are simple, but except for $\mathbb{Z}_2$ these have odd order. It is a deep fact of group theory that these are the only finite simple groups of odd order. Let $A_n$ denote the subgroup of $S_n$ consisting of the even permutations; this is clearly a subgroup. We will show that for $n \geq 5$ $A_n$ is a simple group.

A subgroup $H$ of a group $G$ is called a characteristic subgroup if it is mapped to itself by every automorphism of $G$. If $H$ is a characteristic subgroup it is a normal subgroup, but the converse does not necessarily follow and in fact is false. For example, there are automorphisms of an elementary commutative group which move subgroups. The center of a group $G$ is an example of a characteristic subgroup; if $xa = ax$ for all $x \in G$ and $\sigma$ is an automorphism then $\sigma(x)\sigma(a) = \sigma(a)\sigma(x)$ for all $x \in G$, and since $\sigma$ is surjective $\sigma(a)$ is in the center.

Theorem 16. Suppose $G$ is group of automorphisms of a finite group $N$, and suppose $G$ is $m$-transitive on $N - \{e\}$.

a. If $m = 1$ then $N$ is an elementary commutative $p$-group.
b. If $m = 2$ then either $p = 2$ or $|N| = 3$.
c. If $m = 3$ then $|N| = 4$.
d. $m = 4$ is impossible.
Proof: If \( p \) is any prime divisor of \( G \) then (unless \( N \) is trivial) there is an element of \( N - \{ e \} \) of order \( p \), whence every element has order \( p \). The center of \( N \) is thus nontrivial, and since \( G \) must stabilize it \( N \) is commutative. It follows that \( N \) is an elementary commutative \( p \)-group. If \( m = 2 \) and \( p > 2 \), if \( x \in N - \{ e \} \) then \( x \neq x^{-1} \); \( \text{Stab}_x \) must fix \( x^{-1} \), so if it is transitive \( N = \{ e, x, x^{-1} \} \). If \( m = 3 \) a subgroup of order 4 cannot be mapped to a set of elements which does not form such a subgroup, and such would exist unless \( |N| = 4 \). The last claim is now immediate.

Corollary 17. Suppose \( G \) is an \( m + 1 \)-transitive permutation group on \( S \) which has a regular normal subgroup \( N \); then the conclusions of the theorem hold.

Proof: \( \text{Stab}_x \) is \( m \)-transitive on \( S - \{ x \} \); by the above mentioned correspondence, it acts as a group of automorphisms on \( N \), \( m \)-transitively on \( N - \{ e \} \).

Theorem 18.

a. \( A_n, n \geq 3 \), is generated by the 3-cycles.

b. \( A_n, n \geq 3 \), is \((n - 2)\)-transitive, but not \((n - 1)\)-transitive.

c. \( A_n, n \geq 5 \), is simple.

d. \( A_n \) is the only normal subgroup of \( S_n \), \( n \geq 5 \).

Proof: Since \((abc) = (ac)(ab)\), products of 3-cycles are in \( A_n \). On the other hand, every pair of cycles is either of this form, or of the form \((ab)(cd) = (adc)(abc)\). This proves part a. For part b, choose any map from \( n - 2 \) elements to \( n - 2 \) elements. It has 2 extensions to permutations, and it is easy to see that exactly one of these is even. For part c, first we show that \( A_5 \) is simple. Suppose \( N \) is a nontrivial normal subgroup of \( A_5 \); since \( A_5 \) is 3-transitive it is primitive, and so \( N \) is transitive. Hence \( 5 \mid |N| \), so \( N \) contains a 5-cycle, which we may suppose is \((12345)\). Now,

\[
(54321)(354)(12345)(345) = (245),
\]

and since \( N \) is normal in \( A_5 \) and \((345) \in A_5\), \((245) \in N\). By 3-transitivity of \( A_5 \), every 3-cycle is in \( N \), whence by part a \( N = A_5 \). That \( A_{n+1} \) is simple for \( n \geq 5 \) now follows by induction, noting that the stabilizer of a point is \( A_n \); by corollary 17 \( A_{n+1} \) has no regular normal subgroup, so it is simple by theorem 5. For part d, if \( N \) is a normal subgroup of \( S_n \) then \( N \cap A_n \) must be \( \{ e \} \) or \( A_n \) by part c. In the first case \( N \) must be trivial, else it is transitive, which is impossible since its order is 1 or 2. In the second case \( N \) is \( A_n \) or \( S_n \).

It can be shown directly that \( A_n, n \geq 5 \), is simple; see [Hall]. \( A_2 \) and \( A_3 \) are cyclic groups, and simple. \( A_4 \) is not simple; it has a copy of \( Z_2^2 \) as its unique Sylow 2-subgroup. The permutations in it are

\[
(12)(34), \quad (13)(24), \quad (14)(23).
\]

This group is in fact obviously normal in \( S_4 \).

6. Cyclotomic polynomials. We will require cyclotomic polynomials in the next section; they are of course an important topic themselves. If \( F \) is a field its prime subfield is either \( F_p \) or \( \mathbb{Q} \). The polynomial \( x^n - 1 \) is a polynomial over any of these fields; we assume \( F \) equals its prime subfield and let \( E \) be the splitting field over \( F \) of \( x^n - 1 \). The roots of \( x^n - 1 \) in \( E \) are called the \( n \)th roots of unity. If \( F \) is of nonzero characteristic \( p \), let \( n = p^m m \) where \( p \nmid m \). Then \( x^n - 1 = (x^m - 1)^p \); the \( n \)th roots of unity are just the \( m \)th roots, and \( E \) is the splitting field of \( x^m - 1 \). Hence we may assume that \( p \nmid n \) when \( f \) has characteristic \( p \neq 0 \).

Theorem 19. The \( n \)th roots of unity form a cyclic subgroup of order \( n \) of the multiplicative group of \( E \).
Proof: By exercise 7.2 \(x^n-1\) has no repeated roots. If \(x^n = 1\) and \(y^n = 1\) then \((xy)^n = 1\), and \((x^{-1})^n = 1\), so the roots form a subgroup, which by the first claim has order \(n\). By theorem 9.8 the group is cyclic.

Theorem 20. If \(\xi\) is an \(n\)th root of unity and \(\xi \neq 1\) then \(1 + \xi + \cdots + \xi^{n-1} = 0\).
Proof: This is immediate from \((1 - x^n)/(1 - x) = 1 + x + \cdots + x^{n-1}\).

A generator of the group of \(n\)th roots of unity is called a primitive \(n\)th root of unity. If \(d|n\) there are \(\phi(d)\) primitive \(d\)th roots of unity among the \(n\)th roots of unity, where \(\phi\) is the Euler function. The product \(\prod_{\xi}(x - \xi)\) where \(\xi\) ranges over the primitive \(n\)th roots of unity is called the \(n\)th cyclotomic polynomial; we denote it \(Q_n(x)\). It is clearly of degree \(\phi(n)\) and monic.

Theorem 21.

a. \(x^n - 1 = \prod_{d|n} Q_d(x)\).
b. \(Q_n(x) = \prod_{d|n} (x^{n/d} - 1)^{\mu(d)}\) where \(\mu\) is the Möbius function.

Proof: Part a follows since both sides are \(\prod_{\xi}(x - \xi)\) where \(\xi\) ranges over all the \(n\)th roots of unity. Part b follows by exercise 6.9, applied in the multiplicative group of the field of fractions of \(\mathbb{Z}[x]\).

Theorem 22. \(Q_n(x) \in F[x]\); if \(F = \mathbb{Q}\) then \(Q_n(x) \in \mathbb{Z}[x]\).
Proof: This follows by induction; the basis \(n = 1\) follows since \(Q_1(x) = x - 1\). For \(n > 1\),

\[x^n - 1 = Q_n(x) \prod_{d|n, d < n} Q_d(x)\]

call the second factor \(f(x)\). Inductively \(f(x)\) is in the required polynomial ring \(R\). By the division law in \(R\), \(x^n - 1 = q(x)f(x) + r(x)\) with \(\deg(r) < \deg(f)\), so \(0 = (Q_n(x) - q(x))f(x) + r(x)\). It follows that \(Q_n(x) = q(x)\), so \(Q_n(x)\) is in \(R\).

Theorem 23. If \(F = \mathbb{Q}\) then \(Q_n(x)\) is irreducible.

Proof: Write \(x^n - 1\) as \(q(x)f(x)\) where \(q(x)\) is the (monic) irreducible polynomial for a primitive \(n\)th root of unity \(\zeta\); \(q(x), f(x) \in \mathbb{Z}[x]\) by Gauss’ lemma. It suffices to show that if \(p\) is prime and \(p \nmid n\) then \(\zeta^p\) is a root of \(q(x)\), since then clearly \(\zeta^{kp}\) is, whence \(\zeta^k\) is for any \(k\) with \(\gcd(k, n) = 1\). If \(\zeta^p\) is a root of \(f(x)\) then \(\zeta\) is a root of \(f(x^p)\), so \(q(x)|f(x^p)\). Writing \(g(x)\) for the quotient, again \(g(x) \in \mathbb{Z}(x)\).

Reducing \(f(x)\) \(\mod p\), \(f(x)^p \equiv f(x^p) \mod p\). Letting \(\tilde{f}\) denote \(f\) with its coefficients reduced \(\mod p\), we have 

\[\tilde{f}(x)^p = \tilde{q}(x)\tilde{g}(x) \mod \mathbb{Z}_p[x].\]

But then \(\tilde{f}(x)\) and \(\tilde{q}(x)\) have a common factor \(\mod \mathbb{Z}_p\) (e.g. any irreducible divisor of \(\tilde{q}(x)\)), so \(x^n - 1 = \tilde{q}(x)f(x)\) has a multiple root \(\mod \mathbb{Z}_p\). Exercise 7.2 shows that this is impossible, so \(\zeta^p\) is not a root of \(f(x)\), and so it must be a root of \(q(x)\).

Theorem 24. Suppose \(F = \mathcal{F}_p\), \(q = p^e\), and \(d\) is the order of \(q \bmod n\), (where \(p \nmid n\)). Over \(\mathcal{F}_q\), \(Q_n(x)\) factors into \(\phi(n)/d\) distinct irreducible factors of degree \(d\). The splitting field over \(\mathcal{F}_q\) of any such factor contains all the \(n\)th roots of unity.

Proof: Let \(\zeta\) be a primitive \(n\)th root of unity, in the splitting field over \(\mathcal{F}_q\) of \(x^n - 1\). Then \(\zeta \in \mathcal{F}_{q^k}\) for an integer \(k\) iff \(\zeta^q = \zeta\) iff \(\zeta^{q^k - 1} = 1\) iff \(n|q^k - 1\). The smallest \(k\) for which this holds is \(d\), and the theorem follows.

7. Wedderburn’s theorem.

Theorem 25. A finite division ring is a field.
Proof: Let $K$ be the division ring and $Z$ its center; $Z$ is a finite field $\mathbb{F}_q$. $K$ is readily verified to be a vector space over $Z$, say of dimension $d$. For $a \in K$ let $N(a)$ be the normalizer of $a$ in the multiplicative group $K^\neq$. It is readily verified that $N(a) \cup \{0\}$ is a subring of $K$, and so has $q^{d(a)}$ elements for some $d(a)$ where also $q^{d(a)} - 1 | q^d - 1$ so $d(a)|d$. Now, $a \in Z$ iff $d(a) = d$; since $|Z^\neq| = q - 1$ the class equation yields

$$q^d - 1 = q - 1 + \sum_{a \in S} \frac{q^d - 1}{q^{d(a)} - 1}$$

(1)

where the sum is over a system of representatives $S$ of conjugacy classes, excluding the center, so that $d(a) < a$ for $a \in S$. We claim that equation (1), where $d(a)|d$ and $d(a) < d$, is impossible unless $d = 1$ and $S = \emptyset$. Let $Q_d(x)$ be the $d$th cyclotomic polynomial over $\mathbb{Q}$. Then $Q_d(x)|(x^d - 1)/(x^{d(a)} - 1)$, whence $Q_d(q)|(q^d - 1)/(q^{d(a)} - 1)$. By equation (1), $Q_d(q)|q - 1$; to show that this is impossible it suffices to show that $|q - \xi| > q - 1$ where $\xi$ is a root of unity other than 1, because $|Q_d(q)| = \prod_\xi |x - \xi|$ where $\xi$ ranges over the primitive $d$th roots of unity. But if $\xi = e^{2\pi k/d}$ then

$$|q - \xi|^2 = q^2 - 2q \cos(2\pi k/d) + 1,$$

and $\cos(2\pi k/d) < 1$ unless $k = 0$.

This theorem is called Wedderburn’s theorem; it can be proved from equation (1) using the fact that, if $a \geq 2$, $n \geq 3$ then there is a prime $p$ such that $p|a^n - 1$ but $p \nmid a^i - 1$ for $i < n$, except in the case $a = 2$, $n = 6$. See [Artin] for a proof of this, and another application. Many other proofs of Wedderburn’s theorem are known; see [LidNei] for a survey.
15. Characters.

1. Characters of commutative groups. Suppose $G$ is a commutative group and $F$ is a field. A character on $G$ in $F$ is defined to be a homomorphism $\chi$ from $G$ to the multiplicative group of $F$. These maps form a group $\text{Char}_F(G)$ where multiplication is pointwise, that is, $\chi_1\chi_2(x) = \chi_1(x)\chi_2(x)$ for $x \in G$ and $\chi_1, \chi_2 \in \text{Char}_F(G)$. The identity element of $\text{Char}_F(G)$, that is, the map which assigns to each element of $G$ the identity 1 of $F$, is called the trivial character, and any other character is called nontrivial. The trivial character is often denoted $\chi_0$.

If $x \in G$ and $x^n = 1$, then for $h \in \text{Char}_F(G)$ $(h(x))^n = h(x^n) = 1$; that is, $h(n)$ is an $n$th root of unity in $F$. When $F$ is the complex numbers $\mathbb{C}$, $|\chi(x)| = 1$ if $x$ has finite order; $|\chi(x)|$ is often required to be 1 for all $x$. The group of characters in the complex numbers of norm 1 is called the dual group of $G$; we will denote it $G^*$. Note that in $G^*$ the inverse $\chi^{-1}$ of an element $\chi$ equals $\chi^*$, the pointwise complex conjugate of $\chi$; that is,

$$\chi^{-1}(x) = \chi^*(x) = (\chi(x))^* = 1/\chi(x) = \chi(x^{-1}).$$

**Theorem 1.** For any finite commutative group $G$, $G^*$ is isomorphic to $G$.

**Proof:** We first prove the claim for $G$ cyclic. Let $n$ denote the order of $G$, with generator $a$; and let $U$ denote the multiplicitive group of complex $n$th roots of unity, with generator $\zeta = e^{2\pi i/n}$. We have already observed that the range of a character $h$ is contained in $U$. We claim that the map $h \mapsto h(a)$ is an isomorphism from $G^*$ to $U$; since $U$ is cyclic this proves the claim, indeed, the map taking $a$ to $\zeta$ generates $G^*$. The map $h \mapsto h(a)$ is a homomorphism, since $h_1h_2(a) = h_1(a)h_2(a)$; it is injective, since the kernel consists of the trivial map; and it is surjective, since clearly there is an $h$ with $h(a) = \zeta^i$ for any $i$ with $0 \leq i < n$. For the general case, by corollary 8.13 it suffices to show that $G_1^* \times G_2^*$ is isomorphic to $(G_1 \times G_2)^*$. The isomorphism maps $(h_1, h_2) \mapsto h$ where $h(x, y) = h_1(x)h_2(y)$. This is a homomorphism, since

$$(g_1(x)g_2(y))(h_1(x)h_2(y)) = g_1h_1(x)g_2h_2(y).$$

The map $h \mapsto (h_1, h_2)$ where $h_1(x) = h(x, 0)$ and $h_2(x) = h(0, y)$ is readily verified to be a two sided inverse.

**Theorem 2.** Suppose $G$ is a commutative group of order $n$.

a. $\sum_{g \in G} \chi(g) = \begin{cases} n & \text{if } \chi = \chi_0 \\ 0 & \text{otherwise} \end{cases}$

b. $\sum_{\chi \in G^*} \chi(g) = \begin{cases} n & \text{if } g = 0 \\ 0 & \text{otherwise} \end{cases}$

**Proof:** Clearly $\sum_{g \in G} \chi_0(g) = n$. If $\chi \neq \chi_0$ then $\chi(h) \neq 1$ for some $h \in G$. Then

$$\sum_{g \in G} \chi(g) = \sum_{g \in G} \chi(hg) = \chi(h) \sum_{g \in G} \chi(g);$$

hence $\sum_{g \in G} \chi(g) = 0$. For $g \in G$ the map $\chi \mapsto \chi(g)$ is readily seen to be a character on $G^*$, yielding the usual map from $G$ to $G^{**}$. If this could be shown to be injective (hence bijective), equivalently if $\chi(g) = 1$ for all $\chi$ implies $g = 0$, then part b follows by part a. Now, if for a subgroup $H \subseteq G$ $\chi \upharpoonright H$ is trivial then $\chi$ induces a character on $G/H$. Letting $H$ be the subgroup generated by $g$, it follows that there are at most $|G|/|H|$ characters on $G$, so $H$ must be trivial and $g = 0$.

Theorem 2.a is equivalent to the statement that $\sum_{g} \chi_1(g)\chi_2(g)$ equals $n$ if $\chi_1 = \chi_2$, else 0, and theorem 2.b to the statement that $\sum_{\chi} \chi^*(g)\chi(\chi)$ equals $n$ if $g = h$, else 0. For an alternative proof of theorem 2.b, consider the matrix $A$ whose rows (columns) are indexed by the elements of $G^*$ ($G$), where $A_{\chi g} = \chi(g)$. Let $B = (1/n)A^{**}$; by theorem 2.a, $AB = I$ where $I$ is the identity matrix. Thus $BA = I$, which implies that

$$\sum_{x} \chi(h - g) = \sum_{x} \chi^*(h)\chi(g) = \begin{cases} n & \text{if } h = g \\ 0 & \text{otherwise} \end{cases}.$$
Theorem follows by exercise 10.1 and the fact that \( \psi, \phi \). In this case, for \( \ast \), that part b follows.

Theorem 3. For finite \( G \), \( G^\ast \) forms an orthonormal basis for \( C^G \) with the inner product \( \ast \).

Proof: By the remarks following theorem 2 if \( \chi_1, \chi_2 \) are distinct they are orthogonal, and \( \chi \ast \chi = 1 \). The theorem follows by exercise 10.1 and the fact that \( C^G \) is \( n \)-dimensional.

Given \( \phi \in C^G \), the coefficient of \( \chi \) in the expression of \( \phi \) as a linear combination of the \( \chi \in G^\ast \) is easily seen to be \( \psi \ast \chi \). The sum \( \sum \chi (\psi \ast \chi) \chi \) is called the Fourier expansion of \( \phi \).

If \( G \) is the additive group of \( \mathbb{Z}_n \), which is cyclic of order \( n \), then the characters may be indexed by \( \mathbb{Z}_n \) in a natural way, namely, \( \chi_i(s) = \zeta^{is} \) where \( \zeta = e^{2\pi i/n} \). This is just the usual Fourier expansion, mod \( n \). We may write \( \hat{\phi}(r) \) for \( \phi \ast \chi_r \), and call the function \( \hat{\phi} \) the Fourier transform of \( \phi \). Explicitly, \( \hat{\phi}(r) = (1/n) \sum_{s=0}^{n-1} \phi(s) \zeta^{rs} \); the factor \( 1/n \) is often omitted from the definition.

If \( G \) is the elementary commutative group \( \mathbb{Z}_p \), where \( p \) is prime the characters may be indexed by \( \mathbb{Z}_p \) in a natural way. Writing \( \zeta \) for \( e^{2\pi i/p} \), \( r \) for \( \langle r_1, \ldots, r_e \rangle \) and \( s \) for \( \langle s_1, \ldots, s_e \rangle \), and \( r \cdot s \) for the dot product over \( \mathbb{Z}_p \), \( \chi_r(s) = \zeta^{rs} \). Again the function \( \hat{\phi}(r) = \phi \ast \chi_r \) is called the Fourier transform of \( \phi \).

Theorem 4. If \( G \) is a commutative group and \( H \) is a subgroup of \( G \) a character \( \chi \) on \( H \) can be extended to a character on \( G \).

Proof: It suffices to show that \( \chi \) can be extended to the group \( H_a \) generated by \( H \) and \( a \) where \( a \in G - H \), by Zorn’s lemma. That is, the pairs \( \langle D, \psi \rangle \) where \( H \subseteq D \subseteq G \) and \( \psi \) is a character on \( D \) extending \( \chi \) are partially ordered by the relation which holds if \( D \subseteq D' \) and \( \psi' \) extends \( \psi \). This partial order is inductive, so by Zorn’s lemma there is a maximal element. This must have domain \( G \). So let \( m \) be least such that \( a^m \in H \), let \( \chi(a^m) = t \), and let \( u \) be any complex number with \( u^m = t \); if \( a^m \notin H \) for any \( m \) let \( u \) be arbitrary. Each element of \( H_a \) has a unique expression \( ba^i \) where \( b \in H \); if two such were equal a smaller power of \( a \) would be in \( H \). Let \( \psi(ba^i) = \chi(b)u^i \); this is clearly well-defined, and is readily verified to preserve multiplication.

Theorem 5. Distinct characters on a commutative group \( G \) in a field \( F \) are linearly independent, in the vector space of functions from \( G \) to \( F \).

Proof: Suppose \( \sum_{i=1}^r a_i \chi_i = 0 \) is a nontrivial linear combination, where the characters \( \chi_i \) are distinct and \( r \) is as small as possible. Let \( x \in G \) be such that \( \chi_1(x) \neq \chi_2(x) \). Then \( \sum a_i \chi_i(x) \chi_i = 0 \); dividing the second sum by \( \chi_1(x) \) and subtracting from the first yields a shorter nontrivial linear combination, a contradiction.

2. Norm and trace. Suppose \( F \) is a field. Let \( A_n \) be the algebra of \( n \times n \) matrices over \( F \). The determinant is a monoid homomorphism from the multiplicative monoid of \( A_n \) to that of \( F \); and the trace is an \( F \)-linear transformation from \( A_n \) to \( F \). Suppose \( E \supseteq F \) is a finite extension of fields, of degree \( n \), and choose a basis for \( E \) over \( F \). If \( a \in E \) let \( \theta_a \) be the matrix, with respect to the chosen basis, of the linear transformation \( x \mapsto ax \) of \( E \). The map \( a \mapsto \theta_a \) is an \( F \)-algebra homomorphism from \( E \) to \( A_n \); further it is injective, since \( a \neq 0 \) then \( \theta_a \neq 0 \). The image of this map is an isomorphic copy of \( E \) in the algebra of \( n \times n \) matrices over \( F \).
Since det does not depend on the choice of basis, det(θₐ) depends only on a and not the basis. This function is called the norm, from E to F, and denoted N₁ᵋE:F or simply N if E, F are understood. The norm is a monoid homomorphism from the multiplicative group of E to that of F. Similarly, Tr(θₐ) depends only on a. This function is called the trace and denoted Tr₁ᵋE:F. It is an F-linear transformation from E to F.

**Theorem 6.** Let pᵅ and pₑ be the minimal and characteristic polynomials of θₐ; let d, n be their degrees.

a. pᵅ is the irreducible polynomial of a.

b. pₑ = pᵅᵈ.

**Proof:** For part a, let p be any polynomial; then

\[ p(a) = 0 \iff \forall x \in Aₙ(p(a)x = 0) \iff \forall x \in Aₙ(p(θₐ)x = 0) \iff p(θₐ) = 0. \]

For part b, we may choose the basis so that θₐ is in rational canonical form. In each invariant submodule, the identity Bx = ax holds, where B is the block. The blocks are thus all companion matrices of pᵅ.

The eigenvalues of θₐ are the conjugates of a, which we denote a₁, ..., aᵦ. Each occurs with the same multiplicity, which we denote μ. It is easily seen that

\[ N(a) = \prod_{i=1}^{ᵦ} a_i^{μ_i}, \quad \text{Tr}(a) = \sum_{i=1}^{ᵦ} μa_i, \]

where the field operations are carried out in the normal closure K of E. If E ⊇ F is separable, d = r. Otherwise d = p⁺ᵦ for some e > 0 where p is the characteristic; the trace is then identically 0. Let σ₁, ..., σₑ be the distinct embeddings of E over F in K, so that pᵦs = n for some f. It is easily seen that

\[ N(a) = \prod_{i=1}^{ₑ} \sigma_i(a)^{p^f}, \quad \text{Tr}(a) = \sum_{i=1}^{ₑ} p^f \sigma_i(a). \] (1)

Indeed, it suffices to note that the σᵢ map a to each aᵢ equally often, because the automorphism group of K over F is transitive on the aᵢ.

**Theorem 7.** Suppose D ⊇ E ⊇ F is a tower of finite extensions. Then N₁ᵋD:F = N₁ᵋD:E \circ N₁ᵋE:F, Tr₁ᵋD:F = Tr₁ᵋD:E \circ Tr₁ᵋE:F.

**Proof:** This follows by exercise 9.3.b and (1). Write f as f₁ + f₂ for appropriate f₁, f₂, and σᵢ as τᵢρᵦ for appropriate τᵢ, ρᵦ. Since the maps are all homomorphisms, the theorem follows by algebraic manipulation.

**Theorem 8.** Suppose E ⊇ F is finite and separable.

a. Tr is surjective.

b. The map a → ψₐ where ψₐ(x) = Tr(ax) is an isomorphism of the vector spaces E and LᵢᵋF(E; F).

**Proof:** For part a, it suffices to show that Tr is not identically 0. This follows by theorem 6 and (1), since each σᵢ is a character on the multiplicative group of E in K. For part b, that Tr(ax) is an F-linear function from E to F is readily verified. If a ≠ 0 then the function is not identically 0; thus, the map a → ψₐ is injective. It is readily verified to be F-linear, and since E, LᵢᵋF(E; F) both have dimension n the theorem is proved.

If E ⊇ F is finite and separable, let e₁, ..., eᵦ be a basis for E over F. Let e’₁, ..., e’ᵦ be a basis for LᵢᵋF(E; F), such that e’ᵢ(eⱼ) = δ(i, j). Let fᵢ be such that e’ᵢ(x) = Tr(fᵢx), all x. Then f₁, ..., fᵦ is a basis for E with Tr(eᵢfⱼ) = δ(i, j); further it is the unique such basis. It is called the dual basis to e₁, ..., eᵦ.
If \( F = \mathcal{F}_q, q = p^r, p \) prime, is a finite field, and \( E \supseteq F \) has degree \( n \), the norm and trace are given by
\[
N(x) = x^{(q^n-1)/(q-1)}, \quad \text{Tr}(x) = x + x^{q} + \cdots + x^{q^{n-1}}.
\]
This follows by (1) and the fact that \( x \mapsto x^q \) generates the cyclic group of automorphisms of \( E \) over \( F \). The norm is surjective; indeed, clearly \( N(x) \neq 0 \) for \( x \neq 0 \) since none of the conjugates of \( x \) is 0. Considering \( N \) as a group homomorphism on \( F^\times \), its kernel has at most \( t = (q^n-1)/(q-1) \) elements since such elements satisfy the equation \( x^t = 1 \). Its image therefore has at least \( q-1 \) elements.

The kernels of the norm and trace can be characterized for finite fields. In fact, they can be characterized for cyclic extensions, where a finite extension \( E \supseteq F \) is called cyclic if it is Galois and the automorphism group of \( E \) over \( F \) is cyclic.

**Theorem 9 (Hilbert’s Theorem 90).** Suppose \( E \supseteq F \) is cyclic, of degree \( n \), and let \( \sigma \) be a generator of the automorphism group.

a. \( N(a) = 1 \) iff \( a = b/\sigma(b) \) for some \( b \).

b. \( \text{Tr}(a) = 0 \) iff \( a = b - \sigma(b) \) for some \( b \).

**Proof:** That \( N(a) = 1 \) when \( a = b/\sigma(b) \) follows because the norm is the product of the \( \sigma_i(a) \), and the \( \sigma_i(b) \) and \( \sigma_{i+1}(b) \) run through the same values. Conversely given \( a \), let \( a_i = a\sigma(a)\cdots\sigma^{i-1}(a), 0 \leq i \leq n-1; \) then \( a\sigma(a_i) = a_{i+1} \) for \( i < n - 1 \) and \( a\sigma(a_{n-1}) = N(a) \). Let \( \rho = \sum_{i=0}^{n-1} a_i\sigma^i \); if \( N(a) = 1 \) then \( a\sigma \rho = \rho \).

By theorem 4 there is a \( c \) with \( \rho(c) \neq 0 \); let \( b = \rho(c) \). Then \( a\sigma(b) = b \) and part a is proved. For part b, if \( a = b - \sigma(b) \) then \( \text{Tr}(a) = 0 \) since \( \text{Tr}(a) \) is the sum of the \( \sigma^i(a) \) and \( \sigma_i(b) \) and \( \sigma_{i+1}(b) \) run through the same values. Conversely given \( a \) let \( a_i = a + \sigma(a) + \cdots + \sigma^{i-1}(a) \), let \( c \) be such that \( \text{Tr}(c) \neq 0 \), and let \( b = (\sum_{i=1}^{n-1} a_i\sigma^i(c))/\text{Tr}(c) \). If \( \text{Tr}(a) = 0 \) one verifies that \( b - \sigma(b) = a \).

Note that one can also write \( a = \sigma(b)/b \), by considering \( 1/a \); and \( a = \sigma(b) - b \) in the additive case. If \( F = \mathcal{F}_q \) then \( N(a) = 1 \) iff \( a = b^{q-1} \) for some \( b \), and \( \text{Tr}(a) = 0 \) iff \( a = b^{q} - b \) for some \( b \). In this case the converse direction of part b may also be seen by taking \( b \) to be a root of \( x^q - x - a = 0 \), in some extension of \( E \). Then \( a = b^{q} - b \), and \( \text{Tr}(a) \) is readily seen to equal \( b^{q^2} - b \), so if \( \text{Tr}(a) = 0 \) then \( b \in E \).

**3. Characters on finite fields.** Suppose \( q = p^r \) where \( p \) is a prime. A character on the additive group of the finite field \( \mathcal{F}_q \) is called an additive character, and one on the multiplicative group of nonzero elements is called a multiplicative character. Let \( \zeta = e^{2\pi i/p} \), and let \( \zeta_1 = e^{2\pi i/(q-1)} \); let \( g \) be a primitive element of \( \mathcal{F}_q \).

**Theorem 10.** The additive characters on \( \mathcal{F}_p \) are the functions
\[
\chi_a(x) = \zeta^{\text{Tr}(ax)}, \quad a \in \mathcal{F}_q.
\]
The multiplicative characters are the functions
\[
\psi_a(g^x) = \zeta_1^{ax}, \quad a \in \mathbb{Z}_{q-1}.
\]
**Proof:** It is easily seen that \( \chi \) is an additive character iff it is of the form \( \zeta^l(x) \) where \( l \) is an \( \mathcal{F}_p \)-linear function from \( \mathcal{F}_q \) to \( \mathcal{F}_p \). The claim for additive characters follows by theorem 8. The claim for multiplicative characters is a special case of the characterization of the characters of the cyclic group given above.

Note that \( \text{Tr}(ax) \) is a bilinear form, and may be replaced by any nondegenerate bilinear form. The orthogonality relations for finite fields may be written as follows.
\[
\sum_x \chi_a^*(x)\chi_b(x) = q\delta(a, b), \quad \sum_a \chi_a^*(x)\chi_a(y) = q\delta(x, y),
\]
\[
\sum_x \psi_a^*(x)\psi_b(x) = (q - 1)\delta(a, b), \quad \sum_a \psi_a^*(x)\psi_a(y) = (q - 1)\delta(x, y).
\]
4. Dirichlet characters. Characters may be defined on the additive group or on the multiplicative group of units of the ring \( \mathbb{Z}_n \) of the integers mod \( n \). The additive characters are discussed above; they are the functions \( \chi_i(j) = \zeta^{ij} \) where \( \zeta = e^{2\pi i/n} \). The group of multiplicative characters is isomorphic to the group of units of \( \mathbb{Z}_n \), which we will denote \( U_n \). If \( n = q_1 \cdots q_k \) is the prime power decomposition of \( n \) then the map \( \phi(x) = (x_1) \) from \( U_n \) to \( U_{q_1} \times \cdots \times U_{q_k} \) where \( x_i = x \mod q_i \) is an isomorphism; see the remarks following theorem 6.8.

A system of coset representatives for \( \mathbb{Z}_n \) is called a system of residues, mod \( n \). A reduced system of residues is any set of representatives of the cosets which are units in \( \mathbb{Z}_n \); that is, any set of integers \( \{ x \} \) such that \( \{ x \mod n \} = \{ y : 0 \leq y < n, \gcd(y, n) = 1 \} \).

**Lemma 11.**

a. If \( p \) is an odd prime then there is a primitive element \( g \mod p \) such that \( g^{p-1} \not\equiv 1 \mod p^2 \). Any such is a generator of the cyclic group \( U_{p^e} \) for any \( e > 0 \).

b. If \( e \geq 3 \) then \( U_{2^e} \) is isomorphic to \( \mathbb{Z}_2 \times \mathbb{Z}_{2^{e-2}} \), and \( S = \{ (-1)^a5^b : a = 0, 1, 0 \leq b < 2^{e-2} \} \) is a reduced residue system mod \( 2^e \).

**Proof:** For part a, if \( g \) is a primitive element then \( g + p \) is also. By the binomial theorem \( (g + p)^{p-1} \equiv g^{p-1} + (p-1)pg^{p-2} \mod p^2 \), and since \( p^2 \not\mid (p-1)pg^{p-2} \), at most one of \( g^{p-1} \), \( (g + p)^{p-1} \) can be congruent to \( 1 \mod p^2 \). Next, we claim that

\[
\text{if } a \equiv b \mod p^e \text{ then } a^p \equiv b^p \mod p^{e+1}, \text{ provided } e \geq 1; \text{ the case } p = 2 \text{ is included.}
\]

Indeed, write \( b = a + cp^l \), then

\[
b^p = a^p + pb^{p-1}cp^l + \sum_{i=2}^{p} \binom{p}{i} b^{p-i}c^ip^l.
\]

The second term is clearly divisible by \( p^{l+1} \), and the remaining terms are also since \( l \geq 1 \) so \( 2l \geq l + 1 \). Using (1), it follows by induction that for odd \( p \),

\[
(1 + ap)^{p^l} \equiv 1 + ap^{l+1} \mod p^{l+2}.
\]

The basis \( l = 0 \) is trivial, and for the induction step,

\[
(1 + ap)^{p^{l+1}} \equiv (1 + ap^{l+1})^p \equiv 1 + pap^{l+1} + \sum_{i=2}^{p-1} \binom{p}{i} a^ip^{l(i+1)} + ap^pp^{l+1},
\]

congruence being mod \( p^{l+3} \). It is easily checked that the terms of the sum and the last term are divisible by \( p^{l+3} \), proving the claim. By hypothesis \( g^{p-1} = 1 + ap \) where \( p \nmid a \). Using (2) with \( l = e - 1 \) and \( l = e - 2 \) it is easy to see that the order of \( 1 + ap \mod p^e \) is \( p^{e-1} \). From this it is easy to see that the order of \( g \mod p^e \) is \( (p-1)p^{e-1} \) (write the order as \( n = p^{e-1}t \) and note that \( g^n \equiv g^t \mod p \)), and part a is proved. For part b, we first claim that for any \( l, 5^2 \equiv 1 + 2l+2 \mod 2^{l+3} \); this follows easily by induction using (1). It follows that the order of 5 mod \( 2^e \) for \( e \geq 3 \) equals \( 2^{e-2} \). We next claim that the elements of \( S \) are distinct mod \( 2^e \), which proves they are a reduced residue system. Indeed, if \( (-1)^a5^b \equiv (-1)^c5^d \mod 2^{e} \), it follows that \( (-1)^a \equiv (-1)^b \mod 4 \), so \( a = c \) and \( 5^b \equiv 5^d \mod 2^{e} \), and so \( b = d \). Finally the residue classes of the elements of \( S \) form a group under multiplication, which is obviously isomorphic to \( \mathbb{Z}_2 \times \mathbb{Z}_{2^{e-2}} \).

Since the reduced residue system mod 2 equals \( \{ 1 \} \) and that mod 4 equals \( \{ 1, -1 \} \), we have given the group \( U_q \) for any prime power \( q \), and hence the group \( U_n \) for any \( n \). The proof of theorem 1 and remarks above show that the map from \( U_{q_1}^\ast \times \cdots \times U_{q_k}^\ast \) to \( U_n^\ast \) which takes \( \langle \chi_1, \ldots, \chi_k \rangle \) to \( \chi \), where \( \chi(x) = \chi_1(x_1) \cdots \chi_k(x_k) \) and \( x_i = x \mod q_i \), is an isomorphism.
Let \( \hat{U}_n \) denote \( \{ x \in \mathbb{Z} : \gcd(x, n) = 1 \} \). Given a character \( \chi \) on \( U_n \) consider the function \( \hat{\chi} : \mathbb{Z} \mapsto \mathcal{C} \) where

\[
\hat{\chi}(x) = \begin{cases} 
\chi(x \mod n) & \text{if } x \in \hat{U}_n \\
0 & \text{otherwise}.
\end{cases}
\]

Such a function is called a Dirichlet character mod \( n \). Equipped with pointwise multiplication these form a group, which we denote \( D^n \); the map \( \chi \mapsto \hat{\chi} \) is clearly a group isomorphism from \( U_n^* \) to \( D_n \). A Dirichlet character mod \( n \) is easily seen to be completely multiplicative (i.e., \( f(xy) = f(x)f(y) \) for all \( x, y \) and \( f \) is not identically 0), to have period \( n \) (i.e., \( f(x + n) = f(x) \) for all \( x \)), and to satisfy \( f(x) = 0 \) if \( x \notin \hat{U}_n \). On the other hand if \( f : \mathbb{Z} \mapsto \mathcal{C} \) is a function with these properties then \( f = \hat{\chi} \) where \( \chi(x) = f(x) \) for \( x \in U_n \); indeed \( \chi \) is a homomorphism and \( f(1) \neq 0 \) by definition.

Suppose \( \hat{\chi} \) is a Dirichlet character mod \( n, d|n \) where \( d > 0 \), and whenever \( x, y \in \hat{U}_n \) and \( x \equiv y \mod d \) then \( \hat{\chi}(x) = \hat{\chi}(y) \). Then \( d \) is called an induced modulus for \( \hat{\chi} \). Let \( D^d_n \) denote the Dirichlet characters mod \( n \) which have \( d \) as an induced modulus. \( D^d_n \) is clearly a subgroup of \( D_n \). Also, if \( c|d \) then \( D^c_n \subseteq D^d_n \); \( D^1_n = \{ \chi_0 \} \); and \( D^n_n = D_n \).

**Lemma 12.** If \( \gcd(w, d) = 1, d > 0 \), then for any \( n > 0 \) there is a \( t \) such that \( \gcd(w + td, n) = 1 \).

**Proof:** Let \( \{ p_i \} \) be the prime divisors of \( n \). It is readily seen that for each \( i \) there is a \( t_i \) such that \( p_it \mid w + td; \) if \( p_i \mid d \) then \( p_i \mid t \) so let \( t_i = 0 \), and if \( p_i \mid d \) let \( t_i \) be any value other than \( -wd^{-1} \mod p \). By the Chinese remainder theorem there is a \( t \) with \( t \equiv t_i \mod p_i \) for all \( i \), and clearly \( p_i \mid t \mid w + td \).

**Theorem 13.** Suppose \( d, e|n, d, e > 0 \), and let \( \chi_0 \) be the trivial character mod \( n \).

a. The map \( \hat{\psi} \mapsto \hat{\psi}\hat{\chi}_0 \) is an isomorphism from \( D_d \) to \( D^d_n \).

b. Under this isomorphism \( D^d_n \) gets mapped onto \( D^e_n \), for \( c|d \).

c. \( D^d_n \cap D^e_n = D^n_n \) where \( c = \gcd(d, e) \).

**Proof:** Since \( \hat{\psi}\hat{\chi}_0 = \hat{\psi}\hat{\chi}_0 \) the map of part a is a homomorphism. Given a character \( \hat{\psi} \) in \( D^d_n \), define \( \psi(x) \) for \( x \in U_d \) to be \( \chi(y) \) where \( y \in U_n \) and \( y \equiv x \mod d \); such a \( y \) exists by lemma 12, and \( \psi \) is well-defined since \( \chi \in D^d_n \). \( \psi \) is readily verified to be a homomorphism. The map \( \hat{\psi} \mapsto \hat{\psi} \) is readily verified to be a two-sided inverse to the map \( \hat{\psi} \mapsto \hat{\psi}\hat{\chi}_0 \). For part b, suppose \( \hat{\chi} = \hat{\psi}\hat{\chi}_0 \) and \( x \equiv 1 \mod c \). If \( c \) is an induced modulus of \( \psi \) and \( x \in \hat{U}_n \) then \( \chi(x) = 1 \) since \( \chi_0(x) = 1 \) and \( \psi(x) = 1 \) because \( x \in \hat{U}_d \). If \( c \) is an induced modulus of \( \chi \) and \( x \in \hat{U}_t \) then \( \psi(x) = 1 \) because \( x + td \equiv 0 \mod c \), and \( \psi(x) = \chi(x + td) \). For part c, we observed above that \( D^c_n \subseteq D^d_n \). Suppose both \( d, e \) are induced moduli of \( \hat{\chi} \), and let \( f = \gcd(d, e) \). Now, if \( 1 + ad + be \in \hat{U}_f \) then \( 1 + ad, 1 + be \in \hat{U}_f \), since if, say \( p|1 + ad, f \) then \( p|e \) so \( p|1 + ad + be \). It follows using part b that \( c \) is an induced modulus of \( \chi \).

**5. Matrices over noncommutative rings.** Over a noncommutative ring \( R \), matrix addition and multiplication are defined as in the commutative case, and also scalar multiplication on the left and right. These obey many usual identities; in particular the \( m \times n \) matrices form a two-sided module, and the \( n \times n \) matrices form a ring. The \( n \times n \) matrices, considered as a left module, do not quite form an algebra, since \( r(mn) = m(rn) \) does not follow.

For a an \( m \times n \) matrix, the map \( x \mapsto ax \) from column vectors to column vectors is a homomorphism of right \( R \)-modules; let \( \hat{a} \) denote it. The homomorphisms from \( R^n \) to \( R^m \), considered as right \( R \)-modules, form a commutative group with pointwise addition; considering the codomain as a left module also, the homomorphisms form a left \( R \)-module with pointwise scalar multiplication. The map \( a \mapsto \hat{a} \) is an isomorphism of left \( R \)-modules. If \( m = n \) it is also a ring homomorphism.

The map \( x \mapsto ax \) from row vectors to row vectors is a homomorphism of left \( R \)-modules, which again is denoted \( \hat{a} \). The homomorphisms from \( R^n \) to \( R^m \), considered as left \( R \)-modules, form a commutative group with pointwise addition; this is frequently made into a left \( R \)-module by defining \( rf \)( \( x \) ) = \( f(\hat{r}x) \). The map
Proof:

and let scalar multiplication is on the right and it is convenient to consider the elements as column vectors.

Clearly

Proof:

be a minimal nonzero left ideal of \( R \), a simple left ideal is defined to be a minimal nonzero left ideal of \( R \). A ring is called semisimple if it is a semisimple \( R \)-module, i.e., a direct sum of simple left ideals. Although the direct sum for a semisimple module may be infinite, for a semisimple ring it must be finite, because 1 is a sum of elements from a finite number of ideals.

A semisimple ring is called simple if all its simple left ideals are isomorphic; this terminology is not completely consistent, but it is standard.

Lemma 14. If \( A = B \oplus C \) and \( B \subseteq D \subseteq A \) then \( D = B \oplus (C \cap D) \).

Proof: Clearly \( B \oplus (C \cap D) \subseteq D \). Suppose \( d \in D \) and \( d = b + c \) where \( b \in B \), \( c \in C \); then \( c = d - b \in D \).

Theorem 15. A nonzero \( R \)-module \( M \) is semisimple iff every submodule of \( M \) is a summand of \( M \).

Proof: Suppose \( M \) is semisimple, say \( M = \oplus_{i \in I} S_i \) where \( S_i \) is simple, and \( N \subseteq M \) is a submodule. The subsets \( J \subseteq I \) such that \( N \cap \bigoplus_{j \in J} S_j = 0 \) form an inductive family; let \( J \) be a maximal such, let \( N' = \bigoplus_{j \in J} S_j \), and let \( M_1 = N \oplus N' \). If \( j \in J \) then clearly \( S_j \subseteq M_1 \). If \( i \notin J \) then since \( J \) is maximal, \( N \cap (N' \oplus S_i) \) is nonempty. It follows (by writing an element of the intersection as a sum of elements of the \( S_j \) and \( S_i \)) that \( M_1 \cap S_i \subseteq S_i \), so again \( S_i \subseteq M_1 \). It follows that \( M_1 = M \). Conversely let \( J \) be maximal such that \( \sum_{j \in J} S_j = \bigoplus_{j \in J} S_j \) (such \( J \) are readily verified to be inductive). Let \( M' = \bigoplus_{j \in J} S_j \), and let \( N \) be such that \( M = M' \oplus N \). We claim that if \( N \neq 0 \) then \( N \) contains a simple submodule, a contradiction. Indeed, if \( x \in N \), \( x \neq 0 \), let \( N' \) be a maximal submodule of \( N \) not containing \( x \); by the hypothesis and lemma 14 \( N = N' \oplus T \) for some \( T \). If \( T_1 \subseteq T \) is a submodule then again using lemma 14 \( T = T_1 \oplus T_2 \) for some \( T_2 \). By the maximality of \( N' \) one of \( T_1, T_2 \) must be empty, so \( T \) is simple.

Corollary 16. Any submodule or quotient module of a semisimple module is semisimple. If \( M \) is the sum of simple submodules then \( M \) is semisimple.

Proof: A submodule is by lemma 14. A quotient module \( M/N \) is isomorphic to \( N' \) where \( M = N \oplus N' \). A submodule of a sum of simple modules is a summand.

Lemma 17 (Schur’s Lemma). If \( L \) and \( M \) are simple \( R \)-modules and \( h : L \to M \) is a nonzero homomorphism then \( h \) is an isomorphism.
Proof: The image of \( h \) must be \( M \) since it is not 0; the kernel must be 0 since it is not \( L \).

Lemma 18. Suppose \( M \) is a simple \( R \)-module.

a. If \( L \) is a simple left ideal of \( R \) and \( LM \neq 0 \) then \( M \) is isomorphic to \( L \).

b. If \( x \in M \) is nonzero then \( Rx = M \).

Proof: For part a. If \( rx \neq 0, r \in L, x \in M \), the map \( s \mapsto sx \) is a nonzero module homomorphism from \( L \) to \( M \). For part b, \( Rx \subseteq M \) and \( Rx \neq 0 \) so \( Rx = M \).

In a ring \( R \), a nonzero element \( e \) such that \( e^2 = e \) is called an idempotent. We next make some observations about left ideals of the form \( Re \) for an idempotent \( e \). We call a set \( \{e_1, \ldots, e_n\} \) of idempotents orthogonal if \( e_i e_j = 0 \) for \( i \neq j \), and complete if in addition \( 1 = e_1 + \cdots + e_n \). The following are readily verified.

- If \( e \) is an idempotent then \( \{e, 1 - e\} \) is a complete set of idempotents.
- If \( e_1, e_2 \) are orthogonal idempotents then \( \{e_1, e_2, \ldots, e_n\} \) is an orthogonal system of idempotents iff \( \{e_1 + e_2, \ldots, e_n\} \) is.
- If \( \{e_1, e_2, \ldots, e_n\} \) is an orthogonal system of idempotents then \( R(\sum_i e_i) = \oplus_i Re_i \), and the map \( x \mapsto xe_i \) is the projection onto the \( i \)th component.

It follows that if \( x \in Re \) then \( xe = x \).

On the other hand suppose \( Re = \oplus_i L_i \) where \( e \) is an idempotent and \( L_i \) is a left ideal. By suitably numbering the \( L_i \) we may assume \( e = e_1 + \cdots + e_n \) where \( e_i \in L_i \). If \( x \in L_i \) is nonzero then \( x = xe = xe_1 + \cdots + xe_n \); since the sum is direct \( i \) must be one of \( 1, \ldots, n \), and \( x = xe_i \), \( xe_j = 0 \) for \( j \neq i \). In particular \( Re = \oplus_{i=1}^n L_i \) and \( \{e_1, \ldots, e_n\} \) is an orthogonal set of idempotents.

When \( R \) is a semisimple ring, any left ideal \( L \) equals \( Re \) where \( e \) is an idempotent, since \( R = L \oplus L' \) for some left ideal \( L' \). Also, \( R \) is a direct sum \( \oplus Re_i \) of simple left ideals, where \( \{e_i\} \) is a complete set of idempotents. This observation can be refined. Call an idempotent central if it is in the center of \( R \). If \( e \) is a central idempotent \( Re \) is readily verified to be a two-sided ideal, and also a ring with +, \times the restrictions from \( R \) and multiplicative identity \( e \).

Theorem 19. Suppose \( R \) is a semisimple ring. Every \( R \)-module is semisimple. There is a complete set of central idempotents \( \{e_i\} \), such that each \( Re_i \) is a simple ring. Further every simple \( R \)-module \( M \) is isomorphic to a simple left ideal of exactly one \( Re_i \), and in this case for \( x \in M \) \( e_i x = x \) and \( e_j x = 0 \) for \( j \neq i \).

Proof: Any \( R \)-module is a quotient of \( \oplus_{i \in I} R \) for some \( I \), so is semisimple. Define \( R_i, i \in I \), to be the sum of the simple left ideals in an isomorphism class (that is, the collection of finite sums of elements from the ideals). By lemma 18 if \( L_1, L_2 \) are nonisomorphic simple left ideals then \( L_1 L_2 = 0 \); it follows that \( R_i R_j = 0 \) if \( i \neq j \). Since \( R \) is semisimple it is the sum of the \( R_i \). There are therefore values of \( i \), which we write as \( 1, \ldots, n \), and \( e_i \in R_i \), such that \( 1 = e_1 + \cdots + e_n \). If \( x \in R_i \) then \( x = x_1 = 0 \) unless \( i \in \{1, \ldots, n\} \), when \( x = xe_i \). Since \( R_i \) is a left ideal it follows that \( R_i = Re_i \). Now, \( e_j xe_i = 0 \) unless \( j = i \), so \( xe_i = 1xe_i = e_i xe_i = e_i x1 = ex_i \).

It is readily verified that if \( L \) is a simple left ideal in \( Re_i \) then it is one in \( R \), whence all such are isomorphic.

If \( M \) is a simple \( R \)-module then \( RM \neq 0 \), so \( Re_i M \neq 0 \) for some \( i \), so \( LM \neq 0 \) for some simple left ideal \( L \) of \( Re_i \). By lemma 18 \( M \) is isomorphic to \( L \). Clearly \( i \) is unique. For the last claim, it suffices to observe that this holds for the left ideal; certainly \( e_j x = 0 \), and \( e_i x = xe_i = x \).

Theorem 20. Suppose \( A \) is a simple ring.

a. If \( L_1, L_2 \) are two simple left ideals then for some \( u \in A \) \( L_2 = L_1 u \).

b. \( A \) has no nonzero proper two sided ideals; a semisimple ring is simple iff this is so.

c. \( A \) is the ring of \( n \times n \) matrices over a division ring which is a subring of \( A \).
The exercises give sufficient conditions for $M$ simple ring and $\text{End}_R(M)$ for all $\langle e \rangle \mapsto e$ (exercise). By previous claims there are $\alpha_i, \beta_i$ such that $\lambda_i = e_i \alpha_i = \beta_i e_1$ is nonzero; note that $e_i \lambda_i = \lambda_i$ and $\lambda_i e_1 = \lambda_i$. $A \lambda_i = A e_i$, so for some $\gamma_i \gamma_i \lambda_i = e_i$. Let $\rho_i = e_i \gamma_i e_i$; then $\rho_i \lambda_i = e_i, e_1 \rho_i = \rho_i, \rho_i e_1 = \rho_i$, and $\rho_j \lambda_i = 0$ for $j \neq i$. Since $(\rho_i \lambda_i)^2 = e_i \lambda_i \rho_i \neq 0$; also $(\lambda_i \rho_j)^2 = e_i$ and $\lambda_i \rho_i \in D_1; \lambda_i \rho_i = e_i$ follows. Let $e_i j = \lambda_i \rho_j$. Then $e_i j e_k l = 0$ unless $k = j$ when it equals $e_i; e_i$; that these are the rules for multiplying members of the standard basis for the matrix ring. Let $D = \{d : d \in D\}$ where $\hat{d} = \sum_i \lambda_i d \rho_i$. The map $d \mapsto \hat{d}$ is a ring homomorphism, so $D$ is a division ring isomorphic to $D_1$; the multiplicative identity is 1. Also, for $d \in D_1; \hat{d} e_{ij} = \hat{d} \lambda_i \rho_j = \lambda_i d \rho_j$. If $x \in A$ then $x = \sum_i x_{ij}$ where $x_{ij} = e_i x e_j = \lambda_i d_{ij} \rho_j$ where $d_{ij} = \rho_i \lambda_j \in D_1$, and $x_{ij} = \hat{d}_{ij} e_{ij}$. Further $d_{ij} = \rho_i x_{ij} \lambda_j$ so the expression is unique.

**Theorem 21.** Let $A$ be the ring of $n \times n$ matrices over a division ring $D$.

- **a.** $A$ is simple.
- **b.** The center of $A$ consists of the matrices $d I$ where $I$ is the identity matrix and $d$ is in the center of $D$ (a matrix of the form $d I$ is called a scalar matrix).

**Proof:** Let $e_{ij}$ be the matrix which has 1 in the $i, j$ entry and 0 elsewhere. For a matrix $m$, $m e_{ij}$ has its $j$th column equal to the $i$th column of $m$. Also, $e_{ii}$ is an idempotent. It is readily verified that $A e_{ii}$ is a left ideal; further the action of $A$ on $A e_{ii}$ by left multiplication is transitive, whence $A e_{ii}$ is a simple left ideal. Since $A$ equals $\oplus_i A e_{ii} A$ is semisimple. Clearly the $A e_{ii}$ are isomorphic $A$-modules; it follows that $A$ is simple (exercise). For part b, if $m$ is in the center then $m e_{ij} = e_{ij} m$. It follows that $m_{ii} = 0$ unless $t = i, m_{jj} = 0$ unless $s = j$, and $m_{ii} = m_{jj}$; thus, $m = d I$ for some $d$, which clearly must be in the center of $D$.

It follows from theorems 20 and 21 that if $A$ is a simple ring then $A^{\text{op}}$ is. Indeed, if may take $A$ to be the ring of $n \times n$ matrices over a division ring $D$. Replacing $x$ in $A^{\text{op}}$ by $x^t$ shows that $A^{\text{op}}$ is also the ring of $n \times n$ matrices over $D$. Clearly $(\oplus_i A_i)^{\text{op}} = \oplus_i A_i^{\text{op}}$, so if $A$ is semisimple then $A^{\text{op}}$ is. Note that $A^{\text{op}}$ is semisimple iff $A$ is a direct sum of simple right ideals, so that a semisimple ring $A$ has this property. Indeed, every right $A$-module is a direct sum of simple right $A$-modules, and the latter are isomorphic to simple right ideals.

The number of simple left ideals in a decomposition of a simple ring is determined. Given two decompositions, the ideals may be paired off with isomorphisms, and if $R$ is isomorphic to a proper left ideal then an infinite descending chain of left ideals may be found in $R$. Since $n$ is the number of simple left ideals in the ring of $n \times n$ matrices over a division ring, the value $n$ of theorem 20.c is determined.

If $R$ is a ring and $S \subseteq R$ is a subset, the centralizer $S^*$ of $S$ is $\{r \in R : rs = sr\}$ for all $s \in S\}$. Clearly $S \subseteq S^{**}$ (if $r \in S$ and $s \in S^*$ then $rs = sr$). Consider the ring $R$ of additive group homomorphisms of the $A$-module $M$ where $A$ is a ring. Let $S$ be the set of principal homomorphisms. Then $S^*$ is clearly just $\text{End}_A(M)$. $S^{**}$ is called the bicentralizer of $M$; it is those $f \in \text{End}_A(M)$ such that $f \circ g = g \circ f$ for all $g \in \text{End}_A(M)$. Letting $D$ denote $\text{End}_A(M)$ and recalling that $M$ is a $D$-module with the action $\langle f, x \rangle \mapsto f(x)$, the bicentralizer is just the ring $\text{End}_D(M)$. The map $a \mapsto \psi_a$ is a ring homomorphism from $A$ to $\text{End}_D(M)$.

As mentioned previously, $M$ is called faithful if $a \mapsto \psi_a$ is injective; if it is bijective $M$ is called balanced. The exercises give sufficient conditions for $M$ to be a balanced $A$-module. In particular this is so if $A$ is a simple ring and $M$ is a simple $A$-module $Ae$. It follows that the division ring $D$ of theorem 20.c is determined. $D$, which equals $e A e$, is anti-isomorphic to $\text{End}(Ae)$; the exercises give an explicit anti-isomorphism.
7. Group representation. Suppose $G$ is a group, suppose $F$ is a field, and $F[G]$ is the free $F$-algebra over $G$ or “group ring”. As mentioned in chapter 10.3, by a representation of $F[G]$ is meant an $F$-algebra homomorphism $\rho : F[G] \to \text{End}_F(V)$ where $V$ is a finite dimensional vector space over $F$. Two such are isomorphic if they yield isomorphic actions. $\rho$ is determined by $\rho[G]$, which takes values in the group of invertible linear transformations, which is called the general linear group of $V$. The restricted map is called a representation of $G$. Letting $V = F^n$, the codomain of $\rho$ may be considered to be the algebra of $n \times n$ matrices over $F$; we call $n$ the dimension of $\rho$. Clearly every representation is isomorphic to such a representation. From hereon we will only be concerned with finite $G$.

Theorem 22 (Maschke’s Theorem). If $G$ is a finite group, of order $n$, and $F$ has characteristic $p$ where $p \nmid n$, then $F[G]$ is a semisimple ring.

Proof: Suppose $N \subseteq F[G]$ is an $F[G]$-module. Then $N$ is a subspace over $F$ of $F[G]$; let $\pi$ be an ($F$-linear) projection from $F[G]$ to $N$. By hypothesis $1/n$ exists, so we may define

$$p(x) = \frac{1}{n} \sum_{g \in G} g\pi(g^{-1}x)$$

where we have written $gy$ for the action of $G$ on $F[G]$. It is readily verified that $p$ is $F[G]$-linear, because

$$\sum_g g_0 g\pi(g^{-1}x) = \sum g_1 \pi(g_1^{-1}g_0 x).$$

If $x \in N$ then $g\pi(g^{-1}x) = x$ so $p(x) = x$. Thus, over $F[G]$ $p$ is a left inverse to the inclusion map of $N$ in $F[G]$, so $N$ is a summand.

If the hypothesis of the theorem holds, a representation is called ordinary; otherwise it is called modular. We suppose for the rest of this section that a representation is ordinary. The action of $G$ on $F[G]$ by left multiplication is called the left regular representation. By the theorem this is completely reducible, and the irreducible (also called simple) representations, that is, those which have no invariant subspaces, are isomorphic to the action of $G$ on a simple left ideal of $F[G]$. The elements of $G$ in the left regular representation are permutation matrices when the elements of $G$ are taken as a basis for the vector space $F[G]$.

Given any representation, there is a basis in which irreducible representations occur as blocks along the diagonal. An irreducible representation $\rho$, by $n \times n$ matrices, is isomorphic to the algebra of $m \times m$ matrices over a division ring $D$. Clearly $n = md$ where $d = [D : F]$; indeed, we may choose a basis so that the matrices of $\rho$ may be viewed as $m \times m$ matrices of blocks, where each block is the image of some member of $D$ under an obvious representation by $d \times d$ matrices. The dimension of the $F$-algebra of matrices is $dim^2 = n^2/d$.

We fix the following notation. Write $F[G]$ as $\oplus_i R_i$ where $R_i$ is a simple ring, $e_i$ for a central idempotent generating $R_i$, $L_i$ for some simple left ideal of $R_i$, $\rho_i$ for some representation on $L_i$, and $n_i$ for the degree of $\rho_i$. We assume $\rho_1$ is the trivial representation $g \mapsto 1$, and write $r$ for the number of simple rings. Write $d_i$ for the dimension of the division ring. Clearly $|G| = \sum n_i^2/d_i$. In the left regular representation, $\rho_i$ occurs $n_i/d_i$ times (more properly a representation isomorphic to $\rho_i$) since this is the number of times that $L_i$ occurs in the decomposition of the module $F[G]$ into simple left ideals.

If $E \supseteq F$ is an extension, a vector space $V_E$ over $F$ with basis $B$ may be extended to a vector space $V_E$ with basis $B$. In particular a representation $\rho_F$ may be so extended. The result is a representation $\rho_E$ of $E[G]$, where the representation of an element of $F[G]$ is a matrix over $F$. Even if $\rho_F$ is irreducible, $\rho_E$ need not be. A representation is called absolutely irreducible if it is irreducible over any extension of $F$.

We claim that $\rho_F$ is absolutely irreducible iff the division ring $D$ equals $F$. If $D = F$ then $\rho$ is the algebra of $m \times m$ matrices over $F$, and clearly $\rho_E$ is the algebra of $m \times m$ matrices over $E$. Otherwise let
A be in \( D - F \); since \( D \) is finite dimensional over \( F \), by the usual argument there is a polynomial \( p \in F[x] \) such that \( p(A) = 0 \). Let \( E \) be an extension where \( p \) splits, and let \( x - \alpha \) be a linear factor; then \( A - \alpha I \) is a zero divisor in the algebra \( D_E \) of matrices over \( E \) corresponding to \( D \). It follows that \( \rho_E \) is not irreducible, for if it were its division ring would include \( D_E \).

Say that a field \( F \) if every irreducible representation of \( G \) is absolutely irreducible. By the above argument if \( F \) is algebraically closed then \( F \) splits any finite group. We will show in chapter 20 that given a field \( F \) there is a unique field \( F' \) which is algebraic over \( F \) and algebraically closed; \( F' \) is called the algebraic closure of \( F \). It is not necessary to take the algebraic closure, however, to split a given group \( G \). The step of adjoining a root increases the number of simple submodules, so must terminate after a finite number of steps in a finite extension which splits \( G \).

The left regular representation has entries in the prime subfield. It follows from this and the foregoing that every irreducible representation is isomorphic to one with entries in a finite extension of the prime subfield, i.e., a finite field of characteristic \( p \) or a finite extension of \( Q \). It then follows that any representation is isomorphic to such a representation.

The character \( \chi \) of a representation \( \rho \) is defined to be the function \( \chi: G \to F \) where \( \chi(x) = \text{Tr}(\rho(x)) \). A character is a linear map from \( F[G] \) to \( F \), but is not in general a multiplicative homomorphism from \( G \) to \( F \). If \( \rho \) has dimension 1, the character is the same as the representation, and \( \chi \) is a multiplicative homomorphism from \( G \) to \( F \). Two representations by matrices over \( F \) are isomorphic iff their matrices are similar, in which case they have the same character. A character \( \chi \) is called simple, or irreducible, if \( \rho \) is.

We write \( \chi_i \) for the character of \( \rho_i \). Clearly any character \( \chi \) has at least one expression \( \sum_{i=1}^n a_i \chi_i \) where \( a_i \in \mathcal{N} \).

**Lemma 23.** If \( j \neq i \) and \( x \in R_j \) then \( \rho_i(x) = 0 \); and \( \rho_1(e_i) \) is the identity matrix.

**Proof:** For any \( y \in L_i \), \( \rho_i(x)y = xy = 0 \). That is, \( \rho_i(x) \) is the 0 map on \( L_i \). A similar argument proves the second claim.

**Theorem 24.** Suppose \( F \) has characteristic 0.

- The expression of a character \( \chi \) as \( \sum_i a_i \chi_i \), \( i \in \mathcal{N} \), is unique.
- Representations having the same character are isomorphic.
- The simple characters are linearly independent functions from \( F[G] \) to \( F \).

**Proof:** Applying \( \chi \) to \( e_i \), we see that if \( \chi = \sum_i a_i \chi_i \) then \( \chi(e_i) = a_i \chi_i(e_i) = a_i n_i \), so \( a_i = \chi(e_i)/n_i \). Part b follows, because the representations have the multiplicity of \( \rho_i \) determined. Part c follows, because the identically 0 character has a unique expression.

If \( \rho \) is a representation and the group element \( g \in G \) has order \( t \) then \( \rho(g)^t = I \). The eigenvalues of \( \rho(g) \) are thus \( t \)th roots of unity, in some finite extension of \( F \), and for the corresponding character \( \chi \), \( \chi(g) \) is a sum of \( t \)th roots of unity. In particular if \( t \) is the least exponent for \( G \) then the eigenvalues of every \( \rho(g) \) lie in \( F[\xi] \) where \( \xi \) is a primitive \( t \)th root of unity. In the case \( F = Q \), this field in fact splits \( G \). This is a fairly nontrivial theorem, and may be found in [CR], theorem 41.1.

If \( G \) is Abelian and \( F \) splits \( G \) then an irreducible submodule must have dimension 1, since the matrices of dimension greater than 1 over any field do not form a commutative ring. The characters are thus characters in \( F \) in the sense of section 1; \( \chi(g) \) is a \( t \)th root of unity when \( g \) has order \( t \). Note that the proof of theorem 1 holds because of the assumption that the characteristic of \( F \) does not divide \( |G| \), so that the simple characters form a group isomorphic to \( G \) under pointwise multiplication.

For any representation \( \rho \) and group element \( g \), \( \rho(g) \) is diagonalizable (over some finite extension of \( F \)), because \( \rho \) induces a representation on the cyclic group generated by \( g \). If \( F \) splits \( G \) then \( \rho \) is diagonalizable over \( F \), and the eigenvalues of \( \rho(g) \) lie in \( F \), since a basis can be chosen where \( \rho(g) \) is diagonal.
If $\rho$ is a representation the function $\rho^*(g) = (\rho(g^{-1}))^t$ is a representation, called the contragredient representation; that $\rho^*(gh) = \rho^*(g)\rho^*(h)$ is readily verified. It is also readily verified that if $\rho, \sigma$ are isomorphic then $\rho^*, \sigma^*$ are; that $\rho^*$ is irreducible if $\rho$ is; and that $\rho^{**} = \rho$. We write $\chi^*$ for the character of $\rho^*$, so that $\chi^*(g) = \chi(g^{-1})$.

If $F$ is a subfield, which as we observed above involves no loss of generality for characteristic 0) then $\rho(g^{-1}) = \rho(g)^t$ where $M^t$ is the adjoint of the matrix $M$. To see this, note that it is clearly true if $\rho(g)$ is diagonal, and for unitary $U U^{-1} = U^\dagger$ and $(UMU^\dagger)^t = U^t M^t U^\dagger$. Thus, $\chi^*(g) = \chi(g)^*$, where the operation on the right is complex conjugation; this can also be seen directly. Clearly $|\chi(g)| \leq n$ where $n$ is the dimension of $\rho$. If equality holds the eigenvalues are all equal, to $\xi$ say, so $\rho(g)$ is similar to $\xi I$, whence $\rho(g) = \xi I$.

Let $C_1, \ldots, C_s$ denote the conjugacy classes of $G$, where $C_1 = \{1\}$. Let $c_j$ denote the element $\sum_{g \in C_j} g$. Let $C$ denote the set of the $c_j$.

**Lemma 25.** $C$ is a basis for the center $Z$ of $F[G]$.

**Proof:** If $c \in C$ and $g \in G$ then $gcg^{-1} = c$, or $gc = cg$, and it follows that $c \in Z$. Conversely suppose $x = \sum_a a_g g$ is in $Z$. For $h \in G$ $x = xh = \sum_g a_g hgh^{-1}$, so $a_{gh^{-1}} = a_g$; $x$ is thus a linear combination of elements of $C$.

We might write $F[C]$ for $Z$. $F[C]$ is a subalgebra of $F[G]$, since the center is a multiplicative group. We write $\sigma_{ijk}$ for the structure constants, so that $c_i c_j = \sum_k \sigma_{ijk} c_k$. $\sigma_{ijk}$ is an element of the prime subring of $F$, since as an element of $F[G]$ the left side has coefficients in the prime ring and the $C_k$ are disjoint. Thus, more generally for any subring $R$ of $F R[C]$ is a subalgebra of $R[G]$, and the same argument as that of the lemma shows that it is the center.

A function $f$ on $G$ is called a class function if $f(g) = f(h)$ whenever $g, h$ are conjugate. As usual a class function may be extended to $F[G]$ by linearity; the domain may be considered to be $F[C]$. A character $\chi$ is a class function, because $\rho(hgh^{-1})$ is similar to $\rho(g)$.

As above we use $r$ to denote the number of simple rings and $s$ the number of conjugacy classes. The idempotents $e_i$ are in the center $Z$ of $F[G]$, and from the fact that they are orthogonal it follows that they are linearly independent; thus, $r \leq s$. $Z$ equals the product of the centers of the $R_i$, as is readily verified. If $F$ splits $G$ each of these centers has dimension 1, so $Z$ has dimension $r$, and $r = s$ in this case.

For $\alpha, \beta : G \rightarrow F$ let

$$\alpha \ast \beta = \frac{1}{|G|} \sum_{g \in G} \alpha(g^{-1}) \beta(g);$$

this is permissible since $|G|$ is not divisible by the characteristic of $F$. The function $\alpha \ast \beta$ is readily verified to be a symmetric bilinear form on the vector space of functions from $G$ to $F$.

**Theorem 26.** If $F$ splits $G$ then $\chi_i \ast \chi_j = \delta(i, j)$.

**Proof:** Let $\chi$ be the character of the left regular representation; then $\chi(1) = |G|$, and $\chi(g) = 0$ for $g \neq 1$ since the permutation induced by $G$ fixes no element. Now, $e_i = \sum_g a_{ig} g$ for some $a_{ig}$, and

$$\chi(e_i g^{-1}) = \chi(\sum_h a_{ih} h g^{-1}) = \sum_h a_{ih} \chi(h g^{-1}) = |G| a_{ig},$$

$$\chi(e_i g^{-1}) = \sum_{j=1}^s n_j \chi_j(e_i g^{-1}) = n_i \chi_i(e_i g^{-1}) = n_i \chi_i(g^{-1}).$$

Thus, $a_{ig} = n_i \chi_i(g^{-1})/|G|$; since we cannot have $a_{ig} = 0$ for all $g$, $n_i$ is not divisible by the characteristic of $F$. Also,

$$\chi_j(e_i) = \frac{n_i}{|G|} \sum_g \chi_i(g^{-1}) \chi_j(g) = n_i \chi_i \ast \chi_j.$$

167
By lemma 23 $\chi_j(e_i) = n_i \delta(i, j)$ and the theorem follows since $n_i \neq 0$ in the prime subfield.

For the rest of this section we suppose that $F$ splits $G$. It follows from the theorem that the $\chi_i$ are linearly independent functions from $G$ to $F$. In the proof, we showed that $n_i$ is not divisible by the characteristic of $F$. We will show below that if $F$ has characteristic 0 then $n_i$ divides $|G|.

Let $\Xi$ denote the simple characters; and let $F[\Xi]$ denote the $F$-linear combinations. $F[\Xi]$ is an $s$-dimensional subspace of the vector space of functions from $G$ to $F$, and consists of class functions. In fact $F[\Xi]$ is an algebra; indeed, the product of two characters is a character, so the structure constants are in $R$.

The matrix whose $k,i$ by the inverse on the right yields the equation

$$\sum_{j} F_{\sigma W} e \chi_{i+k} \sum_{j} b_j \chi_{j+i} = \delta(i, j).$$

The proof of theorem 26 only requires $\alpha \star \beta$ to be defined on $R_0[\Xi]$ where $R_0$ is the prime subring. It might be desirable to extend $\star$ differently to all the functions. In particular if $F = C$ we define as usual

$$\alpha \star \beta = \frac{1}{|G|} \sum_{g} \alpha(g)^* \beta(g);$$

this is a Hermitian form which agrees with the previous definition on $Z[\Xi]$.

The character table of $G$ is defined to be the $s \times s$ matrix whose $i, j$ entry is $\chi_i(g)$ where $g$ is any element of $C_i$; we write $\chi_{ij}$ for this, and $\chi_{ij}^*$ for $\chi_i(g^{-1})$. Letting $h_k$ denote $|C_k|$, from theorem 26,

$$\frac{1}{|G|} \sum_{k} h_k \sum_{j} \chi_{ik}^* \chi_{jk} = \delta(i, j).$$

The matrix whose $k,i$ entry is $(h_k/|G|) \chi_{ik}^*$ is the inverse to the matrix of the character table; multiplying by the inverse on the right yields the equation

$$\frac{h_k}{|G|} \sum_{i} \chi_{ik}^* \chi_{il} = \delta(k, l).$$

$c_{j} e_{i}$ is in the center of $R_i$, so equals $b_i e_i$ for some $b_i \in F$. Then $\chi_i(c_{j} e_{i}) = b_i n_i$, and $\chi_i(c_{j} e_{i}) = \sum_{g \in C_i} \chi_i(g e_i) = h_j \chi_{ij}$, so $b_i = h_j \chi_{ij}/n_i$. Multiplying $c_{j} c_{j} = \sum_{k} \sigma_{ijkl} c_k$ by $e_i$, it follows that $b_i b_k = \sum_{k} \sigma_{ijkl} b_i$. Writing $b_i$ for the column vector of $b_i$, and $\Sigma_j$ for the matrix of $\sigma_{ijkl}$ the equations for fixed $i, j$ may be rewritten $b_i b_i = \Sigma_j b_i$.

Suppose $F = C$. If we could solve these equations for the $b_i$ then the character table could be determined. We assume that the $\sigma_{ijkl}$ and $b_i$ are given, and also $|G|$; these can be determined from the multiplication table of the group. To determine the $n_i$ from the $b_i$, observe that $\sum_{i} b_i b_i / k = |G|/n_i^2$.

Burnside [Burnside] gives the following method for solving the equations. Introduce a variable $\lambda_j$ for $1 \leq j \leq s$; multiplying the equations for given $j$ by $\lambda_j$ and summing over $j$ yields the equations $\beta_i b_i = \Sigma b_i$ where $\beta_i = \sum_j \lambda_j b_i$ and $\Sigma = \sum_j \lambda_j \Sigma_j$. The $\beta_i$ are thus eigenvalues of the matrix $\Sigma - \beta I$ where $\beta$ is a further variable. Now, if $F = Q[\xi]$ where $\xi$ is a primitive $t$th root of unity and $t$ is the least exponent of $G$, then $\chi_{ij} \in F$, so $b_i \in F$. The equation $\det(\Sigma - \beta I) = 0$, which is homogeneous of degree $s$, thus splits into linear factors in $F$, since the coefficients of the $\lambda_j$ in a linear factor are exactly the $b_i$ for some $i$.

A complex number is called an algebraic integer if it is the root of a monic polynomial with coefficients in $Z$.

- The algebraic integers form a subring of $C$.
- An algebraic integer in $Q$ is in $Z$.

168
- The norm or trace of an algebraic integer (in \( \mathcal{O} \) from some finite extension containing it) is an integer.
- If \( \gcd(m, n) = 1 \) and \( \alpha, m\alpha/n \) are algebraic integers then \( \alpha/n \) is an algebraic integer.

We will prove the first two facts in chapter 20. The third follows by theorem 6, since (up to sign) the sum and product of the conjugates of an algebraic number are coefficients of its irreducible polynomial. The fourth follows since \( am/n + b = 1/n \) for some integers \( a, b \).

A root of unity is an algebraic integer, whence a sum of roots of unity is. A conjugate of a root of unity is a root of unity, of the same order. If \( \chi = \sum_{i=1}^{n} \xi_i \) is a sum of roots of unity and \( |\chi| = n \) then \( |\psi| = n \) for any conjugate \( \psi \) of \( \chi \); this follows because equality holds iff the \( \xi_i \) are all equal, and because \( \psi \) will be the sum of the same conjugate of the \( \xi_i \). It follows that if \( |\chi/n| < 1 \) then \( |N(\chi/n)| < 1 \) where \( N(x) \) is the norm (from some finite extension containing the \( \chi \)).

Suppose \( F = \mathcal{C} \); then a character has algebraic integers for values. We claim that \( b_{ij} \) is an algebraic integer; this follows since it is an eigenvalue of the matrix \( \Sigma_j \) which has integer entries. Thus, \( \sum_j b_{ij} \chi_{ij} \) is an algebraic integer. But this equals \( |G/n_i| \), so the latter must be an integer. This proves that when \( F = \mathcal{C} \), \( n_i \) divides \( |G| \). As observed earlier, if the characteristic is 0 there is no loss of generality in assuming \( F = \mathcal{C} \), but exercise 20.2 gives a proof which avoids this assumption.

**Theorem 27.** Suppose \( G \) is simple, and \( F \) is a finite extension of \( \mathcal{O} \) which splits \( G \).

a. If \( G \) has a nontrivial one dimensional representation then \( G \) is Abelian.
b. This is so if for some \( i > 1, g \neq 1, \rho_i(g) \) is a scalar matrix.
c. This in turn is so if for some \( i, j > 1, \gcd(n_i, h_j) = 1 \) and \( \chi_{ij} \neq 0 \).

**Proof:** A one dimensional representation \( \rho \) is identically 1 on the derived group \( G' \) (see the next chapter for the derived group). \( \rho \) thus induces a one-dimensional representation on \( G/G' \), which is nontrivial if \( \rho \) is, in which case \( G/G' \neq \{1\} \), and so \( G' \subseteq G \). If \( G \) is simple then \( G' = \{1\} \) and \( G \) is commutative. This proves part a. For part b, the subgroup of those \( g \) such that \( \rho_i(g) \) is a scalar matrix is a normal subgroup, which if nontrivial must be all of \( G \). This induces a one-dimensional representation, which is nontrivial since \( i \neq 1 \).

For part c, if \( \gcd(n_i, h_j) = 1 \) then since \( b_{ij} = h_j \chi_{ij}/n_i \), \( \chi_{ij}/n_i \) is an algebraic integer. Now, \( |\chi_{ij}/n_i| \leq 1 \), and if \( |\chi_{ij}/n_i| < 1 \) then \( N(\chi_{ij}/n_i) \) must be 0, whence \( \chi_{ij} = 0 \).

**Exercises.**

1. Suppose \( d \mid n \) and let \( \rho : \mathbb{Z}_n \mapsto \mathbb{Z}_d \) map \( x \) to \( x \mod d \). Show the following.
   a. \( \rho \) maps \( U_n \) to \( U_d \).
   b. The restriction \( \rho_1 \) of \( \rho \) to \( U_n \) has range \( U_d \). Hint: Use lemma 12.
   c. For \( w \in U_d | \rho_1^{-1}[w]| = \phi(n)/\phi(d) \) where \( \phi(t) = |U_t| \) is the Euler function. Hint: Choose any \( u \in \rho_1^{-1}[w] \), and let \( u' \) be its multiplicative inverse mod \( n \). The map \( x \mapsto u'x \) is a bijection from \( \rho_1^{-1}[w] \) to \( \rho_1^{-1}[1] \).

2. Show that the addition operation of the category \( \mathcal{C}^{op} \) is the same as that of \( \mathcal{C} \). Hint: It is an addition operation.

3. Show the following.
   a. Suppose \( R \) is a ring and \( e \) is an idempotent. Then \( eRe \) is a ring, with \( e \) as the multiplicative identity and \( + \) inherited from \( R \).
   b. Suppose \( R \) is a simple ring. If \( Re \) is a simple left ideal then \( eRe \) is a division ring, and conversely. Hint: If \( eRe \) is nonzero then using simplicity \( Rer = Re \), so for some \( s \) \( sere = e = (ese)(ere) \). For the converse, if \( Re \) is not simple it contains orthogonal idempotents.

4. Suppose for a ring \( R \) that \( R = \oplus L_i \) where the \( L_i \) are isomorphic simple left ideals; show \( R \) is simple. Hint: If \( L \) were a simple left ideal not isomorphic to the \( L_i \) then \( L \cap L_i = 0 \), whence \( L \cap R = 0 \).

5. Show that a two-sided ideal in a semisimple ring is a sum of some of the \( Re_i \) in a decomposition as in theorem 19. Hint: Use theorem 20.a.
6. Suppose $A$ is a ring and $e$ an idempotent. For $\phi \in \text{End}_A(Ae)$ let $\hat{\phi} = \phi(e)$. Show that the map $\phi \mapsto \hat{\phi}$ is an anti-isomorphism of rings from $\text{End}_A(Ae)$ to $eAe$.

7. Let $A$ be a ring and $M$ an $A$-module. Show the following.

a. If $M = \sum_{i \in I} S_i$ where $S_i$ is simple then $M = \oplus_{i \in I} S_i$ where $J \subseteq I$.

b. If $M = \oplus_i S_i$ where $S_i$ is simple then every simple submodule of $M$ is isomorphic to some $S_i$. The sum of the simple submodules isomorphic to some $S_i$ is called an isotypic submodule.

c. If $M$ is isotypic it is the direct sum of its isotypic submodules.

d. If $M$ is isotypic and $M = \oplus_{i \in I} S_i$ then the cardinality of $I$ is determined.

Hint: For part a, the collection of $J$ such that $\sum_{j \in J} S_j = \bigoplus_{j \in J} S_j$ is inductive; consider an expression for 0 in the sum over the union of a chain. Hence there is a maximal such $J$, and $S_i$ cannot have empty intersection with the sum over $J$, so is contained in it. For part b, $S = \eta[M]$ for an epimorphism $\eta$; each $\eta[S_i]$ can be nonzero for only one $i$, and is then isomorphic to $S$. For part c, the intersection of an isotypic component with the sum of the remaining ones cannot contain a simple submodule. For part d, the claim for finite $I$ follows using results of the next chapter and the fact that a decomposition of $M$ into a direct sum of finitely many simple submodules yields a composition series for $M$. For infinite $I$, if $M = \oplus_i S_i$ where the $S_i$ are isomorphic simple submodules each $S_j$ in a second such sum is contained in the direct sum of finitely many of the $S_i$, because $S = Ax$ for any $X \in S_j$.

8. Let $M$ be an $A$-module, and let $M'$ be the direct sum of $I$ copies of $M$. Show that the bicentralizers of $M$ and $M'$ are isomorphic. Hint: Let $\pi_i : M' \to M$ and $\mu_i : M \to M'$ be the projection and injection maps, and let 0 be a fixed element of $I$. First, $f \mapsto \sum_i \mu_i f \pi_i$ (where the sum may be defined pointwise since $\pi_i(x)$ is nonzero only finitely often) maps the bicentralizer of $M$ to that of $M'$. Indeed, if $g \in \text{End}_D(M')$ then

$$f'g\mu_i = \sum_j \mu_j f \pi_j g \mu_i = \sum_j \mu_j \pi_j g \mu_i f = g \mu_i f$$

$$= g \sum_j \mu_j f \pi_j \mu_i = g f' \mu_i.$$  

Now write $f'g = \sum_i f'g\mu_i \pi_i$. Second, $f' \mapsto \pi_0 f' \mu_0$ maps the bicentralizer of $M'$ to that of $M$. Indeed, if $g \in \text{End}_D(M)$ then

$$g f = \pi_0 g \pi_0 f' \mu_0 = \pi_0 f' \mu_0 g \pi_0 \mu_0 = f'g.$$  

Finally, the maps are inverse.

9. Suppose $A$ is a ring, $M$ an $A$-module, and $D = \text{End}_A(M)$. A generator in $\text{Mod}_A$ is an $A$-module $M$ such that any $A$-module is the homomorphic image of some power of $M$; for example $A$ is a generator. Show the following.

a. If $A$ is a summand of $M$ then $M$ is balanced.

b. If $M$ is a generator then $M$ is balanced.

c. If $M$ is a generator $M$ is finitely generated over $D$.

d. If $A$ is a simple ring then $D$ is a matrix ring over a division ring.

Hint: For part a, clearly $M$ is faithful. Write $M = A \oplus N$ and suppose $\lambda \in \text{End}_D(M)$. Since the projections $p_1, p_2$ are in $D$, $\lambda(a + n) = \lambda(a) + \lambda(n)$ where $a \in A$, $n \in N$. Since $a' \mapsto a'a$ is in $D$ $\lambda(a) = \lambda(1)a = \lambda(1)a$. Let $\theta_a$ be the map $a + n' \mapsto an$; then $\theta_a \in D$, and $\lambda(n) = \lambda(1) = \theta_n(1) = \theta_n(\lambda(1)) = \lambda(1)n$. Now, if $M$ is a generator then $A$ is a homomorphic image of $M^n$ for some finite $n$, since 1 is a linear combination of finitely many elements of $M$. Also, $M^n = A \oplus N$ for some $N$, since $A$ is a free $A$-module. Part b now follows by part a and exercise 8. For part c, $D^n$ is isomorphic (as a $D$-module) to $\text{Hom}_A(M^n, M)$, where the latter is considered a $D$-module by composition on the left, whence to $\text{Hom}_A(A, M) \oplus \text{Hom}_A(N, M)$. Then prove that the $D$-modules $\text{Hom}_A(A, M)$ and $M$ are isomorphic. For part d, by Schur’s lemma $D$ is a division ring. $A$ is isomorphic to $\text{End}_D(M)$ where the simple left ideal $M$ of $A$ is finitely generated over $D$. 

170
10. Suppose $A$ is a ring, $M$ a semisimple $A$-module, $D = \text{End}_A(M)$, and $f \in \text{End}_D(M)$, and $x_1, \ldots, x_n \in M$. Show that there is an $a \in A$ such that $f(x_i) = ax_i$ for all $i$. Hint: First suppose $n = 1$. Let $\pi$ be the projection from $M$ to $Ax$; then $f\pi = \pi f$, whence $f(x) = f\pi(x) \in Ax$. In general, let $f' \in \text{End}(M^n)$ be as in exercise 8; then $f'((x_1, \ldots, x_n)) = (f(x_1), \ldots, f(x_n))$. 

1. Basic definitions. A sequence of groups $G_0 \supset G_1 \cdots$ is called a descending series, and is said to descend from $G_0$. A sequence of groups $G_0 \subset G_1 \cdots$ is called an ascending series, and is said to ascend from $G_0$. A finite series $G_0 \supset \cdots \supset G_n$ thus descends from $G_0$, and ascends from $G_n$. We call it a series from $G_0$ to $G_n$, and say it is of sidxlength of series length $n$, where this may be 0. Note that this corresponds with its length as a chain of subgroups in the partial order by inclusion of the subgroups.

A descending series where $G_{i+1} \triangleleft G_i$ for all $i$ (with the understanding that $i < n$ if the series terminates at $G_n$) is called a subnormal series. $H$ is called a subnormal subgroup of $G$ if there is a subnormal series from $G$ to $H$.

A proper normal subgroup $N$ of a group $G$ is a maximal proper normal subgroup if, as usual, there is no larger proper normal subgroup of $G$. This is easily seen to be the case iff $G/N$ is nontrivial and simple. Indeed, if $M$ were a larger proper normal subgroup then $M/N$ would be a nontrivial proper normal subgroup of $G/N$. Conversely if $M$ were a proper normal subgroup of $G/N$ its inverse image under the canonical epimorphism would be a proper normal subgroup of $G$ larger than $N$.

If in a finite subnormal series $G_{i+1}$ is a maximal proper normal subgroup of $G_i$ for all $i < n$, the series is called maximal, or a composition series. Clearly a subnormal series is a composition series iff the quotient groups $G_{i+1}/G_i$ are all simple. If $G$ is finite, there is a composition series from $G$ to $\{1\}$; to obtain one, simply repeat the step of taking a maximal proper normal subgroup. There may or may not be a composition series if $G$ is infinite.

Recall that $N(S)$ denotes the normalizer of the set $S$. Say that two subsets $A, B$ of a group $G$ commute if $AB = BA$. Many authors say that $A, B$ are permutable. Also, if $G = AB$ for subgroups $A, B$ then $G$ is said to be factorizable. If, in addition, $AB = BA$ and $A \cap B = \{1\}$ then $G$ is isomorphic to the direct product of $A$ and $B$.

**Lemma 1.** Suppose $A, B, C$ are subgroups of $G$.

a. $AB \subseteq BA$ iff $AB = A \cup B$ iff $AB$ is a subgroup iff $AB = BA$.

b. If $B \subseteq N(C)$ then $B, C$ commute.

c. If $B \subseteq A$ and $B, C$ commute then $A \cap (B \cup C) = B \cup (A \cap C)$.

d. If $A \subseteq N(C)$ and $A \cup B \subseteq N(C)$.

e. If $A \subseteq N(B)$ and $A \subseteq N(C)$ then $A \subseteq N(B \cup C)$

f. If $A \triangleleft C$ and $B \triangleleft C$ then $A \cup B \triangleleft C$.

g. If $A \triangleleft B$ then $A \cap C \triangleleft B \cap C$.

**Proof:** Write $a, a'$ for members of $A$, etc. For part a, if any $ba$ can be rewritten as $a'b'$ then any product $a_1 b_1 a_2 \cdots$ can be rewritten as an element of $AB$. Thus, if $BA \subseteq AB$ then $A \cup B \subseteq AB$, so $A \cup B = AB$. In particular, $AB$ is a subgroup. On the other hand if $AB$ is a subgroup then $A \cup B \subseteq AB$ since $A, B \subseteq A \cup B$. Finally, if $A \cup B = AB$ then $BA \subseteq AB$. For part b, $bc = bcb^{-1}b = c'b$ and $cb = b^{-1}cb = bc'$. For part c, if $B \subseteq A$ then $B, A \cap C \subseteq A \cap (B \cup C)$. For the opposite inclusion, if $x \in A \cap (B \cup C)$ then $x = a = bc$, using part b. Since $c = b^{-1}a$ and $B \subseteq A$, $c \in A$, whence $c \in A \cap C$, whence $bc \in B \cup (A \cap C)$. For part d, let $u = a_1 b_1 \cdots a_k b_k$, and rewrite $u^{-1}cu$ as an element of $C$. For part e,

$$a^{-1}b_1 c_1 \cdots b_k c_k a = a^{-1}b_1 a a^{-1}c_1 a \cdots a^{-1}b_k a a^{-1}c_k a.$$  

Part f follows using part d.

**Lemma 2.** Suppose $a \triangleleft A$, $b \triangleleft B$. Then

$$a \triangleleft (A \cap b) \triangleleft a \triangleleft (A \cap B), \quad (A \cap b) \triangleleft a \triangleleft (a \cap B) \triangleleft A \cap B,$$

and

$$a \triangleleft (A \cap B) \triangleleft a \triangleleft (A \cap b) \triangleleft a \triangleleft (A \cap B).$$
Proof: The first claim follows because $a$ and $A \cap B$ both normalize $a, A \cap b$. The second claim follows because $a \cap B$ and $A \cap b$ are both normal in $A \cap B$. Write $\alpha$ for $a \sqcup (A \cap b)$ and $\beta$ for $a \sqcup (A \cap B)$; clearly $\alpha \sqcup \beta = a \sqcup (A \cap B)$, so the third claim follows from $\alpha \cap \beta = (a \cap B) \sqcup (A \cap b)$. This follows by lemma 1c, since $A \cap b$ and $a$ commute.

**Theorem 3.** Suppose $H = A_n \triangleleft \cdots \triangleleft A_0 = G$ and $H = B_m \triangleleft \cdots \triangleleft B_0 = G$ are subnormal series. Let $A_{ij} = A_i \cup (A_{i-1} \cap B_j)$, $0 < i \leq n$, $0 \leq j \leq m$; $B_{ji} = B_j \cup (B_{j-1} \cap A_i)$, $0 < j \leq m$, $0 \leq i \leq n$. Then $A_{i,j-1}/A_{ij}$ and $B_{j,i-1}/B_{ji}$ are isomorphic.

Proof: Apply lemma 2 with $\{A, B\} = \{A_{i-1}, B_{j-1}\}$ in both possible ways.

**Corollary 4.** If $A_n \triangleleft \cdots \triangleleft A_0$ in the theorem is a composition series then $m \leq n$.

Proof: In the series from $A_{i-1}$ to $A_i$, all but one of the quotient groups are trivial. In the series of all $B_{ji}$ there are $n$ nontrivial quotients, and $m \leq n$ follows.

**Corollary 5 (Jordan-Holder Theorem).** If both series in the theorem are composition series then $m = n$. Further the quotients $A_i/A_{i-1}$ and $B_i/B_{i-1}$, $1 \leq i \leq n$, are isomorphic in pairs.

In a subnormal series, if $N$ is a normal subgroup of $G$, we say $N$ is normal in $G$. We say that $N$ is normal in $G$ if $N$ is normal in $G$. Suppose we write $G = N \oplus H$. The intersection or join of a collection of $\Omega$-subgroups is a normal subgroup. If $\Omega$ is a monoid and $G$ a group an action of $\Omega$ on $G$ is defined to be a homomorphism from $\Omega$ to $\text{Hom}(G; G)$. An action is thus given by a map $\langle \omega, g \rangle \mapsto \omega g$ such that

$$\omega(xy) = (\omega x)(\omega y), \quad 1x = x, \quad (\omega_1 \omega_2)x = \omega_1(\omega_2 x).$$

Fixing $\Omega$ we may define a subcategory $\Omega$-Grp of Grp by defining an $\Omega$-homomorphism $f : G \to H$ to be a homomorphism such that $f(\omega x) = \omega f(x)$.

An $\Omega$-subgroup of $G$ is a subgroup $H$ which is closed under the action of $\Omega$, or using the usual notation, such that $\Omega H \subseteq H$.

We leave it as an exercise to verify the following.

- The forgetful functor from $\Omega$-Grp to Grp creates limits and colimits.
- The intersection or join of a collection of $\Omega$-subgroups is a $\Omega$-subgroup.
- If $N$ is a normal $\Omega$-subgroup then $\Omega$ acts on $G/N$ via $\omega(x + N) = \omega x + N$.
- The homomorphism theorems hold for $\Omega$-subgroups and $\Omega$-homomorphisms.

From these it is easy to see that lemma 2 holds for $\Omega$-subgroups. Thus if we restrict all groups to be in $\Omega$-Grp, theorem 3 and its corollaries still hold. The isomorphisms of corollary 5 are $\Omega$-isomorphisms.

If $\Omega$ is the group of inner automorphisms of a group $G$ the $\Omega$-subgroups are the normal subgroups. An $\Omega$-homomorphism between two subgroups is called a central homomorphism. A descending subnormal series from $G$ is called a normal series. That is, a normal series is a series of normal subgroups, such that $G_{i+1} \triangleleft G_i$. A maximal normal series, i.e., one where $G_{i+1}$ is maximal among the normal subgroups of $G_i$ which are normal in $G$, is called a principal series. If $\Omega$ is the group of all automorphisms of $G$, a descending subnormal series is called a characteristic series.

Suppose $N_i$, $1 \leq i \leq n$, are normal subgroups of a group $G$. Write $N_i'$ for $\sqcup_{j \neq i} N_j$. $G$ is said to be the internal direct product of the $N_i$ if $G$ is their join, and $N_i \cap N_i' = \{1\}$ for all $i$. 173
Lemma 7. If $G$ is the internal direct product of $N_1, \ldots, N_n$ then $G$ is isomorphic to $\times_i N_i$. On the other hand $\times_i N_i$ is the internal direct product of the standard embeddings of the $N_i$.

Proof: Exercise.

If $G$ has a composition series then by corollary 6 any normal series is finite, and it follows that $G$ has a principal series.

Theorem 8. $G$ has a composition series if and only if it has a principal series where each quotient $G_i/G_{i+1}$ is a product of finitely many isomorphic simple groups.

Proof: Let $N$ be minimal among the normal subgroups of $G_i$ which properly contain $G_{i+1}$. If $N = G_i$ we are done. Otherwise, the conjugates of $N$ in $G_i$ are all normal subgroups of $G_i$; their union is a normal subgroup of $G_i$ and so equals $G_i$. Choose a sequence $N_0, N_1, \ldots$ of the conjugates of $N$ so that $N_s \not\subseteq \cup_{i<s} N_i$; let $L_s = \cup_{i<s} N_i$. The $L_s$ are normal subgroups of $G_i$, and the sequence must terminate at $G_i$ after finitely many steps since $G$ has a composition series. We may delete an $N_s$ contained in the join of the remaining ones, until there are no such; the $L_s$ still terminate at $G_i$. Each $N_s$ then intersects the join of the remaining ones in $G_i/G_{i+1}$, by the minimality condition on $N$. By lemma 7, $G_i/G_{i+1}$ is the internal direct product of the $N_s/G_{i+1}$. Finally, a proper normal subgroup of $N_i/G_{i+1}$ would be normal in $G_i/G_{i+1}$ (because $N_i/G_{i+1}$ is a direct factor) and so $N_i/G_{i+1}$ is simple. The theorem follows readily.

An $R$-module is a group with operators, satisfying some additional equations; we have already seen that the facts mentioned above hold in this case, whence theorem 3 and its corollaries do. In a descending series, of course, $M_{i+1}$ is a submodule of $M_i$. Note also that $N$ is a maximal nonzero proper submodule of $M$ if and only if $M/N$ is a simple module. A module which has a composition series to 0 is said to have finite length, and to be of length $n$ where $n$ is the length of the series.

A vector space $V$ has finite length if and only if it is finite dimensional, and in this case the length equals the dimension. Indeed, in a composition series the dimension changes by one, and an $n$-dimensional space has a composition series of length $n$.

2. Modular lattices. Lemma 1c implies that the lattice of normal subgroups of a group is modular; a fortiori this is so for the subgroups of a commutative group, the normal $\Omega$-subgroups of a group, or the submodules of a module. In these cases the language of lattice theory is often convenient. In a lattice write $x/y$ for a pair of elements where $y \leq x$. Two pairs $x \sqcup y/x$ and $y/x \sqcap y$ are said to be perspective. Two pairs are called projective if there is a chain of perspectives between them. Note that in the subgroup lattices, the quotients corresponding to two projective pairs are isomorphic. Let $[x/y]$ denote the “interval” $\{z : x \leq z \leq y\}$; this is a sublattice.

Lemma 9. In a modular lattice, the map $u \mapsto u \sqcap y$ is a lattice isomorphism from $[x \sqcup y/x]$ to $[y/x \sqcap y]$, with inverse $v \mapsto x \sqcup v$.

Proof: In any lattice, $u \mapsto u \sqcap y$ is readily verified to be a meet preserving map between the two sublattices, and $v \mapsto x \sqcup v$ and $v \mapsto x \sqcup v$ a join preserving map in the opposite direction. It therefore suffices to show that the maps are inverse to each other (since being order preserving they then preserve inf’s and sup’s). By the modular law, $x \sqcup (u \sqcap y) = u \sqcap (x \sqcap y) = u$; similarly $(x \sqcup v) \sqcap y = v \sqcap (x \sqcap y) = v$.

Lemma 10. In a modular lattice, suppose $a \leq A, b \leq B$. Then $a \sqcup (A \sqcap B)/a \sqcup (A \sqcap b)$ and $A \sqcap B/(A \sqcap b) \sqcup (a \sqcap B)$ are perspective.

Proof: The proof of lemma 2 goes through; that $a \cap b = (a \cap b) \sqcup (A \cap b)$ follows by modularity.
Theorem 3 and its corollaries go through, with appropriate modification of terminology; namely, replace “subnormal series” by “chain”, “maximal proper normal subgroup” by “maximal properly smaller element”, “composition series” by “maximal chain”, and “isomorphic” by “projective”. A length function \( l(x/y) \) may be defined on the pairs; this is a finite integer if there is a maximal chain from \( x \) to \( y \), namely the length of any such; or \( \infty \) otherwise.

**Theorem 11.** Suppose a modular lattice \( L \) has a greatest element \( U \) and a least element \( 0 \). Then \( l(U/0) \) is finite iff \( L \) satisfies the ascending and descending chain conditions.

**Proof:** If \( l(U/0) \) is finite, by corollary 4 it is an upper bound on the length of any ascending or descending chain. Conversely, let \( x_0 = U \), and proceeding inductively if \( x_k \neq 0 \) we can find a maximal properly smaller element since \( L \) satisfies the ascending chain condition. Since \( L \) satisfies the descending chain condition we must eventually reach 0.

From hereon we assume that \( L \) is a modular lattice with a least element \( 0 \), and further that \( l(x/0) \) is finite for all \( x \in L \). We write \( l(x) \) for \( l(x/0) \). Two elements \( x, y \) are called projective if the pairs \( x/0, y/0 \) are.

**Lemma 12.** If \( y \leq x \) then \( l(x/y) = l(x) - l(y) \). Also \( l(x \lor y) = l(x) + l(y) - l(x \land y) \).

**Proof:** There is a maximal chain from \( y \) to 0 containing \( x \) by corollary 6, and the first claim follows. The second claim follows because \( l(x \lor u/x) = l(y/x \land y) \) by lemma 9.

If \( x = \sqcup_i y_i \) and \( y_i \land y'_i = 0 \) where \( y'_i = \sqcup_{j \neq i} y_j \) we say that \( x \) is the direct join of the \( y_i \). Note that a group is the direct join of some normal subgroups exactly if it is their internal direct product. If an element \( x \in L \) can be written as a direct join \( y \sqcup z \) where \( y \land z = 0 \) we say that \( x \) is decomposable, else indecomposable. If \( x \) is the direct join of some \( y_i \), where \( y_i \) is indecomposable, we call \( \sqcup_i y_i \) a decomposition of \( x \).

Note that under the hypotheses on \( L \) any element \( x \) has a decomposition; indeed, it suffices that \( L \) satisfy the descending chain condition. If there is an element without a decomposition let \( x \) be a minimal such; \( x \) can neither be indecomposable nor decomposable. It follows that any expression of \( x \) as direct join can be broken up into a decomposition. The proof of the following theorem follows [Hall].

**Theorem 13.** Suppose \( \sqcup_{i=1}^m y_i \) and \( \sqcup_{j=1}^m z_j \) are two decompositions of \( x \). Then for each \( i \) there is a \( j \) such that \( y_i \) and \( z_j \) may be replaced by each other; such \( y_i, z_j \) are projective. Finally \( m = n \).

**Proof:** We prove the theorem by induction on \( l(x) \), the basis \( x = 0 \) being trivial. Using the notation \( y'_i \) as above, suppose \( y_i \sqcup z'_j = y'_i \sqcup z_j = x \). Then

\[
l(y_i) = l(x) - l(z'_j) + l(y_i \land z'_j) \geq l(z_j),
\]

and similarly \( l(z_j) \geq l(y_i) \), whence \( l(y_i) = l(z_j) \); it follows that \( y_i \land z'_j = y'_i \land z_j = 0 \). Suppose \( y_i \sqcup z'_j < x \); renumbering we suppose \( j = 1 \), and we will find a \( j > 1 \) such that \( y_1 \) and \( z_j \) can replace each other. Write \( y, z \) for \( y_1, z_1 \). Let

\[
p_j = y \sqcup z'_j, \quad q_j = p_j \land z_j, \quad r_j = \sqcup_{k \leq j} q_j, \quad s_j = \sqcup_{k \leq j} z_j, \quad t_j = \sqcap_{k \leq j} p_j
\]

for \( 1 \leq j \leq m \). Now, if \( q_1 = z_1 \) then \( p_1 = x \) contrary to hypothesis. Clearly \( r_m \) is the direct join of the \( q_k \). It follows that \( l(r_m) < l(x) \). We claim next that \( r_j = s_j \sqcap t_j \); the proof is by induction on \( j \). The basis \( j = 1 \) reduces to \( q_1 = z_1 \sqcap p_1 \). For the induction step,

\[
r_{j+1} = r_j \sqcup q_{j+1} = (s_j \sqcap t_j) \sqcup (p_{j+1} \sqcap z_{j+1}) = p_{j+1} \sqcap (s_j \sqcap t_j) \sqcup z_{j+1},
\]

where
the last step following by modularity since

\[ s_j \cap t_j \leq s_j \leq z^j_{j+1} \leq p_{j+1}. \]

But \( z_{j+1} \leq b'_k \leq p_k \) for \( k \leq j \), so \( z_{j+1} \leq t_j \), whence again by modularity

\[ r_{j+1} = p_{j+1} \cap (t_j \cap (z_{j+1} \cup s_j)) \]

and the induction step follows. Write \( r \) for \( r_m \). Since \( y \leq p_j \) for all \( j \), \( y \leq t_m \); since \( s_m = x \), \( r = t_m \); and so \( y \leq r \). By modularity \( (r \cap y') \cup y = r \), and clearly this join is direct. By the main induction hypothesis applied to \( r \) there is a \( j \) such that \( y \) and an indecomposable submodule \( e \) of \( q_j \) can replace each other. Since \( e \cap y' = e \cap (r \cap y') = 0 \) and \( d(e) = d(y) \), \( x \) is the direct join of \( e \) and \( y' \). Since \( e \leq q_j \leq z_j \), it follows that \( e = z_j \); indeed \( e \cap (y' \cap z_j) = 0 \), and using modularity \( e \cup (y' \cap z_j) = z_j \), and \( z_j \) is indecomposable. Since \( z_j = e \leq q_j \leq z_j \), \( z_j = q_j = (y \cup z'_j) \cap z_j \), whence \( z_j \leq y \cup z'_j \), so \( y \cup z'_j = x \); note that \( j > 1 \) follows. The join is direct because \( l(y \cap z'_j) = 0 \) since \( l(y) = l(z_j) \) and \( l(y \cup z'_j) = l(x) \). There remains the case when \( a_i \cup b'_k = x \) for all \( j \) but \( a_i \cup b_j < x \) for all \( j \). By renumbering and applying the previous case we may assume that \( i = 1 \) and \( y_n, z_m \) can replace each other. By lemma 9 the map \( \phi \) where \( \phi(u) = (u \cap y_n) \cup y'_n \) is a lattice isomorphism from \([b'_m/0]\) to \([a'_n/0]\), and clearly \( \sqcup_{j=1}^{m-1} y_i \) and \( \sqcup_{j=1}^{m-1} \phi[z_j] \) are two decompositions of \( y'_n \). By induction there is a \( j \leq m - 1 \) such that \( y_i \) and \( \phi[z_j] \) can replace each other. Now,

\[ z_j \cup y_n = \langle z_j \cup y_n \rangle \cap \langle y'_n \cup y_n \rangle = \langle z_j \cup y_n \rangle \cap y'_n \cup y_n = \phi[z_j] \cup y_n. \]

It follows that \( z_j \cup y'_i = \phi[z_j] \cup y'_i = x \), a contradiction which eliminates this case. The first claim of the theorem is thus proved. If \( y_i \) and \( z_j \) can replace each other they are both perspective to \( x/y'_i \). To see that \( m = n \), successively replace \( y_1, \ldots \), by \( z_j \)'s; a \( z_j \) can only be used once, and all of them must get used.

For decompositions into internal direct products of normal subgroups the theorem is known as the Wedderburn-Remak-Schmidt theorem. The factors \( y_i \), and \( z_j \) are centrally isomorphic, since this is true of the quotient groups of perspective pairs in the lattice of normal subgroups of a group. We leave it as an exercise to show that if \( \phi : y_i \rightarrow z_j \) is a central isomorphism then \( u^{-1} \phi(u) \) is in the center of \( x \), for all \( u \in y_i \).

3. Solvable groups. A descending series \( G = G_0 \supset G_1 \cdots \) is called Abelian, cyclic, or central if for all \( i \) \( G_i/G_{i+1} \) is Abelian, cyclic, or in the center of \( G/G_{i+1} \) respectively. A group is called
- solvable if it has an Abelian finite subnormal series;
- supersolvable if it has a cyclic finite normal series; and
- nilpotent if it has a central finite normal series.

Clearly a supersolvable or nilpotent group is solvable. We will show below that a nilpotent group is supersolvable if (and only if) its subgroup lattice satisfies the ascending chain condition. In particular this is so for finite groups.

If \( G \) is a group and \( g, h \in G \) the commutator of \( g, h \) is defined to be \( g^{-1}h^{-1}gh \). The derived subgroup is the subgroup generated by the commutators. The derived subgroup is a characteristic subgroup, since any automorphism (indeed any endomorphism) maps a commutator to a commutator. We let \( G' \) denote the derived group. \( G' \) is easily seen to be the smallest normal subgroup \( N \) of \( G \) such that \( G/N \) is Abelian. This follows because for a normal subgroup \( N \), \( xyN = yxN \) iff \( x^{-1}y^{-1}xy \in N \). Also, if \( H \subseteq G \) is a subgroup then \( H' \subseteq G' \). We let \( G^{(k)} \) be the \( k \)th derived group, i.e., the result of applying the derivation step \( k \) times; \( G^{(k)} \) is clearly a characteristic subgroup for all \( k \).

Theorem 14. The following are equivalent for a group \( G \).

a. \( G^{(k)} = \{1\} \) for some \( k \).

b. \( G \) has a finite Abelian normal series.

c. \( G \) is solvable.

176
supersolvable. Applying the canonical epimorphism to a series for $P$ are characteristic subgroups since they are the unique subgroups of the given order. The inverse images of $P$ under the canonical epimorphisms may be inserted in the given series, and the resulting series is still supersolvable. It suffices to show this for $Q$ and let $Q = Q_0 \supseteq \cdots$ be a finite Abelian subnormal series for $G/N$. Applying $\eta^{-1}$, and continuing with a finite Abelian subnormal series for $N$, shows that $G$ is solvable.

A supersolvable group has a normal series in which every quotient is either infinite cyclic or has prime order. Indeed, if there is a quotient $Q = G_i/G_{i+1}$ of order $p_1 \cdots p_r$ where the $p_i$ are (not necessarily distinct) primes, then there is a descending series $Q = P_r \supseteq \cdots \{1\}$, where $P_i$ has order $x \leq i p_j$. The $P_j$ are characteristic subgroups since they are the unique subgroups of the given order. The inverse images of the $P_i$ under the canonical epimorphisms may be inserted in the given series, and the resulting series is still normal.

A subgroup $H$ of a supersolvable group $G$ is supersolvable. Given a cyclic finite normal series for $G$, intersect each group with $H$. This is a normal series for $H$, and $H \cap G_i/G_i \cap G_{i+1} \cong (H \cap G_i)G_{i+1}/G_{i+1}$. The latter group is a subgroup of $G_i/G_{i+1}$ and so is cyclic. A quotient $G/N$ of a supersolvable group is supersolvable. Applying the canonical epimorphism to a series for $G$ yields a series for $G/N$. A finite direct product of supersolvable groups is supersolvable. It suffices to show this for $G \times H$; if $G_0 \supseteq \cdots$ is a series for $G$ then a series for $G \times H$ is obtained by following $G_0 \times H \supseteq \cdots$ by a series for $H$.

The lattice of subgroups of a group satisfies the ascending chain condition iff every subgroup is finitely generated; the proof is essentially the same as that of theorem 6.4.a. If $G$ is supersolvable then its lattice of subgroups has the ascending chain property. By the foregoing it suffices to show that $G$ is finitely generated. But this follows readily from the definition.

On the other hand suppose $G$ is nilpotent and its lattices of subgroups satisfies the ascending chain condition. A quotient group of a central finite normal series $G_0 \supseteq \cdots$ is then a finitely generated Abelian group, so may be refined to a finite series of cyclic groups. If the given series is refined accordingly, the resulting series is normal. Indeed, letting $G_{ij}$ denote one of the inserted groups, $G_{ij}/G_{i+1} \triangleleft G_i/G_{i+1}$ since the given series is central. It can be shown that it suffices to assume that $G$ is finitely generated (see [Hall]).

The commutator $[X, Y]$ of two subgroups of a group $G$ is the subgroup generated by $\{x^{-1}y^{-1}xy : x \in X, y \in Y\}$. For example, $G' = [G, G]$. Let $D_0 = G$, and inductively $D_{i+1} = [D_i, G]$. We claim that given any descending central normal series $G = G_0 \supseteq \cdots$, $D_i \subseteq G_i$. It follows that $G$ is nilpotent iff $D_i = \{1\}$ for some $i$. In this case the series of $D_i$ is called the lower central series; note that it is characteristic. To prove the claim, first note that $[G_i, G] \subseteq G_{i+1}$ if (and only if) $G_i/G_{i+1} \subseteq Z(G/G_{i+1})$; the claim follows easily by induction.

Let $Z_0 = \{1\}$, and inductively $Z_{i+1}$ is the inverse image of $Z(G/Z_i)$ under the canonical epimorphism. We claim that in any ascending series $\{1\} = G_0 \supseteq \cdots$ of normal subgroups of $G$ such that $G_{i+1}/G_i \subseteq Z(G/G_i)$ for all $i$, $G_{i+1} \subseteq Z_i$. It follows that $G$ is nilpotent if $Z_i = G$ for some $i$. In this case the series of $Z_i$ is called the upper central series; note that it is characteristic. To prove the claim, first note that by an observation above $[G_i+1, G] \subseteq G_i$. Secondly, $G_{i+1}Z_i/Z_i \subseteq Z(G/Z_i)$ if (and only if) $[G_i+1, G] \subseteq Z_i$. The claim follows by induction.

\[177\]
If $G$ is nilpotent the lengths of the upper and lower central series are equal. Indeed, if $n$ is least such that $Z_n = G$, and $m$ least such that $D_m = \{1\}$, then then $D_i \subseteq Z_{n-i}$, so $D_n = \{1\}$ and $m \leq n$; similarly $n \leq m$. The length of these sequences is called the nilpotence class of $G$. Note that a group has nilpotence class 1 iff it is Abelian.

If $G$ has nilpotence class $n$, $H \subseteq G$ is a subgroup, and $N \triangleleft G$ then $H, G/N$ have nilpotence class at most $n$. This is readily seen by considering the lower central series.

Following is a list of some other basic facts about solvable, supersolvable, and nilpotent groups. Some of these are proved in the exercises. For proofs of the remaining ones and other facts see [Hall] and [Bechtel].

1. Suppose $M, N \triangleleft G$, $M$ has nilpotence class $n$, and $m$ has nilpotence class $m$; then $NM$ has nilpotence class at most $m + n$.
2. If $G$ is supersolvable then $G'$ is nilpotent.
3. In a principal series for a finite solvable group the quotient groups are elementary Abelian. The orders of the quotients are called the chief factors of the group.
4. A finite $p$-group is nilpotent.
5. For finite $G$ the following are equivalent. (for any $G$ a$\Rightarrow$b$\Rightarrow$c).
   a. $G$ is nilpotent.
   b. A proper subgroup of $G$ is a proper subgroup of its normalizer.
   c. Every maximal subgroup is normal.
   d. $G$ is the internal direct product of its Sylow subgroups.
6. For finite $G$, $G$ is supersolvable iff every maximal subgroup has prime index.

For the last fact, we give some definitions. Let $\pi$ be a finite set of prime numbers. A $\pi$-number is one whose prime divisors are in $\pi$ (if $\pi$ is empty is the only $\pi$ number). A $\pi$-group is one where the order of every element is a $\pi$-number. If $G$ is a finite group and $\pi$ is a subset of the prime divisors of $|G|$ let $\pi'$ denote the remaining prime divisors. A Hall $\pi$-subgroup of $G$ is a $\pi$-subgroup whose index is a $\pi'$-number. For example, for a prime $p$ a Hall $p$-subgroup is the same as a Sylow $p$-subgroup. A Hall $p'$-subgroup is also called a $p$-complement. By a Sylow base for a finite group $G$ is meant a collection of $p$-Sylow subgroups, one for each prime divisor $p$ of $|G|$, such that any two of the subgroups commute.

**Lemma 16.** Suppose $G$ is a finite group and $A, B, C$ are subgroups.

a. If each pair of $A, B, C$ commute then $AB, C$ commute.

b. $[B : A \cap B] \leq [A \cup B : A]$, with equality iff $A, B$ commute.


**Proof:** For part a, $abc = acb = cab$. For part b, if $x, y \in B$ then $xy^{-1} \in A$ iff $xy^{-1} \in A \cap B$. The inequality follows, and if $A, B$ commute then coset representatives of $A$ in $BA$ can be chosen in $B$ so equality follows. Also, for $x, y \in B$ the cosets $xA, yA$ in $A \cup B$ are distinct, whence if equality holds these exhaust $A \cup B$ and $A \cup B \subseteq BA$. For part c, $[A \cup B : A]$ and $[A \cup B : B]$ are relatively prime, and $[A \cup B : A][A : A \cap B] = [A \cup B : B][B : A \cap B]$, so $[A \cup B : A]$ divides $[B : A \cap B]$. By part b these are equal, whence $A, B$ commute. If $[G : AB]$ were greater than 1 then $[G : A], [G : B]$ would have a common factor. Thus $G = AB$ and the last claim follows by algebra from $[A \cup B : A \cap B] = [A \cup B : A][A : A \cap B]$.

The finiteness condition is clearly unnecessary for part a, and can be relaxed for the other two parts; see the exercises. Now, if $G$ has a Sylow base then for every subset $\pi$ of the prime divisors of $G$ there is a Hall $\pi$-subgroup, namely the join of the appropriate Sylow subgroups (this is a subgroup by part a of the lemma). On the other hand suppose that for every prime divisor $p$ of $G$ there is a Hall $p'$-subgroup; choose one such $C_p$. Let $S_p = \cap_{q \neq p} C_p$. It follows by the lemma that $\{S_p\}$ is a Sylow base.

7. A finite group $G$ is solvable iff $G$ has a Sylow base. Further in this case the following hold.

a. Any two Hall $\pi$-subgroups are conjugate.
b. Any π-subgroup is contained in a Hall π-subgroup.
c. Any two Sylow bases are conjugate.
d. Suppose \( m \) is the order of a Hall π-subgroup. The number of Hall π-subgroups may be expressed as a product of factors \( f_i \), where each \( f \) divides a chief factor and is congruent to 1 mod some prime factor of \( m \).

The proof that a group with a finite base is solvable uses the following theorem, due to Burnside.

**Theorem 17.** A group \( G \) of order \( p^aq^b \), \( p, q \) prime, is solvable.

**Proof:** A subgroup or quotient group of such a group is such a group. Also, if \( e = 0 \) or \( f = 0 \) \( G \) is solvable. Thus, we need only show that \( G \) is not simple when \( e, f > 0 \); we may also assume that \( G \) is not Abelian. Let \( H \) be a Sylow \( p \)-subgroup, and let \( h \neq 1 \) be in the center of \( H \); then \( H \) is contained in the centralizer \( C(h) \) of \( h \). If \( C(h) = G \) the center of \( G \) contains \( h \) and is therefore a nontrivial proper normal subgroup. Otherwise \( h_j = [G : C(h)] = q^n \) for some \( a > 0 \) where \( h \in C_j \), using the notation of section 15.7. Now, \( \sum_i n_i^{2}\equiv 0 \mod q \) and \( n_1 = 1 \), so for some \( i > 1, (q | n_i) \). Suppose \( \chi_{ij} = 0 \) whenever \( q \nmid n_i, i > 1 \); then

\[
0 = \chi_1 * \chi_i = 1 + \sum_{i=2}^s n_i \chi_{ij} = 1 + q\alpha
\]

where \( \alpha \) is an algebraic integer, which is impossible. Thus, there is an \( i > 1 \) with \( q \nmid n_i \) and \( \chi_{ij} \neq 0 \); \( \gcd(n_i, h_{ij}) = 1 \) follows readily, and the theorem follows by theorem 15.27.

**4. Regular rings.** A module is said to be Artinian if its lattice of submodules satisfies the descending chain condition. Exercise 8.3 is readily verified to hold, with Noetherian replaced by Artinian. By definition, a module \( M \) has finite length iﬀ its lattice of submodules does; by theorem 11 this is iﬀ it is Artinian and Noetherian. If \( M \) is semisimple, it has finite length iﬀ it is either Artinian or Noetherian, as is readily verified. There are modules which are Artinian but not Noetherian. Consider the polynomials \( F[x^{-1}] \) as an \( F[x] \) module, letting \( x^i x^{-i} = 0 \) if \( j > i \). The submodules are readily verified to be the \( F[x]x^{-i} \); letting \( i = 0, 1, \ldots \) yields an infinite ascending chain, but there are no infinite descending chains.

A proof of this theorem which does not use character theory may be found in [KuSt]. The remaining sections of this chapter concern the length of a ring \( R \), considered as an \( R \)-module. This section defines a type of ring where finite length implies semisimplicity. An Artinian ring is Noetherian; this is proved in section 6. The notion of the radical of a ring is needed for this; this is discussed in section 5.

A ring \( R \) is called regular if for all \( x \in R \) there is an element \( x' \) such that \( x = xx'x \). \( x' \) is a kind of pseudo-inverse; for example if \( x \) has an inverse then it may serve as \( x' \). The ring of endomorphisms of any vector space \( V \) over any division ring is regular. This follows because there is a linear map \( x'' \) from the image of \( x \) to \( V \), which is a right inverse to \( x; x' \) may be any endomorphism of \( V \) extending any such \( x'' \).

**Theorem 18.** A ring \( R \) is semisimple iﬀ it is regular and has finite length.

**Proof:** Suppose \( R \) is semisimple. Suppose \( x \in R \); then \( R = Rx \oplus L \) for some left ideal \( L \). Let \( e_1, e_2 \) be a complete system of idempotents determined by the decomposition, and let \( x' \) be such that \( e_1 x = x'x \). Then \( x = xx'x \), showing \( R \) to be regular. That \( R \) has finite length follows because it is a direct sum of finitely many simple left ideals, and a composition series is easily derived. Conversely suppose \( R \) is regular. Given a simple left ideal \( L \) and an element \( x \in L \), let \( x' \) be such that \( x = xx'x \) and let \( e = x'x \). Then \( e \) is an idempotent, and \( Re \subseteq Rx \subseteq L \), so since \( L \) is simple \( L = Re \). Further \( R = Re \oplus R(1 - e) \). If \( R \) is Noetherian, there is a simple left ideal \( L_1 \), and \( R = L_1 \oplus L' \) for some \( L' \). We may continue inductively; since the ring is Artinian (in fact, since it is Noetherian) we must reach \( 0 \) after a finite number of steps.
Suppose $R$ is a finite dimensional vector space over a division ring; then $R$ has finite length, because the submodules are subspaces and the dimension increases (decreases) in an ascending (descending) chain. It follows that $R$ is semisimple if it is regular. This need not be the case; for example let $R$ be the group algebra of a finite group $G$ over a field $F$, where the characteristic of $F$ divides the order of $G$. Let $x$ be the element $g_1 + \ldots + g_n$, where $G = \{g_1, \ldots, g_n\}$. It is readily verified that $xx'x = 0$ for all $x' \in F[G]$. To conclude this section, another property of regular rings is proved.

**Theorem 19.** If $R$ is a regular ring a finitely generated left ideal is principal.

**Proof:** First, a principal left ideal $Ra$ equals $Re$ where $e = a'a$ whenever $a = aa'a$; also $e$ is idempotent. Second, if $e, f$ are idempotents with $ef = fe = 0$ then $Re + Rf = R(e + f)$; clearly $R(e + f) \subseteq Re + Rf$, and $xe + yf = (xe + yf)(e + f)$. Given $Ra + Rb$, let $Ra = Re$ where $e$ is idempotent; then $b, be \in Re + Rb$, and $be, b - be \in Re + Rb(1 - e)$, so $Re + Rb = Rb(1 - e)$. Let $f_1$ be an idempotent such that $f_1 = b'(1 - e)$ for some $b'$; then $Rb(1 - e) = Rf_1$, and $f_1e = 0$. Let $f = (1 - e)f_1$; then $f \in Re + Rf_1$, and $f_1 \in Re + Rf$, so $Re + Rf = Re + Rf_1$. Finally $f^2 = f$ and $fe = ef = 0$ are readily verified, so $Ra + Rb = R(e + f)$. The theorem follows inductively.

5. The radical of a ring. The radical $\text{Rad}(R)$ of a ring $R$ is defined to be the intersection of all maximal left ideals of $R$; it is clearly a left ideal. It will be shown shortly that it is a two-sided ideal, and also the intersection of all maximal right ideals.

**Lemma 20.** $\text{Rad}(R)$ is the intersection of all annihilators of simple $R$-modules.

**Proof:** Let $S$ denote the intersection of the annihilators. First, we claim that if $M$ is a simple $R$-module and $m \in M$, then $\text{Ann}(m)$ is a maximal left ideal. Indeed, since $M$ is simple the map $r \mapsto rm$ is an epimorphism. Its kernel is $\text{Ann}(m)$, whence $M$ is isomorphic to $R/\text{Ann}(m)$; and since $M$ is simple $\text{Ann}(m)$ must be maximal. From this, if $r \in \text{Rad}(R)$ and $M$ is simple, then $r \in \text{Ann}(m)$ for any $m \in M$, whence $r \in \text{Ann}(M)$. This show that $\text{Rad}(R) \subseteq S$. Second, we claim that if $I \subseteq R$ is a maximal left ideal then $R/I$ is a simple $R$-module; this follows from $R(r + I) = Rr + I = R$ if $r \notin I$. So if $r \in S$ then $r(R/I)$ equals the 0 element of $R/I$, whence $rR \subseteq I$ and $r \in I$, which shows that $S \subseteq \text{Rad}(R)$.

**Lemma 21.** The following are equivalent.

a. $x \in \text{Rad}(R)$.

b. $1 + rx$ has a left inverse for all $r \in R$.

c. $1 + rx$ has a two-sided inverse for all $r \in R$.

**Proof:** If $1 + x$ has no left inverse then $1 + x \in I$ for some maximal left ideal $I$, and $x \notin I$ else $1 \in I$; thus, $x \notin \text{Rad}(R)$, and a $\Rightarrow$ b follows because $\text{Rad}(R)$ is a left ideal. If $x \notin \text{Rad}(R)$ then $x \notin I$ for some maximal left ideal $I$, and $Rx + I = R$, so $1 - rx \in I$ for some $r \in R$, and $1 - rx$ has no left inverse; that is, b $\Rightarrow$ a. Finally, suppose $x \in \text{Rad}(R)$ and let $a$ be such that $a(1 + x) = 1$, or $1 - a = ax$. Then $1 - a \in \text{Rad}(R)$, so $1 - (1 - a) = a$ has a left inverse, which must be $1 + x$ since this is a right inverse; thus, $a$ is a two-sided inverse for $1 + x$.

From lemma 20 we can conclude that $\text{Rad}(R)$ is a two-sided ideal, since it is the intersection of a family of such. By lemma 21, $\text{Rad}(R)$ is the largest two-sided ideal $I$ such that $1 + I \subseteq \text{Units}(R)$. Adapting the argument shows that this is also true of the intersection of all maximal right ideals, which thus equals $\text{Rad}(R)$.

**Theorem 22.** $\text{Rad}(R)$ is the smallest two-sided ideal $I$ such that $\text{Rad}(R/I) = 0$. 

180
Proof: If $J$ is the intersection of all maximal left ideals of $R$ which contain $I$ then $\text{Rad}(R/I)$ equals $J/I$. If $I = \text{Rad}(R)$ then $J = I$, so $\text{Rad}(R/I) = 0$. If $\text{Rad}(R) \nsubseteq I$ then $(\text{Rad}(R) + I)/I \neq 0$. We claim that $(\text{Rad}(R) + I)/I \subseteq \text{Rad}(R/I)$; it follows that if $\text{Rad}(R/I) = 0$ then $\text{Rad}(R) \subseteq I$. Now, if $M$ is an $R/I$-module it may be considered an $R$-module via the canonical epimorphism; we denote the $R$-module as $M'$. Since a submodule of $M'$ is a submodule of $M$ (because the homomorphism is surjective), if $M$ is simple $M'$ is. If $r$ annihilates $M'$ then $r + I$ annihilates $M$, and the claim follows using lemma 20.

Theorem 23. Suppose $R$ is commutative. Suppose $I \subseteq R$ is an ideal, and $M$ is a finitely generated $R$-module.

a. If $[IM] = M$ then for some $x \in I$, $(1 + x)M = 0$.

b. (Nakayama’s lemma) If also $I \subseteq \text{Rad}(R)$ then $M = 0$.

Proof: Part a is proved by induction on the number $t$ of generators; for the basis, $t = 0$ and $x$ may be taken as 0. For $t > 0$, suppose $M = \sum_{i=1}^{t} Ra_i$ and let $M' = M/Ra_i$. Clearly $[IM'] = M'$, and the $a_i + Ra_i$, for $i < t$ generate $M'$, so by induction $(1 + y)M' = 0$ for some $y \in I$. Then $(1 + y)M \subseteq Ra_i$, so $(1 + y)M = (1 + y)[IM'] \subseteq IRa_i = Ia_i$, and $(1 + y)a_i = za_i$ for some $z \in I$. We may now let $1 + x = (1 + y)(1 + y - z)$. Part b follows by lemma 21.

6. Artinian rings. In this section the notation $[S]$ for the ideal generated by $S$ will be used. An element $x \in R$ is called nilpotent if $x^n = 0$ for some $n \in \mathbb{N}$. A left, right, or two-sided ideal is called nil if its elements are all nilpotent. As a consequence of lemma 21, $\text{Rad}(R)$ contains any nil ideal $I$. Indeed, for every $x \in I$, $r \in R$, $(rx)^n = 0$ for some $n$, so $1 - rx$ has the inverse $\sum_{i=0}^{n-1}(rx)^i$. An ideal is called nilpotent if $[I^n] = 0$ for some $n \in \mathbb{N}$; nilpotent ideals are nil, but the converse need not hold.

Lemma 24. If $R$ is Artinian then $\text{Rad}(R)$ is nilpotent.

Proof: Write $I$ for $\text{Rad}(R)$; the sequence $I, [I^2], \ldots$ is eventually constant, say $[I^n] = [I^n]$ for $j \geq n$. Write $J$ for $[I^n]$, and suppose $J \neq 0$. We claim that there are nonzero $x, y \in J$ such that $yx = x$. Let $K$ be a minimal left ideal such that $K \subseteq J$ and $[JK] \neq 0$; there is such a $K$ since $J$ satisfies the requirements and $R$ is Artinian. Choose a nonzero $x \in K$ such that $Jx \neq 0$. Then $Jx \subseteq J$ and $[Jx] = Jx \neq 0$, and since $Jx \subseteq K$ $Jx = K$, and the required $y$ exists. But then $(1 - y)x = 0$, so $1 - y$ is a unit since $y \in \text{Rad}(R)$, which is a contradiction. Thus, $\text{Rad}(R)^n = 0$.

Suppose $R$ is Artinian. By the theorem and remarks following theorem 22, $\text{Rad}(R)$ is the largest nilpotent two-sided ideal. Also any nil ideal, whether left, right, or two-sided, is contained in $\text{Rad}(R)$ and is thus nilpotent.

If $R$ is commutative then $\text{Rad}(R)$ equals the intersection of the maximal ideals. If $x$ is nilpotent then $Rx$ is nil, and $x \in \text{Rad}(R)$. $\text{Rad}(R) = 0$ iff $R$ contains no nonzero nilpotent elements. Also, there are only finitely many maximal ideals. Indeed, if $[IJ] = I$ then $J \subseteq I$, so if $M_1, M_2, \ldots$ are maximal ideals then $M_1 \supseteq [M_1M_2] \supseteq \cdots$, so there can only be finitely many $M_i$. $\text{Rad}(R)$ equals $[M_1M_2 \cdots]$, by theorem 6.8.

As observed earlier, if $R$ is semisimple then $R$ is Artinian. It is also easy to see that $\text{Rad}(R) = 0$. If $R = \oplus_{j \neq i} L_j$ where the $L_j$ are simple left ideals, and $x \neq 0$, then $x = \sum x_j$ where $x_j \in L_j$ and, say, $x_i \neq 0$. Then $\oplus_{j \neq i} L_j$ is a maximal left ideal not containing $x$.

Lemma 25. If $R$ is an Artinian ring and $I \subseteq R$ is a left ideal which is not nilpotent then $I$ contains an idempotent.

Proof: Let $J$ be a minimal left ideal such that $J \subseteq I$ and $J$ is not nilpotent; $J$ exists since $I$ satisfies the requirements. Since $[J^2] \subseteq I$ is not nilpotent $[J^2] = J$. It follows as in lemma 24 that there are nonzero $x, y \in J$ with $yx = x$. Let $y_0 = y$, and inductively $w_i = y_i^2 - y_i$ and $y_{i+1} = y_i + w_i - 2y_iw_i$. We claim that
some $y_i$ is idempotent. First, $y$ is not nilpotent since $y^i x = x$ for all $i$. Second, $W = \{ w \in J : w x = 0 \}$ is a left ideal properly contained in $J$ (since $y \in J - W$), so $W$ is nilpotent; in particular $w_0$ is nilpotent. Inductively, if $y_{i+1}$ were nilpotent then $y_i$ would be, since it is a sum of commuting nilpotent elements; and $w_{i+1} = 4w_i^3 - 3w_i^2$ is nilpotent. Further $w_i$ is divisible by $w_0^2$. Hence $w_i$ equals 0 for some $i$, and $y_i^2 = y_i$.

**Theorem 27.** $R$ is semisimple iff $R$ is Artinian and $\text{Rad}(R) = 0$.

**Proof:** One direction was already observed. Conversely suppose $R$ is Artinian and $\text{Rad}(R) = 0$. If $L$ is a left ideal then it is not nilpotent, by the hypothesis that $\text{Rad}(R) = 0$. By lemma 25 $L$ contains an idempotent $e$. If $L$ is simple then $R = Le$. We may thus form a sequence $L_1, \ldots$ of simple left ideals, where there is a left ideal $K_i$ such that $R = \oplus_{j \leq i} L_i \oplus K_i$. The $K_i$ form a descending sequence, which must eventually reach 0.

**Exercises.**

1. Complete the following alternate proof of the Jordan-Hoelder theorem.

   a. If $G = G_0 \supset \cdots \supset G_n = \{0\}$ is a composition series and $H_0 \supset \cdots \supset H_r$ is any subnormal series of subgroups of $G$ then $r \leq n$. Hint: Induction on $n$. If $H_0 = G$, if $H_1 \subseteq G_1$ the claim is immediate; otherwise $H_1 G_1 / G_1 \cong H_1 / G_1 \cap H_1$, so the latter is simple. By induction $G_1 \cap H_1$ has a composition series, and this has length at most $n - 2$. $H_1$ thus has a composition series of length at most $n - 1$, so $r - 1 \leq n - 1$.

   b. The Jordan-Hoelder theorem holds. Hint: Induction on $n$. If $H_1 = G_1$ the claim is immediate; otherwise $G_1 \cap H_1$ is maximal in both. The composition series $G_0 \supset G_1 \supset G_1 \cap H_1 \cdots$ and $H_0 \supset G_1 \supset G_1 \cap H_1 \cdots$ have identical quotients.

   2. Prove lemma 7.

   3. Suppose $L$ is a lattice with least element 0. Suppose for each element $x$ there is a maximum value for the length of a chain from $x$ to 0; write $l(x)$ for this. Suppose finally that $l(x \cup y) = l(x) + l(y) - l(x \cap y)$. Show that $L$ is modular.

   4. Suppose $h : A \to B$ is a central homomorphism between two normal subgroups of a group $G$. Show that $u^{-1} \phi(u)$ is in the centralizer of $\phi[A]$. Prove the statement made about the central isomorphism of the Wedderburn-Remak-Schmidt theorem.

   5. Prove the following of the facts about solvable groups stated in the text.

   2. Hint: If $G = G_0 \supset \cdots$ is a cyclic normal series let $H_i = G_i \cap G'$. This series is cyclic and normal. The conjugations of $H_i / H_{i+1}$ by elements of $G / H_{i+1}$ form an Abelian group, whence the commutator of two such elements centralizes $H_i / H_{i+1}$.

   3. Hint: A quotient is the direct product of isomorphic simple groups. The quotients are also solvable.

   4. Hint: The center of a $p$-group is nontrivial.

   a$\Rightarrow$b, for any $G$. Hint: If $H \subseteq G$ let $i$ be such that $Z_i \subseteq H$, $Z_{i+1} \not\subseteq H$. Since $[H, Z_{i+1}] \subseteq Z_i$, for $z \in Z_{i+1} - H$ $[H, z] \subseteq H$ and so $z \in N(H)$.

   b$\Rightarrow$c, for any $G$.

   c$\Rightarrow$d. Hint: If $H$ is a Sylow subgroup then $N(H) = N(N(H))$, so $N(H) = G$. 182
5. d⇒a. Hint: Extend the above argument for item 4.

6. Show that for groups \( A \supseteq B \supseteq C \), \([A : C] = [A : B][B : C]\) where the equation is understood to be in the extended positive integers, so that \( x\infty = \infty \). Show that lemma 16 holds if \( A, B \) have finite index in \( G \).

7. Show that if \([I^n] = 0\) and \([J^m] = 0\) then \([I + J]^{n+m} = 0\). Hint: In the expansion of \( \prod x_i + y_i \) each term contains at least \( n \) \( x \)'s or \( m \) \( y \)'s. Suppose the former; each \( x \), multiplied by the \( y \)'s preceding it, is in \( I \).
17. Topological spaces.

1. Basic definitions. If $X$ is a set a topology on $X$ is a collection $T$ of subsets of $X$, called the open sets, such that

- $\emptyset, X \in T$;
- if $S \subseteq T$ then $\bigcup S \in T$;
- if $S \subseteq T$, $S$ finite, then $\bigcap S \in T$.

$T$ must be closed under arbitrary unions and finite intersections; the first requirement is redundant, considering $\emptyset$ to be the empty union, and $X$ the empty intersection. Alternatively, closure under arbitrary union and pairwise intersections, and $X \in T$, may be required.

A topological space is a set $X$, together with a topology on it. We use the usual notational convenience of denoting the space by $X$ when we don’t need a name for the topology. When we do need a name, we frequently use $T_X$. Given topological spaces $X, Y$, a function $f : X \rightarrow Y$ is said to be continuous if $f^{-1}[U] \in T_X$ whenever $U \in T_Y$. We say, $f^{-1}[U]$ is open whenever $U$ is, omitting the relevant space since this is clear. Since $(gf)^{-1}[C] = f^{-1}[g^{-1}[C]]$, the composition of continuous functions is continuous. The identity function from $X$ to $X$ is clearly continuous. Thus, the topological spaces and continuous functions form a (concrete) category, which we denote $\text{Top}$.

The topologies on a set $X$ are partially ordered by inclusion. A topology with more sets is called stronger, and one with fewer sets weaker. The intersection of any family of topologies is again a topology, as is readily verified. The topologies thus form a complete lattice, indeed a closure system. The weakest topology has only $\emptyset, X$ open and is called the indiscrete topology. The strongest topology has all sets open and is called the discrete topology.

Some authors reverse the meaning of the terms stronger and weaker; others prefer “finer” for more open sets, and “coarser” for fewer. There are other common variations of usage in topology; further examples will occasionally be noted.

If $S$ is any collection of subsets of $X$ there is a weakest topology $T$ containing the sets of $S$. This may be described as follows. Let $S_1$ be the intersection closure of $S$; let $T$ be the sets which are unions of sets in $S_1$. Certainly any topology containing $S$ contains $T$, so we need only show that $T$ is a topology. $T$ is certainly closed under union, and it is closed under finite intersection because $(\bigcup_i C_i) \cap (\bigcup_j D_j) = \bigcup_{i,j} (C_i \cap D_j)$. We may call $T$ the topology generated by $S$. Note that the join of a collection of topologies is the topology generated by their union.

$S$ is said to be a base for $T$ if $T$ is the unions of subcollections of $S$. It is easily seen that $S$ is a base for $T$ iff $\bigcup S = T$ and for all $U, V \in S$ and all $x \in U \cap V$ there is a $W \in S$ with $x \in W$. $S$ is said to be a subbase for $T$ if $T$ is generated by $S$, that is, if the intersection closure of $S$ is a base for $T$.

A set $K \subseteq X$ is called closed if $K^c = X - K$ is open. A topology may be given by giving the closed sets. The collection of closed sets is closed under arbitrary intersection and finite union. Writing co-$S$ for the complements of the sets in the collection $S$, $S$ is a called a base (subbase) for the closed sets co-$T$ if co-$S$ is a base (subbase) for the topology $T$.

It is easily seen that a function $f : X \rightarrow Y$ is continuous iff the inverse image of a closed set is closed. Also, $f$ is continuous iff the inverse image of any set in a subbase, or base, for the topology (closed sets) of $Y$ is open (closed).

An isomorphism in $\text{Top}$ is called a homeomorphism, that is, a homeomorphism is a bijection $f$ such that both $f$ and $f^{-1}$ are continuous. A function $f : X \rightarrow Y$ between topological spaces is called open (closed) if $f[S]$ is open (closed) whenever $S$ is open (closed). Clearly, a continuous bijection is a homeomorphism iff it is open iff it is closed.

Since the intersection of any family of closed sets is closed there is a smallest closed set $A^1$, called the
closure, containing any subset \( A \) of a topological space \( X \). Clearly, \( x \in A^{cl} \) iff any open set containing \( x \) intersects \( A \). A point \( x \in X \) is called a limit point of \( A \) if any open set containing \( x \) intersects \( A - x \). It is easy to see that if \( x \in A^{cl} \) then either \( x \in A \) or \( x \) is a limit point of \( A \).

The map \( A \mapsto A^{cl} \) is called the Kuratowski closure operator. The following facts are left to the reader (exercise 1a).

- \( A \) is closed iff \( A^{cl} = A \) iff \( A \) contains its limit points.
- If \( x \) is a limit point of \( A \cup B \) then \( x \) is a limit point of \( A \) or of \( B \).
- \( \emptyset^{cl} = \emptyset \), \( A \subseteq A^{cl} \), \( (A^{cl})^{cl} = A^{cl} \), \( (A \cup B)^{cl} = A^{cl} \cup B^{cl} \) (Kuratowski closure axioms).
- \( X^{cl} = X \), and if \( A \subseteq B \) then \( A^{cl} \subseteq B^{cl} \) (these follow from the axioms).
- If \( A \subseteq A^{cl} \) satisfies the axioms the sets \( A \) such that \( A^{cl} = A \) are the closed sets of a topology on \( X \), for which \( A \mapsto A^{cl} \) is the closure operator.
- \( f : X \mapsto Y \) is continuous iff \( f[A^{cl}] \subseteq (f[A])^{cl} \) for \( A \subseteq X \) iff \( (f^{-1}[B])^{cl} \subseteq f^{-1}[B^{cl}] \) for \( B \subseteq Y \).
- If \( U, V \) are disjoint open sets then \( U^{cl} \) and \( V \) are disjoint.

Given any subset \( S \subseteq X \), the union of the open sets contained in \( S \) is the largest open set contained in \( S \). This is called the interior of \( S \); we denote it \( S^{int} \). Clearly a set is open iff it equals its interior.

To conclude this section some additional commonly used definitions will be given. Suppose \( x \) is a point in a topological space \( X \). An open set containing \( x \) is called an open neighborhood of \( x \). A subset \( S \subseteq X \) is called a neighborhood of \( x \) if there is an open set \( U \) with \( x \in U \) and \( U \subseteq S \) (i.e., if \( x \in S^{int} \)). Some authors use the term “neighborhood” for “open neighborhood”.

Suppose \( X \) is a topological space, and \( \{A_j\} \) is a collection of subsets. \( \{A_j\} \) is said to cover \( X \), or be a cover, if \( X = \bigcup_j A_j \). More generally \( \{A_j\} \) covers a subset \( S \subseteq X \) if \( S \subseteq \cup_j A_j \). A cover is called open if its sets are open, and closed if its sets are closed. Suppose \( C \) is a cover of a space \( X \). If \( C' \) is a subset of \( C \) and \( C' \) is a cover then \( C' \) is said to be a subcover of \( C \). If \( C' \) is such that for all \( S' \in C' \) there is an \( S \in C \) with \( S' \subseteq S \), and \( C' \) is a cover, then \( C' \) is said to be a refinement of \( C \).

A collection \( \{A_j\} \) of subsets of a topological space \( X \) is said to be locally finite if every \( x \in X \) has an open neighborhood which intersects only finitely many \( A_j \). \( \{A_j\} \) is said to be point finite if each \( x \in X \) is a member of only finitely many \( A_j \) (this definition applies for \( \{A_j\} \) a collection of subsets of a set \( X \)).

2. Induced and coinduced topologies. Suppose \( \{X_j\} \) is a collection of topological spaces, with \( T_j \) the topology on \( X_j \), and \( Y \) is a set. If functions \( f_j : Y \mapsto X_j \) are given, the topology on \( Y \) generated by \( \{f_j^{-1}[U] : U \in T_j\} \) is clearly the weakest making all the \( f_j \) continuous. It is called the topology induced by the \( f_j \). If functions \( f_j : X_j \mapsto Y \) are given, the collection \( \{U : f_j^{-1}[U] \in T_j, \text{ all } j\} \) is clearly the strongest topology on \( Y \) making all the \( f_j \) continuous. It is called the topology coinduced by the \( f_j \).

Observe that the induced topology is the join of those induced by each \( f_j \), and the coinduced topology the meet. If there is only one function \( f \), the induced topology equals \( \{f^{-1}[U] : U \in T_X\} \), since the latter is a topology. Considering \( f^{-1} \) for \( f : Y \mapsto X \) as mapping the collection of subsets of \( X \) to that of \( Y \), the induced topology is the image of \( T_X \), and the coinduced topology the inverse image of \( T_Y \).

**Theorem 1.**

a. Suppose \( Y \) is equipped with the topology induced by the functions \( \{f_j : Y \mapsto X_j\} \). A function \( g : Y' \mapsto Y \) is continuous iff each \( f_jg \) is.

b. Suppose \( Y \) is equipped with the topology coinduced by the functions \( \{f_j : X_j \mapsto Y\} \). A function \( g : Y \mapsto Y' \) is continuous iff each \( gf_j \) is.

**Proof:** One direction is immediate. For the other, for part a, \( g^{-1}[f_j^{-1}[U]] \) is open for \( U \) an open subset of \( X_j \). Since \( \{f_j^{-1}[U]\} \) is a subbase for the topology on \( Y \), \( g \) is continuous. For part b, if \( U \) is open in \( Y' \) then \( f_j^{-1}[g^{-1}[U]] \) is open for all \( j \), whence \( g^{-1}[U] \) is open.

185
Corollary 2. Suppose $G$ is the forgetful functor from Top to Set, and $H \in \text{Top}^J$ for some small category $J$. If $(Y, f_j)$ is a limit (colimit) cone in $\text{Set}$ for $GH$, and $Y$ is equipped with the topology induced (coinduced) by the $f_j$, then $(Y, f_j)$ is a limit (colimit) cone in Top for $H$.

Proof: By the theorem, the map induced in Set is continuous.

The property of theorem 1a in fact characterizes the induced topology. Since $\iota_Y$ is continuous each $f_\alpha$ is. If $Y$ is not the induced topology let $Y'$ be $Y$ with a weaker topology which makes the $f_\alpha$ continuous; the identity map would not be continuous from $Y'$ to $Y$, a contradiction. A similar claim holds for theorem 1b.

If $Y \subseteq X$ the topology induced on $Y$ by the inclusion map is called the subspace topology, and is said to be inherited from $X$. $Y$ is called a subspace when it is equipped with this topology. $Y$ may be said to have a property of subsets (e.g., open subspace), or of spaces (e.g., compact subspace; see below). The following facts are left to the reader (exercise 1b).

- $T_Y = \{ U \cap Y : U \in \mathcal{T}_X \}$.
- The closed sets of $Y$ are those of the form $K \cap Y$ where $K$ is closed in $X$.
- The collection $\{ U \cap Y \}$, where $U$ ranges over a base (subbase) for $\mathcal{T}_X$ forms a base (subbase) for $T_Y$.
- If $S \subseteq Y$ is open (closed) in $X$ then $S$ is open (closed) in $Y$.
- If $Y$ is open then $S \subseteq Y$ is open in $X$ iff $S$ is open in $Y$; and $T_Y \subseteq \mathcal{T}_X$.
- If $Y$ is closed then $S \subseteq Y$ is closed in $X$ iff $S$ is closed in $Y$; and co-$T_Y \subseteq \text{co-} \mathcal{T}_X$.
- If $Z \subseteq Y$ then the topologies $Z$ inherits from $Y$ and $X$ are the same.
- If $f : X \rightarrow Y$ is continuous and $X' \subseteq X$ is a subspace then $f| X'$ is continuous.
- The coimage-image factorization in $\text{Set}$ of $f : X \rightarrow Y$ is a coimage-image factorization in Top.

If the index category $J$ of corollary 2 is discrete, the limit is as usual called the product, and denoted $\times_J X_j$. The sets $U \times (\times_{j \in K} X_j), U \in \mathcal{T}_I$, form a subbase for the product topology. The sets $(\times_{j \in K} U_j) \times (\times_{j \in K} X_j)$ for $K$ a finite set of indices and $U_j \in \mathcal{T}_I$ for $j \in K$ form a base. $U_j$ may be restricted to be from a base or subbase for $T_j$. If $J$ is not discrete, the limit is a subset of the Cartesian product, and it is easily seen that its topology is that inherited from the product topology.

The projection map $\pi_j : \times_J X_j \rightarrow X_j$ is readily seen to be an open map. Suppose $U$ is a basic open set, equal to $U_j$ for $j \in F$ for $F$ a finite set of indices, and $X_j$ otherwise. Then $\pi_j[U]$ is either $U_j$ or $X_j$, depending on whether $j \in F$ or not.

Given a set $X$, an equivalence relation $\equiv$ yields the canonical epimorphism $\eta : X \twoheadrightarrow X/\equiv$ where $X/\equiv$, called the quotient, is the set of equivalence classes. A surjection $h : X \twoheadrightarrow Y$ yields the equivalence relation $h(x) = h(y)$; $Y$ is canonically isomorphic (via the map $y \mapsto h^{-1}(y)$, with inverse $[x] \mapsto \eta(x)$) to the quotient, and may be identified with it. A subset $S \subseteq X$ is called saturated if it is a union of equivalence classes. The saturated sets are closed under union, intersection, and difference. The map $S \rightarrow \bigcup S$ is a bijection from the subsets of $X/\equiv$ to the saturated subsets of $X$, and $\eta[\bigcup S] = S$. For $R \subseteq X$, the smallest saturated set containing $R$ (called the saturation of $R$) is the union of the equivalence classes which intersect $R$; this is easily seen to equal $\eta^{-1}\eta[R]$. $R$ is saturated iff $R = \eta^{-1}\eta[R]$, iff $R = \eta^{-1}[S]$ for some $S \subseteq X/\equiv$.

If $X$ is a topological space and $\equiv$ is an equivalence relation on $X$, the topology on $X/\equiv$ coinduced by the canonical epimorphism $\eta$ is called the quotient topology, and $X/\equiv$ is called the quotient space with this topology. The following facts are left to the reader (exercise 1c).

- $S \subseteq X/\equiv$ is open in $X/\equiv$ iff $\bigcup S$ (S being a set of equivalence classes) is open in $X$, iff $S = \eta[R]$ for some saturated open set $R \subseteq X$.
- $\eta$ is open (closed) iff whenever $S$ is open (closed) then its saturation is open (closed).
- Suppose $f : X \rightarrow Y$ is continuous and surjective and let $\equiv$ be the induced equivalence relation. Then $Y$ is homeomorphic to $X/\equiv$ iff $Y$ has the topology coinduced by $f$, iff $f[S]$ is open (closed) for every saturated open (closed) set $S$. 

186
- Under the same hypotheses, if \( f \) is open or closed then \( Y \) has the coinduced topology.

The quotient has the universality property that if \( f : X \mapsto Y \), and \( x_1 \equiv x_2 \) implies \( f(x_1) = f(x_2) \), then \( f \) factors uniquely through the canonical epimorphism \( \eta : X \mapsto X/\equiv \). Indeed, the map must be \( [x] \mapsto f(x) \), which is well-defined. It is also continuous, by theorem 1.

Remarks made in section 4.5 regarding the product of structures apply to the product of topological spaces; namely, in \( \times_i X_i \) (writing \( x \) for an element of the product and \( x_i \) for the \( i \)th component) the relation \( x_i = y_i \) is an equivalence relation, and the quotient is homeomorphic to \( X_i \). Indeed, \( x \mapsto x_i \) respects the relation, so \( [x] \mapsto x_i \) is continuous. It is also open, because \( x \mapsto x_i \) is, and the image of an open set under the former map is the image of a saturated open set under the latter. Finally the \( w \mapsto [x] \) where \( x_i = w \) is the inverse.

If the index category of corollary 2 is discrete then the colimit is called the topological sum. We write \( \oplus_j X_j \) for this. The underlying set is the disjoint union of the \( X_j \), and the disjoint union of the \( T_j \)'s forms a base for the topology. The injection from a summand to the sum is clearly both open and closed. A direct limit in \( \text{Set} \) is a quotient of the direct sum; it is easily seen that a direct limit in \( \text{Top} \) has the quotient topology.

The map taking a set \( X \) to the discrete space is readily seen to be the object function of a left adjoint to the forgetful functor. Indeed, the identity map on \( X \) is a universal arrow from \( X \) to the forgetful functor, since any function \( f : X \mapsto Y \) is continuous when \( X \) has the discrete topology and \( Y \) has any topology.

The map taking a set \( X \) to the indiscrete space is readily seen to be the object map of a right adjoint to the forgetful functor. Indeed, the identity map from a space \( X \) to the indiscrete space \( X_0 \) on the same set is continuous; and any continuous map from \( X \) to an indiscrete space \( Y \) is continuous as a map from \( X_0 \) to \( Y \).

3. The coherent topology. Say that a cover \( \{A_j\} \) of a topological space \( X \) is generating if the topology on \( X \) is coinduced by the inclusion maps from the \( A_j \) equipped with their subspace topologies. If \( \{A_j\} \) is a generating cover \( X \) is said to have the topology coherent with the \( A_j \). The following are left to the reader (exercise 1d).

- \( \{A_j\} \) is generating iff whenever \( U \cap A_j \) is open in \( A_j \) for all \( j \) then \( U \) is open, iff whenever \( U \cap A_j \) is closed in \( A_j \) for all \( j \) then \( U \) is closed.
- If \( \{A_j\} \) is generating and for all \( j \) there is a \( k \) with \( A_j \subseteq B_k \) then \( \{B_j\} \) is generating. In particular this is so if \( \{A_j\} \subseteq \{B_k\} \),
- If \( \{A_j\} \) is generating and for given \( j \) \( \{B_{jk}\} \) is generating for \( A_j \) then \( \{B_{jk}\} \) is generating for \( X \).
- Any open cover is generating.
- Any finite closed cover is generating.

The last fact can be generalized. For any locally finite family \( \{A_j\} \), \( \cup_j A_j^{\text{cl}} = \cup_j A_j^{\text{cl}} \). To see this, suppose \( x \) is a limit point of \( \cup_j A_j \). Then there is an open neighborhood of \( x \) which intersects only finitely many \( A_j \), and it is easily seen that \( x \) is a limit point of \( A_j \) for one of these \( j \). A similar argument shows that any locally finite closed cover is generating.

**Theorem 3.** Suppose \( \{A_j\} \) is a generating cover of \( X \), \( f : X \mapsto Y \), and \( f \restriction A_j \) is continuous for all \( j \); then \( f \) is continuous.

**Proof:** If \( V \subseteq Y \) is open then \( f^{-1}[V] \cap A_j \) is open in \( A_j \) for all \( j \).

**Theorem 4.** Let \( f : \bigoplus_j A_j \mapsto X \) be the map determined in \( \text{Top} \) by the inclusion maps \( A_j \subseteq X \). Then \( X \) has the topology coinduced by the inclusion maps iff it has the topology coinduced by \( f \).

**Proof:** This follows because for \( U \subseteq X \), \( f^{-1}[U] \) is open iff \( A_j \cap U \) is open in \( A_j \) for all \( j \). Indeed, \( f^{-1}[U] = \bigoplus_j (A_j \cap U) \), and \( A_j \) is open in \( \bigoplus_j A_j \).
Theorem 5. Suppose \( \{A_j\} \) is a set of subsets of some set \( X \), each equipped with a topology. Suppose \( \{A_j\} \) is closed under intersection, \( A_j \cap A_k \) has the topology inherited from either \( A_j \) or \( A_k \), and \( A_j \cap A_k \) is closed (open) in both \( A_j, A_k \). Equip \( X \) with the topology coinduced by the inclusion maps. Then \( A_j \) has the inherited topology, and \( A_j \) is closed (open) in \( X \).

Proof: Suppose \( K \subseteq A_j \) is closed in \( A_j \); then by the hypotheses \( K \cap A_k \) is closed in \( A_k \) for all \( k \), whence \( K \) is closed in \( X \). In particular \( A_j \) is closed in \( X \). The argument for open intersections is similar.

4. Properties of topological spaces. In this section and the next properties of topological spaces will be considered. These properties will be “topological”, where a property of topological spaces is topological if whenever \( X \) and \( Y \) are homeomorphic, \( X \) has the property iff \( Y \) does. An example of a non-topological property would be “convex”, for subspaces of \( \mathbb{R}^n \) (convexity is considered in chapter 22).

Let \( X \) be a topological space. \( X \) is called second countable if its topology has a countable base. For \( x \in X \) a neighborhood base at \( x \) is a collection \( C \) of open sets such that given any open set \( V \) containing \( x \) there is a \( U \in C \) such that \( x \in U \) and \( U \subseteq V \). \( X \) is called first countable if \( x \in X \) there is a countable neighborhood base at \( x \). \( X \) is called Lindelof if every open cover has a countable subcover.

Theorem 6 (Lindelof’s theorem). A second countable space is a Lindelof space.

Proof: Let \( B \) be a countable base and \( C \) a cover of the space \( X \). For each \( x \in X \) choose a set \( S \in B \) containing \( x \) and a subset of a set of \( C \). For each such \( S \), of which there are only countably many, choose a superset in \( C \). The resulting subcollection of \( C \) is countable, and covers \( X \).

Corollary 7. In a second countable space any uncountable set \( S \) has a limit point.

Proof: If \( S \) has no limit point then for each \( x \in S \) we may choose a basic open set containing \( x \) but no other point of \( S \).

Let \( S \subseteq X \) be a subset. \( S \) is called dense if every open set contains elements of \( S \). \( X \) is called separable if it has a countable dense subset. A second countable space is separable; take a countable base and choose an element of each set in the base. The resulting set is clearly countable, and is dense because any open set contains a set in the base.

A topological space \( X \) is said to be compact if every open cover has a finite subcover. As mentioned above, a subset \( S \subseteq X \) is called compact if it is a compact space with the subspace topology. It is easily seen that \( S \) is compact iff every open cover of \( S \) has a finite subcover. \( X \) is called locally compact if for every \( x \in X \) there is an open set \( U \) with \( x \in U \) and \( U^{cl} \) compact. \( X \) called paracompact if every open cover has a locally finite refinement.

We leave the following facts to the reader (exercise 1e).
- A closed subspace of a compact space is compact.
- The union of finitely many compact subspaces is compact.
- If \( Y \subseteq X \) is a subspace then \( S \subseteq Y \) is compact in \( Y \) iff it is compact in \( X \).
- If \( X \) is compact and \( f : X \rightarrow Y \) is continuous then \( f[X] \) is compact.
- A compact space is locally compact and paracompact.
- \( X \) is locally compact iff every point has a compact neighborhood.
- A closed subspace of a locally compact space is locally compact.
- If \( X \) is locally compact and \( K \subseteq X \) is compact then there is an open set \( U \) with \( K \subseteq U \) and \( U^{cl} \) compact.

Say that subsets \( A, B \subseteq X \) of a topological space can be separated if there are disjoint open sets \( A', B' \) with \( A \subseteq A', B \subseteq B' \).
- \( X \) is \( T_0 \) if given distinct points \( x, y \) there is an open set containing one but not the other.
- $X$ is $T_1$ if given distinct points $x, y$ there is an open set containing $x$ but not $y$.
- $X$ is $T_2$ or Hausdorff if distinct points $x$, $y$ can be separated.
- $X$ is regular if a closed set $A$ and a point $x \notin A$ can be separated.
- $X$ is $T_3$ if it is $T_1$ and regular.
- $X$ is completely regular if given a closed set $A$ and a point $x \notin A$ there is a continuous function $f : X \mapsto [0, 1]$ such that $f(w) = 0$ for $w \in A$ and $f(x) = 1$ (where $[0, 1]$ denotes the closed interval $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$).
- $X$ is Tychonoff if it is $T_1$ and completely regular.
- $X$ is normal if disjoint closed sets can be separated.
- $X$ is $T_4$ if it is $T_1$ and normal.

Variations in usage include the following. Regular, completely regular, or normal spaces are required to be $T_1$. Paracompact spaces are required to be Hausdorff, or regular.

The implications

$$T_4 \Rightarrow T_3 \Rightarrow \text{Tychonoff} \Rightarrow T_2 \Rightarrow T_1 \Rightarrow T_0,$$

normal $\Rightarrow$ completely regular $\Rightarrow$ regular

hold. To see that a completely regular space is regular, for small $\epsilon$ the intervals $[0, \epsilon)$ and $(1 - \epsilon, 1]$ are open in $[0, 1]$ and disjoint, and their inverse images separate $A$ and $x$. That a normal space is completely regular follows from the next theorem; the remaining implications are virtually immediate.

**Theorem 8 (Urysohn’s Lemma).** $X$ is normal iff, given disjoint closed sets $A, B$ there is a continuous function $f : X \mapsto [0, 1]$ such that $f(w) = 0$ for $w \in A$ and $f(w) = 1$ for $w \in B$.

**Proof:** Suppose $X$ is normal; for each rational $r$ of the form $p/2^n$ for $n \geq 0$ and $0 \leq p \leq 2^n$, a set $A_r$ will be defined, open if $r \neq 0$. Let $A_0 = A$ and $A_1 = B^c$. Inductively, if $t = (2p + 1)/2^{n+1}$ let $r = p/2^n$ and $s = (p + 1)/2^n$, and choose $A_t$ such that $A^t_o \subseteq A_t$ and $A^t_1 \subseteq A_s$. Now define $f(x) = 0$ for $x \in A$, and for $x \in X - A_0$ let $f(x) = \sup \{t : x \notin A_t\}$. Clearly $0 \leq f(x) \leq 1$, and $f(x) = 1$ for $x \in B$. Now, $f(x) > a$ iff $x \in A_t$ for some $t < a$, whence $f^{-1}([0, a]) = \cup t < a A_t$, so $f^{-1}([0, a])$ is open. Also $f(x) > a$ iff $x \notin A_t$ for some $t > 0$, whence $f^{-1}((a, 1]) = \cup t > a (A^t_1)^c$, so $f^{-1}((a, 1])$ is open. Since $\{0, a\} \cup \{(a, 1]\}$ form a base for the topology on $[0, 1]$, $f$ is continuous. In the other direction, take a small $\epsilon$ as in the remarks preceding the theorem.

**Corollary 9 (Tietze Extension Theorem).** If $X$ is a normal space, $K$ is a closed subset, and $f : K \mapsto [-1, 1]$ be a continuous function. Then there is a continuous function $g : X \mapsto [-1, 1]$ such that $g|K = f$.

**Proof:** Let $K_0 = \{x : f(x) \leq -1/3\}$, $L_0 = \{x : f(x) \geq 1/3\}$. Then $K_0, L_0$ are closed and disjoint, so by Urysohn’s lemma there is a continuous function $g_0 : X \mapsto [-1/3, 1/3]$ which is $-1/3$ on $K_0$ and $1/3$ on $L_0$. Let $f_1 = f - g_0|K$; then $f_1 : K \mapsto [-2/3, 2/3]$. Continuing in this way we may obtain a sequence of functions $g_i : X \mapsto [-1/(3i)(2/3)^n, (1/3)(2/3)^n]$ such that, letting $f_i = f - \sum_{j<i} g_j|K$, $f_i : K \mapsto [-1/(3i)(2/3)^n, (1/3)(2/3)^n]$ and we leave the remainder of the proof as an exercise 8.

The converse of corollary 9 holds. If $K$ and $L$ are disjoint closed subsets of $X$ let $f$ be the function on $X$ which is $-1$ on $A$ and $1$ on $B$. If this has an extension $g : X \mapsto [-1, 1]$ then $A$ and $B$ can be separated as usual.

We leave the following to the reader (exercise 1f).
- A space is $T_1$ iff for any point $x$, $\{x\}$ is closed.
- A space is regular iff, for $U$ open and $x \in U$ there is an open $V$ with $x \in V$ and $V^c \subseteq U$.
- A subspace of a $T_0$ space is $T_0$, and similarly for $T_1$, $T_2$, regular, and completely regular spaces.
- A closed subspace of a normal space is normal.

189
- The product of a family of $T_0$ spaces is $T_0$, and similarly for $T_1$, $T_2$, regular, and completely regular spaces.

- If $X$ is any space, $Y$ is Hausdorff, and $f, g : X \to Y$ are continuous then $\{ x \in X : f(x) \neq g(x) \}$ is open.

Thus if $f, g$ agree on some set $S$ they agree they agree on $S^c$.

**Lemma 10 (shrinking lemma).** Suppose $\{ U_j : j \in J \}$ is a point finite open cover of a normal space $X$. Then there is an open cover $\{ V_j \}$ where $V_j^c \subseteq U_j$.

**Proof:** Say that a partial function $\phi$ on $J$ is a partial shrinking if $\phi(j)$ is an open set whose closure is contained in $U_j$, and the sets $\phi(j)$ for $j \in \text{Dom}(\phi)$, together with the sets $U_j$ for $j \notin \text{Dom}(\phi)$, form a cover of $X$. A partial shrinking is a partial function. We claim that the union of a chain of partial shrinkings is a partial shrinking. If $x \in X$ is such that every $U_j$ with $x \in U_j$ is shrunk by some point of the chain, then by local finiteness there is some point where they are all shrunk, so $x$ must be in one of the $\phi(U_j)$ at that point. By the claim Zorn’s lemma may be applied; let $\phi$ be a maximal partial shrinking. Suppose $U_k$ is not in the domain. Then $C = U_k - \cup_{j \neq k} U_j = X - \cup_{j \neq k} U_j$ is closed. By normality there is a $V_k$ with $C \subseteq V_k \subseteq V_k^c \subseteq U_k$. This is a contradiction, because $\phi$ could be extended by assigning $V_k$ to $U_k$.

**Corollary 11.** Under the same hypotheses, there are functions $f_j : U_j \to [0,1]$, such that $f_j(x) = 0$ if $x \notin U_j$, and for each $x \in X$, $\sum_j f_j(x) = 1$ (where by hypothesis the sum is finite).

**Proof:** By the shrinking lemma and Urysohn’s lemma, choose a function $g_j$ which is 1 on a closed subset of $U_j$ and 0 outside $U_j$. Then set $f_j(x) = g_j(x)/\sum_j g_j(x)$.

A collection $\{f_i\}$ of continuous functions from $X$ to $[0,1]$ such that at any $x$ only finitely many $f_i$ are nonzero, and $\sum f_i(x) = 1$, is called a partition of unity. If for a cover $\{U_j\}$, for each $i$ there is a $j$ with $\text{Supp}(f_i) \subseteq U_j$ (where $\text{Supp}(f)$, the support of $f$, is $\{ x : f(x) \neq 0 \}$), the partition is said to be subordinate to the cover. Given a cover of a paracompact Hausdorff space, there is a partition of unity subordinate to it, since the space is normal and the cover has a locally finite (so certainly point finite) refinement. In more specialized settings the functions $f_i$ may be required to have additional properties, in particular on smooth manifolds (see below) they may be required to be smooth; we omit further discussion of this topic (see [Spivak]).

**Lemma 12.**

a. In a Hausdorff space disjoint compact subsets can be separated.

b. A compact subspace of a Hausdorff space is closed.

c. If $X$ is a locally compact Hausdorff space, $U$ is open, and $K \subseteq U$ is compact, then there is an open set $V$ with $K \subseteq V$, $V^c \subseteq U$, and $V^c$ compact.

**Proof:** We first prove part a for a compact set $K$ and a point $x \notin K$. For each $w \in K$ choose disjoint open sets $U_w, V_w$ with $w \in U_w$, $x \in V_w$. There is a finite set of $U_w$ covering $K$, and the intersection of the corresponding $V_w$ contains $x$ and is disjoint from the union of the $U_w$. Now let $L$ be a compact set disjoint from $K$. For each $x \in L$ choose disjoint open sets $U_x, V_x$ with $K \subseteq U_x$, $x \in V_x$. Choose a finite subset of the $V_x$ covering $L$ and proceed similarly. Part b follows, because a single point is compact and so can be separated from a compact set; and thus the complement of a compact set is open. We first prove part c for $X$ compact; $K$ and $X - U$ are disjoint compact sets, so are contained in disjoint open neighborhoods $V, W$. It follows that $V^c$ has the required properties. If $X$ is locally compact, choose an open $U_0$ with $K \subseteq U_0$ and $U_0^c$ compact and replace $U, X$ by $U \cap U_0, (U \cap U_0)^c$.

We note also that a continuous function from a compact space to a Hausdorff space is closed; this follows using exercise 1e. If the function is bijective it is a homeomorphism and both spaces are compact Hausdorff.
**Theorem 13.**

a. If $X$ is Lindelöf and regular then $X$ is normal.

b. If $X$ is locally compact and Hausdorff then $X$ is $T_3$.

c. If $X$ is paracompact and regular then $X$ is normal.

d. If $X$ is paracompact and Hausdorff then $X$ is $T_4$.

**Proof:** For part a, let $K_i$, $i = 0, 1$, be disjoint closed sets. Let $C_i = \{ U : U$ open, $U^{cl} \cap K_{1-i} = \emptyset \}$; since $X$ is regular $C_i$ is an open cover of $K_i$, and since $X$ is Lindelöf and $K_i$ is closed, $C_i$ has a countable subcover $C_i'$ of $K_i$. Write $C_i'$ as $\{ U_{ij} \}$, and let $V_{ij} = U_{ij} - \cup_{l \leq n} U_{1-i,l}^{cl}$. One readily verifies that $V_0$ and $V_1$ are disjoint for all $j, \bar{j}$; also $\{ V_i \}$ is still a cover of $K_i$ so letting $V_i = \cup_j V_{ij}$, $V_0$ and $V_1$ are disjoint and $K_i \subseteq V_i$. For part b, if $K$ is closed and $x \notin K$, then $K^{c}$ is open, so there is an open neighborhood $V$ of $x$ with $V^{cl}$ disjoint from $K$; then $(V^{cl})^{c}$ and $V$ separate $K$ and $x$. For part c, given closed subsets $K, L$, for each $x \in K$ choose an open set $U_x$ with $U_x^{cl} \cap L = \emptyset$. Similarly choose $V_y$ for each $y \in L$. Then $\{ U_x \} \cup \{ V_y \} \cup \{ (K \cup L)^{c} \}$ is an open cover of $X$, so there is a locally finite refinement. The union of the closures of the $V'$ in the refinement for which $V' \subseteq V_y$ for some $y$ may be subtracted from each $U'$ where $U' \subseteq U_x$ for some $x$; The modified $U'$ are still open and still cover $K$. Taking a similar cover of $L$, the unions of these two covers are disjoint. For part d, by part c it suffices to show that $X$ is regular. Let $K$ be closed and $x \notin K^{c}$. For each $w \in K$ choose disjoint open sets $U_w, V_w$ with $w \in U_w$, $x \in V_w$. The open cover $\{ K^{c} \} \cup \{ U_w \}$ has a locally finite refinement $\{ W \} \cup \{ U' \}$ where each $U'$ is a subset of some $U_w$ and $x \in W$. $W$ intersects finitely many $U'$; each of these is contained in some $U_w$. Let $V$ be the intersection of the corresponding $V_w$ and $W$. Let $U$ be the union of all the $U'$. Then $U, V$ separate $K, x$.

By part d if $X$ is compact and Hausdorff then $X$ is $T_4$; this also follows directly using lemma 12a.

An infinite sequence in a topological space is a sequence of points $x_i$, for $i \in \mathbb{N}$. A point $x$ is said to be a limit of the infinite sequence $\langle x_i \rangle$ if every neighborhood of $x$ contains all but finitely many points of the sequence. In this case, the sequence is said to converge to $x$.

**Theorem 14.**

a. If $\langle x_i \rangle$ is a sequence in $X - \{ x \}$ which converges to $x$ then $x$ is a limit point of $\{ x_i \}$.

b. Suppose $S$ is a countable set of points and $x$ is a limit point of $S$; then if there is a countable neighborhood base at $x$ then there is a sequence $\langle x_i \rangle$ in $S - \{ x \}$ which converges to $x$. In particular this follows if $X$ is first countable.

c. In a Hausdorff space a sequence can converge to at most one value.

**Proof:** Part a is immediate from the definitions. For part b, enumerate the neighborhood base as $U_i$, and let $U_i' = \cap_{j \leq i} U_j$. Let $y_i$ be any point in $S \cap U_i' - \{ x \}$. For part c, given distinct points there are disjoint open sets containing them, and a sequence can lie in only one of these all but finitely often.

A topological space is said to be sequentially compact if any infinite subset has a limit point.

**Theorem 15.**

a. If $X$ is compact then $X$ is sequentially compact.

b. If $X$ is second countable and sequentially compact then $X$ is compact.

**Proof:** For part a, suppose $X$ is compact and $S \subseteq X$ has no limit point. For each $x \in X$ choose an open neighborhood $U_x$ of $x$ such that $U_x \cap S = \{ x \}$. The resulting cover has a finite subcover, so $S$ is finite. For part b, let $C$ be a cover, which we may assume is countable, say $\{ U_i \}$. Let $V_0 = U_0$, and inductively let $V_i$ be the first $U_j$ not covered by $V_j$, $j < i$. If this process terminates after a finite number of steps we have found a finite subcover. Otherwise for each $i$ choose a point $x_i$ in $V_i$ but no $V_j$, $j < i$. The resulting infinite set has a limit point, say $x$, which is in some $V_i$ since the $V_i$ clearly form a cover. This is a contradiction, because no $x_j$ with $j > i$ can be in $V_i$, but by theorem 14 there must be such $x_j$. 191
A space is called \(\sigma\)-compact if it has a countable cover by compact subspaces; equivalently, if it is the union of a countable ascending chain of compact subspaces.

A space is called connected if it is not the disjoint union of two nonempty open subsets. Connected spaces are considered in the next section, but it is convenient to introduce them here for the following lemma. Note that if a nonempty subspace of a connected space is both open and closed then it is the entire space.

**Lemma 16.** If \(C\) is a locally finite cover of a connected space then \(C\) is countable (or finite).

**Proof:** Let \(D_0\) be any member of \(C\), and let \(D_{n+1}\) be \(D_n\) with the sets of \(C\) which intersect sets of \(D_n\) added. Let \(D = \bigcup D_n\), and let \(Y = \bigcup D\). \(Y\) is certainly open. If \(U \in C\) intersects \(Y\) then it is in \(D\), and it follows that \(Y\) is closed also. Hence \(Y = X\).

**Theorem 17.** Suppose \(X\) is a locally compact Hausdorff space.

a. If \(X\) is second countable then \(X\) is \(\sigma\)-compact.

b. If \(X\) is \(\sigma\)-compact \(X\) is paracompact.

c. If \(X\) is connected and paracompact then \(X\) is \(\sigma\)-compact.

**Proof:** For part a, let \(U_i\) be a countable base; we claim that the \(U_i\) for which \(U_i^c\) is compact are still a base, and taking the closures of these yields the theorem. Let \(V\) be open and \(x \in V\). Then since \(X\) is locally compact Hausdorff there is a compact neighborhood of \(x\) in \(V\), which proves the claim. For part b, we first claim that there is a cover by compact subspaces \(K_i\) where \(K_i\) is contained in the interior of \(K_{i+1}\). Let \(L_i\) be a cover by compact subspaces. Let \(K_0 = L_0\). Inductively, \(L_i \cup K_{i+1}\) has a cover by finitely many open sets with compact closures; let \(K_{i+1}\) be the union of the closures. Suppose \(\{U_i\}\) is an open cover of \(X\). The collection \(\{U_i \cap K_{i+2}^{int} - K_{i-1}\}\) \((K_{-1} = \emptyset)\) covers \(K_{i+1}^{int} - K_i\), and has a finite subcover \(V_{ij}\). The collection of all \(V_{ij}\) refines \(\{U_i\}\) and covers \(X\). Further, it is locally finite since given \(x \in K_k, K_k^{int}\) is a neighborhood not intersecting \(V_{ij}\) for \(i > k + 1\). For part c, there is a cover of \(X\) by open sets with compact closure. This has a locally finite refinement. By lemma 16 this is countable, so the closures of the sets of the cover are compact and cover \(X\).

5. **Connectedness.** A path in a topological space \(X\) is a continuous function \(f : [0, 1] \rightarrow X\); the points \(f(0), f(1)\) are called the endpoints, and are said to be joined by the path. \(X\) is called

- connected if it is not the disjoint union of two nonempty open subsets;

- path connected if any two points can be joined by a path;

- locally (path) connected if for every open set \(U\) and \(x \in U\) there is a (path) connected open neighborhood \(V\) of \(x\) contained in \(U\).

(The first definition was already given above).

**Theorem 18.**

a. If \(\{S_t\}\) is a family of connected subspaces and \(\cap_t S_t\) is nonempty then \(\cup_t S_t\) is connected.

b. If \(S\) is connected and \(S \subseteq T \subseteq S^{t}\) then \(T\) is connected.

c. The relation of belonging to a common connected subspace is an equivalence relation on \(X\).

d. The equivalence classes, which are called the components of the space, are maximal connected subspaces.

e. The components are closed.

**Proof:** Suppose \(\cup_t S_t \subseteq A \cup B\) where \(A, B\) are disjoint and open. Since each \(S_t\) is connected, it must be a subset of either \(A\) or \(B\) because its intersection with either is open in \(S_t\). Since there is a point common to all \(S_t\) they must all be in either \(A\) or \(B\). This proves part a. For part b proceed similarly; \(S\) must be in \(A\) or \(B\), say \(A\), and \(S^{t}\) must therefore be also. For part c, the relation is obviously reflexive and transitive, and if \(x, y\) are in a common connected subspace and \(y, z\) are then \(x, z\) are by part a. Two points in a connected subspace are equivalent, and an equivalence class is connected since anything in the same class as \(x\) is in a connected
clearly Hausdorff. If \( X \) points, since any larger subset can be disconnected, as is readily verified. A totally disconnected space is a base with elements, and consider distinct

Theorem 21

Theorem 19.

a. The relation of being joined by a path is an equivalence relation on \( X \).

b. The equivalence classes, which are called the path components of the space, are maximal path connected subspaces.

Proof: The constant path joins a point to itself. The “reverse” \( g \) of a path \( f \) joining \( x \) to \( y \) joins \( y \) to \( x \); this is defined by \( g(t) = f(1 - t) \). The “concatenation” \( g \) of paths \( f_1 \) from \( x \) to \( y \), and \( f_2 \) from \( y \) to \( z \), joins \( x \) to \( z \); this is defined by \( g(t) = f_1(2t) \) for \( t \leq 1/2 \) \( g(t) = f_1(2t - 1) \) for \( t \geq 1/2 \). This proves part a. All points in a path connected subspace are equivalent, and an equivalence class is path connected, proving part b.

Theorem 20. The product of a family of connected spaces is connected.

Proof: Suppose \( P = \times_i X_i = A_1 \cup A_2 \) where \( A_1, A_2 \) are open and disjoint. Fix an element \( z \in P \). Let \( S \) be a set of elements of \( P \) which equal \( z \) in all but one coordinate position \( i \). It is easily seen that since \( X_i \) is connected \( S \) must lie in either \( A_1 \) or \( A_2 \). It follows that any two elements of \( P \) which differ in only one place are in the same \( A_i \), whence any two which differ in only finitely many places are. The elements which differ from \( z \) in only finitely many places thus lie in the same \( A_i \) as \( z \). The closure of this set lies in that \( A_i \), and the closure equals \( P \).

We leave the following to the reader (exercise 1g).

- If \( f : X \to Y \) is continuous and \( X \) is (path) connected then \( f[X] \) is (path) connected.
- A (locally) path connected space is (locally) connected.
- A space is locally (path) connected iff the (path) components of every open subspace are open.
- In a locally path connected space the path components are the components.
- The product of a finite family of path connected spaces is path connected.
- A locally connected second countable space has countably many components.

A standard counterexample for connected spaces is the set \( \{0\} \times [-1,1] \cup \{\langle x, \sin(2\pi/x) \rangle : 0 < x \leq 1\} \). This is connected, but not path connected or locally connected. Adding a path from \( \langle 0, -1 \rangle \) to \( \langle 1, 0 \rangle \) which is disjoint from the rest of the set yields a path connected space which is not locally connected.

A set is called clopen if it is both closed and open. \( \emptyset \) and \( X \) are clopen sets, called the trivial ones. \( X \) is connected iff there are no nontrivial clopen sets, since \( S, S^c \) are both open iff \( S \) is clopen. A clopen set is a union of components of \( X \) (consider \( C \cap S, C \cap S^c \) for a component \( C \)). In particular a connected clopen set is a component. If the components of a space are open, for example if the space is locally connected, then they are are clopen (since the complement of a component is the union of the remaining components).

A space is called totally disconnected if for any two distinct points \( x, y \) there are disjoint open sets \( A, B \) with \( x \in A, y \in B, \) and \( X = A \cup B \); equivalently, if there is a clopen set containing \( x \) but not \( y \). The rationals provide an example which is not discrete; if \( r_1 < r_2 \) are rational let \( s \) be an irrational between them, and consider \( \{x : x < s\} \) and \( \{x : x > s\} \). The components of a totally disconnected space are the points, since any larger subset can be disconnected, as is readily verified. A totally disconnected space is clearly Hausdorff. If \( X \) is \( T_0 \) and the clopen sets form a base then \( X \) is totally disconnected. Indeed, given distinct \( x, y \) take an open neighborhood \( U \) of one (say \( x \)) not containing the other; then take a set \( V \) in the base with \( x \in B \subseteq U \).

Theorem 21. If \( X \) is compact and totally disconnected then the clopen sets form a base.

Proof: Suppose \( U \) is open and \( x \in U \); then \( U^c \) is compact. For each \( y \in U^c \) choose a clopen set \( S_y \) containing \( y \) but not \( x \). The \( S_y \) form an open cover of \( U^c \), so there is a finite subcover. Let \( W \) be the union of the finite subcover; then \( W^c \) is a clopen set with \( x \in W^c \) and \( W^c \subseteq U \).
If \( Y \) is a subspace of \( X \) and \( S \) is clopen in \( X \) then \( Y \cap S \) is clopen in \( Y \). It follows that a subspace of a totally disconnected space is totally disconnected. In a product of totally disconnected sets, the subbasic open set \( U \times (\times_{k \neq j} X_j), U \subseteq X_j, U \) clopen, is clopen. It follows readily that a product of totally disconnected spaces is totally disconnected. In particular, any subspace of a product of copies of the discrete two element space is totally disconnected. We will see in the next section that it is also compact. Exercise 5 gives the converse.

6. Tychanoff’s theorem. A collection of sets is said to have the finite intersection property if the intersection of any finite subcollection is nonempty. It is easily seen that a space is compact iff whenever a collection \( C \) of closed sets has the finite intersection property then \( \bigcap C \neq \emptyset \). It is also easy to see that given a base for \( X \), \( X \) is compact iff any cover by basic open sets has a finite subcover, or equivalently any collection of basic closed sets with the finite intersection property has nonempty intersection. In fact, “basic” can be replaced by “subbasic”.

**Theorem 22.** A topological space \( X \) is compact iff every collection of subbasic closed sets with the finite intersection property has nonempty intersection.

**Proof:** One direction is trivial. For the other, we first show that, if \( C \) is a collection of basic closed sets with the finite intersection property, then there is a collection \( C' \) which is maximal among the collections of basic closed sets containing \( C \) and having the finite intersection property. By the maximal principle it suffices to show that the union of a chain \( D \) of such collections is again such a collection. Certainly \( \bigcup D \) is a collection of subbasic closed sets containing \( C \). Any finite subcollection \( F \subseteq \bigcup D \) is a subcollection of some collection in \( D \), and so has nonempty intersection. Next, if \( K \in C' \) then \( K = K_1 \cup \cdots \cup K_t \) for some subbasic closed sets \( K_1, \ldots, K_t \). We claim that one of the \( K_i \) is in \( C' \). This follows because if for each \( K_i \) there was a finite \( F_i \subseteq C' \) with \( K_i \cap \bigcap F_i = \emptyset \) then \( \bigcap F_i = \emptyset \) where \( F' = \bigcup F_i \); hence some \( K_i \) could be added to \( C' \), and since \( C' \) is maximal \( K_i \in C' \). Now, since every \( K \in C' \) contains some subbasic closed set which is also in \( C' \), it follows from the hypothesis that \( \bigcap C' \neq \emptyset \), whence \( \bigcap C \neq \emptyset \).

**Theorem 23 (Tychanoff’s theorem).** The product of a collection of compact spaces is compact.

**Proof:** The sets \( K \times (\times_{k \neq j} X_j), K \subseteq X_j, K \) closed, form a subbase for the closed sets in the product topology. Let \( C \) be a collection of these with the finite intersection property. Let \( C_i = \{ \pi_i(K) : K \in C \} \) where \( \pi_i \) is the projection onto the \( i \)th coordinate. Each \( C_i \) is a collection of closed subsets of \( X_i \) with the finite intersection property, so there is an \( x_i \in \bigcap C_i \). It follows that \( (x_1, \ldots, x_n) \in \bigcap C \).

7. Metric spaces. A pseudo-metric function on a set \( X \) is a function \( d : X \times X \mapsto \mathbb{R} \), such that
- \( d(x, y) \geq 0 \),
- \( d(x, x) = 0 \),
- \( d(x, y) = d(y, x) \), and
- \( d(x, y) + d(y, z) \geq d(x, z) \) (triangle inequality).

Let \( B_{\epsilon x} = \{ w : d(w, x) < \epsilon \} \); such sets are often called “open balls”.

**Theorem 24.**

a. If \( x \in B_{\epsilon W} \) and \( \epsilon < \zeta - d(u, x) \) then \( B_{\epsilon x} \subseteq B_{\mu \zeta} \).

b. The open balls form the base for a topology on \( X \).

**Proof:** If \( z \in B_{\epsilon x} \), then \( d(u, z) \leq d(u, x) + d(x, z) < \zeta - \epsilon + \epsilon \); this proves part a. It follows that if \( x \in B_{\epsilon W} \cap B_{\epsilon x} \) then \( B_{\epsilon x} \subseteq B_{\epsilon W} \cap B_{\epsilon x} \) provided \( \epsilon < \zeta - d(u, x), \xi - d(v, x) \).

The set \( S = \{ w : d(w, x) \leq \epsilon \} \) is closed, since if \( d(w, x) > \epsilon \) there is an open ball around \( w \) disjoint from \( S \). Such sets are called closed balls.
X is called a pseudo-metric space when equipped with a pseudo-metric. It is a topological space when equipped with the topology determined by the open balls, which is called the metric topology. A pseudo-metric is called a metric if \( d(x, y) = 0 \) implies \( x = y \); X is called a metric space when equipped with a metric. A pseudo-metric may be extended to subsets of \( X \); define \( d(S, T) \) to be \( \inf \{ d(x, y) : x \in S, y \in T \} \), with \( d(x, S) \) written for \( d(x, \{ x \}, S) \). As noted in section 10.7, the absolute value \( |x| \) on \( \mathbb{R} \) satisfies the axioms \( |x| \geq 0, |x| = 0 \) iff \( x = 0 \), \( |x + y| \leq |x| + |y| \), and \( |xy| = |x||y| \). It follows that \( d(x, y) = |x - y| \) is a metric on \( \mathbb{R} \).

**Theorem 25.** Let \( X \) be a pseudo-metric space, with pseudo-metric \( d \).

a. If \( f : X \to Y \) where \( X, Y \) are pseudo-metric spaces then \( f \) is continuous iff, given \( x \in X \) and \( \epsilon > 0 \) there is a \( \delta > 0 \) such that \( d(f(x), f(y)) < \epsilon \) whenever \( d(x, y) < \delta \).

b. For \( S \subseteq X \) \( d(x, S) \) is a continuous function of \( x \).

c. For \( S \subseteq X \) \( x \in S^c \) iff \( d(x, S) = 0 \).

d. \( X \) is normal.

e. \( X \) is \( T_1 \) (and hence \( T_3 \)) iff it is a metric space.

f. \( X \) is first countable.

g. If \( X \) is separable it is second countable.

**Proof:** We leave part a as an exercise. For part b, from the triangle inequality it follows that \( d(x, S) \leq d(x, y) + d(y, S) \); also \( d(y, S) \leq d(x, y) + d(x, S) \). Thus \( |d(x, S) - d(y, S)| \leq d(x, y) \) and part b follows. For part c, by part b \( \{ x : d(x, S) = 0 \} \) is closed, and it clearly contains \( S \). If \( x \notin S^c \) then for some \( \epsilon \) \( B_{\epsilon} \cap S = \emptyset \), and \( d(x, S) \geq \epsilon \). For part d, given closed sets \( K, L \) let \( U = \{ x : d(x, K) < d(x, L) \} \), \( V = \{ x : d(x, K) > d(x, L) \} \). By part b and properties if \( R, U, V \) are open, and they are clearly disjoint. Also, \( K \subseteq A \) since \( d(x, L) > 0 \) for \( x \in K \) by part c; similarly \( L \subseteq B \). For part e, if \( X \) is \( T_1 \) and \( x \neq y \) there is an open set containing \( x \) but not \( y \), hence an open ball, so \( d(x, y) > 0 \). Conversely if \( X \) is a metric space and \( x \neq y \) then \( d(x, y) > 0 \) so there is an open ball containing \( x \) but not \( y \). For parts f.g, the balls \( B_{x,1/n} \) for \( n \in \mathbb{N}^+ \) form a neighborhood base at \( x \), so \( X \) is first countable. If \( X \) is separable, we claim the balls \( B_{x,1/n} \) for \( n \in \mathbb{N}^+ \), \( x \in S \) for a base, whence \( X \) is second countable. Indeed given \( B_{uc} \) select a point \( v \in S \cap B_{u,1/n} \) where \( 1/n < \zeta/3 \); then \( u \in B_{v,2/n} \) and \( B_{v,2/n} \subseteq B_{uc} \), which proves the claim.

If \( C \) is an open cover of a metric space \( X \), we call a real number \( \epsilon > 0 \) a Lebesgue number for the cover if for every point \( x \) \( B_{\epsilon} \) is contained in some set in the cover. A metric space is called totally bounded iff for every \( \epsilon > 0 \) there is a finite set \( S \) of points such that \( \{ B_{\epsilon} : x \in S \} \) is a cover of \( X \). An infinite sequence in a metric space is called a Cauchy sequence if for all \( \epsilon > 0 \) there is an \( n \) such that \( d(x_i, x_j) < \epsilon \) if \( i, j \geq n \). A metric space is called complete if every Cauchy sequence converges to some limit.

**Theorem 26.** Suppose \( X \) is a metric spaces.

a. If \( X \) is sequentially compact then \( X \) has a Lebesgue number.

b. If \( X \) is sequentially compact then compact \( X \) is totally bounded.

c. If \( X \) is sequentially compact then \( X \) is compact.

d. If \( X \) is compact then \( X \) is complete.

e. If \( X \) is totally bounded and complete then \( X \) is compact.

f. If \( Y \subseteq X \) is a complete subspace then \( Y \) is closed in \( X \) if it is closed.

g. If \( X \) is complete and \( Y \subseteq X \) is a closed subspace then \( Y \) is complete.

**Proof:** For part a, suppose to the contrary that for each \( n \in \mathbb{N}^+ \) there is a \( w \) such that \( B_{w,1/n} \) is not contained in a set of the cover. The set \( S \) of such \( w \) is clearly infinite, so by assumption has a limit point \( x \). This is contained in some set \( U \) of the cover. But then there are infinitely many points of \( S \) inside \( B_{x,\epsilon} \) where \( B_{x,2\epsilon} \subseteq U \), a contradiction. For part b, choose successively points \( x_i \) such that \( x_i \) is outside \( \bigcup_{j<i} B_{x,\epsilon} \).
This must terminate after a finite number of steps, else we have an infinite set with no limit point. For part c, let $C$ be an open cover of $X$. Let $\epsilon$ be a Lebesgue number, and let $S$ be a finite set of points such that \{ $B_x : x \in S$ \} is a cover. Each $B_x$ is in some set of $C$, and choosing one such for each $x \in S$ yields a finite subcover of $C$. Part d follows, since any sequence either runs through finitely many values and so has a limit; or runs through infinitely many values and so has a limit point $x$ by compactness, and for a Cauchy sequence $x$ is the limit of the sequence. For part e, let $\langle x_i^0 \rangle$ be any sequence; we show that it has a Cauchy subsequence, whence \{ $x_i$ \} has a limit point, so that $X$ is compact. It follows from total boundedness that given an infinite sequence $\langle x_i \rangle$ and an $\epsilon > 0$ there is a $B_x$, in which infinitely many of the $x_i$ lie. Successively choose subsequences $\langle x_i^k \rangle$, $k > 0$, to lie within some $B_{x_1/k}$. The $x_i^k$ form a Cauchy subsequence of the original sequence. For part f, if $Y$ is complete and $y_n$ is a sequence in $Y$ converging to $y$ in $X$, then $y_n$ is a Cauchy sequence, so $y \in Y$. For part g, if $y_n$ is a Cauchy sequence in $Y$ then $y_n$ converges to some $y$ in $X$, and by closure $y \in Y$.

**Theorem 27.** A pseudo-metric space $X$ is paracompact.

**Proof:** Let \{ $C_\alpha : \alpha \in J$ \} be a cover of $X$, and let $\alpha$ be a well order on $J$. Define sets $D_{\alpha n}$ for $n \in \mathbb{N}^+$ by induction on $n$ as follows. $D_{\alpha n}$ is the union of all $B_x$ such that $x$ is not in any $C_\beta$ with $\beta < \alpha$ or $D_{\beta j}$ with $j < n$, and $B_x < \alpha n \subseteq C_\alpha$. \{ $D_{\alpha n}$ \} is a cover, because if $\alpha$ is least such that $x \in C_\alpha$ and $n$ is least such that $B_x < \alpha n \subseteq C_\alpha$ then $x$ will be in $D_{\alpha n}$ unless it is already in some $D_{\beta j}$ with $j < n$. Certainly \{ $D_{\alpha n}$ \} refines \{ $C_\alpha$ \}. Now, given $x$, let $\alpha, n$ be such that $x \in D_{\alpha n}$, and let $m$ be such that $B_x < \alpha n \subseteq D_{\alpha m}$. We claim that $B = B_{x,1/2^{n+m}}$ does not intersect $D_{\beta j}$ if $j > m + n$. Indeed, an open ball which can be included in $D_{\beta j}$ has its center $y$ outside $D_{\alpha m}$ since $j > n$, whence $d(x,y) > 1/2^m$, and the claim follows because $1/2^{(n+m)} + 1/2^j \leq 1/2^m$. Finally we claim that $B$ intersects at most one $D_{\beta j}$ for each $j < n+m$, for which it suffices to show that given points $y_t \in D_{\beta j}$, $t = 1,2$, where $\beta_1 < \beta_2$, $d(y_1,y_2) > 2 \cdot 1/2^{n+m}$. Now, $y_t \in B_{x,1/2^j} \subseteq D_{\beta,1}$ for some $z_i$, $t = 1,2$; and $d(z_1,z_2) > 3/2^t$ since $B_{y_1,1/2^t} \subseteq C_{\beta_1}$ but $z \notin C_{\beta_1}$. Thus $d(y_1,y_2) > 3/2^t$ and the claim follows.

If $d$ is a pseudo-metric on $X$ then $Y \subseteq X$ becomes a pseudo-metric space when equipped with the restriction of $d$ to $Y$. The metric topology on $Y$ is easily seen to be that inherited from the metric topology on $X$. If $d$ is a metric then its restriction is also.

**Lemma 28.** Suppose $\langle X_i, d_i \rangle$, $i \geq 1$, is a countable family of pseudo-metric spaces. The binary function on $\times_i X_i$ given by

$$d(\langle x_i \rangle, \langle y_i \rangle) = \sum_i \min(1, d(x_i, y_i))/2^i$$

is a pseudo-metric, whose metric topology is the product space topology. If the pseudo-metrics are metrics then $d$ is a metric.

**Proof:** First note that $\min(1, d(x,y))$ is a pseudo-metric (exercise 10), and has the same topology since the open balls of radius less than 1 are a base for the topology of either function. We leave the verification that $d$ is a pseudo-metric to exercise 10. Write $x$ for an element of the product, with $x_i$ its $i$th component. If $B = B_{x,1/2^p}$ let $U = \{ y : d_i(x_i, y_i) < 1/2^{p+i+2}, i \leq p+2 \}$. Then $U$ is open in the product topology and contains $x$, and as is readily verified $U \subseteq B$. On the other hand given $U = \{ x : x_i \in V \}$ where $V$ is open in $X_i$, choose $\epsilon$ with $B_{x_i,\epsilon} \subseteq V$. Then in the product $B_{x,\epsilon/2} \subseteq U$. Finally, clearly $d(\langle x_i \rangle, \langle y_i \rangle) = 0$ iff $d(x_i, y_i) = 0$ for all $i$.

**Lemma 29.** Suppose $f_i : X \rightarrow Y_i$ are given maps, $f : X \rightarrow \times_i Y_i$ is the induced map, and $f'$ is the corestriction of $f$ to $f[X]$. Suppose that for any closed subset $K \subseteq X$ and any $x \in K$ there is an $i$ such that $f_i(x) \notin (f[K])^i$. Then $f'$ is open.
Proof: Let $U \subseteq X$ be open, and suppose $x \in U$. Choose $i$ so that $f_i(x) \notin (f_i(U^c))^c$. Choose an open set $V \subseteq Y_i$ with $f_i(x) \in V$ and $V \cap f_i(U^c) = \emptyset$. Then $f(x) \in V \times (\times_{j \neq i} Y_j) \cap f[X] \subseteq f[U]$; since $x$ was arbitrary this proves that $f[U]$ is open in $f[X]$.

A topological space $X$ is called pseudo-metrizable if there is a pseudo-metric defined on $X$ whose metric topology is the given topology. $X$ is called metrizable if there is such a metric; clearly this is so iff $X$ is pseudo-metrizable and $T_1$, and in this case any pseudo-metric with the given topology is a metric. A collection of subsets of a topological space is called $\sigma$-locally finite if it is a countable union of locally finite collections. A topological space is called locally metrizable iff every point has an open neighborhood which is metrizable.

Lemma 30. If $X$ is regular and has a $\sigma$-locally finite base then $X$ is normal.

Proof: Proceed as in the proof of theorem 13a, except define $C'_i$ as follows. Write $C_i$ as $\cup_i C_{ij}$ where $C_{ij}$ is locally finite, and let $C'_i$ be $\{U_{ij}: U_{ij} = \bigcup C_{ij}\}$. Each $U_{ij}^c$ is disjoint from $K_{1-i}$, and the rest of the proof is as before.

Theorem 31. $X$ is pseudo-metrizable iff it is regular and has a $\sigma$-locally finite base.

Proof: Let $X$ denote the topological space and $(X, d)$ $X$ equipped with a pseudo-metric. If $d$ is a pseudo-metric on $X$ the identity map from $X$ to $(X, d)$ is continuous iff $d$ is continuous (exercise). Also, the identity map has the property of lemma 29 iff $d(x, K) > 0$ for any closed $K$ and $x \in K^c$. Thus if we could find a countable collection $\{d_i\}$ of continuous pseudo-metrics on $X$, such that for each closed $K \subseteq X$ and $x \in K^c$ there is an $i$ with $d_i(x, K) > 0$, it would follow by lemmas 28 and 29 that $X$ is pseudo-metrizable. Let $B = \cup_n B_n$ be a base, where each $B_n$ is locally finite. Given $n, m$ define $d_{nm}$ as follows. For $U \in B_m$ let $U_\ast = \cup\{V \in B_n: V^c \subseteq U\}$; by local finiteness $U_\ast^c \subseteq U$. By lemma 29 and theorem 8 there is a continuous function $f_U: X \mapsto [0, 1]$ such that $f_U|U_\ast = 1$ and $f_U|U^c = 0$. Let $d_{nm}(x, y) = \sum_{U \in B_m} |f_U(x) - f_U(y)|$; note that the sum is actually finite, and further $d_{nm}$ is continuous. Now, if $K \subseteq X$ is closed and $x \in K^c$ then $x \in U$ and $U \subseteq K^c$ for some $m$ and $U \in B_m$; and since $X$ is regular $x \in V$ and $V^c \subseteq U$ for some $n$ and $V \in B_n$. It is readily seen that $d_{nm}(x, K) \geq 1$. For the “only if” direction, by theorem 27 $X$ is paracompact. Let $C_n$ be the cover $\{B_{x, 1/n}\}$; each $C_n$ has a locally finite refinement $C'_n$. If $U$ is open and $x \in U$ then $B_{x, 1/n} \subseteq U$ for some $n$; there is a $B \in C'_n$ containing $x$, and $B \subseteq B_{x, 1/n}$. Thus $\cup_n C'_n$ is a base.

Theorem 32. If $X$ is paracompact, Hausdorff, and locally metrizable then $X$ is metrizable.

Proof: We show that $X$ has a $\sigma$-locally finite base; the proof is similar to the “only if” direction of theorem 31. Let $C_n$ be the cover $\{B_{x, 1/n}\}$ where $\alpha$ runs over a cover $\{X^\alpha\}$ of $X$ by metric subspaces and $x$ runs over $X^\alpha$. Given $U$ and $x \in U$, choose $\alpha$ with $x \in X^\alpha$ and replace $U$ by $U \cap X^\alpha$.

8. Completion. As observed previously, $\mathcal{R}$ is a metric space when equipped with the “usual” metric $|x - y|$. The following lemma can be proved by using the fact that $\mathcal{R}$ is the completion of the rationals, in the sense to be given shortly. We will prove it from the fact that every nonempty set of real numbers with an upper (lower) bound has a least upper bound, abbreviated lub (greatest lower bound, glb). This fact may be proved from the set-theoretic definition of the real numbers; or taken as an axiom.

Lemma 33. $\mathcal{R}$ is complete.

Proof: Let $\langle x_n \rangle$ be a Cauchy sequence in $\mathcal{R}$. The set of values taken on by the sequence is bounded above; choose any $\epsilon > 0$, choose an $N$ so that if $i, j \geq N$ then $|x_i - x_j| < \epsilon$, and consider $x_N + \epsilon$. This bounds above $x_N$ for $x \geq N$, so an upper bound can be obtained by considering the maximum of this and the $x_n$ for $n < N$. A similar argument shows that the set of values is bounded below. It is clear that $\text{lub}\{x_k : k \geq n\}$ exists for
each \( n \). Letting \( b_n \) denote this value, it is clear that \( \text{glb}\{b_n\} \) exists. Let \( x \) denote this value; we claim that the sequence \( x_n \) converges to \( x \). Given \( \epsilon \), we may successively choose \( N \leq M \leq L \) so that \( |x - b_N| < \epsilon/3 \), \( |b_N - x_M| < \epsilon/3 \), and \( |x_M - x_n| < \epsilon/3 \) for \( n \geq L \), which proves the claim.

Let \( \text{Met} \) be the category of metric spaces and continuous functions, and let \( \text{CMet} \) be the full subcategory of complete metric spaces. For various applications the collection of arrows is too large, and other categories of metric spaces are considered. In particular, a function \( f : X \to Y \) between metric spaces is said to be uniformly continuous iff, for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that \( d(f(x), f(y)) < \epsilon \) whenever \( d(x, y) < \delta \). It is readily verified that the metric spaces with the uniformly continuous functions form a category, which we denote \( \text{Met}_U \), with \( \text{CMet}_U \) the complete metric spaces.

The uniformly continuous functions preserve Cauchy sequences (which is not the case for arbitrary continuous functions). Indeed, if \( f : X \to Y \) is uniformly continuous, \( \langle x_i \rangle \) is a Cauchy sequence in \( X \), and \( \epsilon > 0 \), choose \( \delta \) so that \( d(f(x_i), f(x_j)) < \epsilon \) whenever \( d(x_i, x_j) < \delta \). Then choose \( N \) so that if \( i, j \geq N \) then \( d(x_i, x_j) < \delta \). Then \( d(y_i, y_j) < \epsilon \) where \( y_i = f(x_i) \).

If \( f : X \to Y \) is continuous, and uniformly continuous on a dense subspace \( X' \subseteq X \), then \( f \) is uniformly continuous. Indeed, given \( \epsilon \) let \( \delta \) be such that (writing \( y \) for \( f(x) \), etc.) \( d(y'_i, y'_j) < \epsilon/3 \) whenever \( x'_i, x'_j \in X' \) and \( d(x'_i, x'_j) < \delta \). Since \( f \) is continuous, given \( x_i \in X \) for \( i = 1, 2 \) we may choose \( x'_i \in X' \) with \( d(y_i, y'_i) < \epsilon/3 \) for \( i = 1, 2 \).

A function \( f : X \to Y \) between metric spaces for which \( d(f(x), f(y)) = d(x, y) \) is called an isometry; an isometry is clearly a uniformly continuous injection. The metric spaces with the isometries are readily verified to form a category, which we denote \( \text{Met}_I \), with \( \text{CMet}_I \) the complete metric spaces. If \( f : X \to Y \) is an isometry on a dense subset of \( X \) then \( f \) is an isometry, as is readily verified.

Suppose \( X \) is a metric space, with metric \( d \); let \( X_1 \) be the set of Cauchy sequences. If \( x = \langle x_i \rangle \) and \( y = \langle y_i \rangle \) are two such, using the triangle inequality, \( d(x_i, y_i) \) is a Cauchy sequence in \( \mathcal{R} \). The limit thus exists; define \( d_1(x, y) \) to be this limit. The function \( d_1 \) is a pseudo-metric on \( X_1 \); the axioms hold because they hold componentwise. Let \( X^{\text{cpl}} \) be the canonical quotient metric space by the congruence relation \( d(x, y) = 0 \); this metric space is called the completion of \( X \). Let \( j_1 \) be the map from \( X \) to \( X_1 \), which takes \( x \) to the sequence \( \langle x_i \rangle \) where \( x_i = x \) for all \( i \).

**Theorem 34.**

a. The map \( j : X \to X^{\text{cpl}} \) induced by \( j_1 \) is an isometry.

b. \( j[X] \) is dense in \( X^{\text{cpl}} \).

c. \( X^{\text{cpl}} \) is complete.

d. \( \langle X^{\text{cpl}}, j \rangle \) is a universal arrow from \( X \) to the forgetful functor from \( \text{CMet}_U \) and \( \text{Met}_U \).

e. The same is true for \( \text{CMet}_I \) and \( \text{Met}_I \).

**Proof:** For part a, clearly \( d(j_1(x), j_1(y)) = d(x, y) \). For part b, if \( x \in X^{\text{cpl}} \) is represented by the Cauchy sequence \( \langle x_i \rangle \), then the elements \( j(x_i) \) converge to \( x \). For part c, suppose \( \langle x_i \rangle \) is a Cauchy sequence in \( X^{\text{cpl}} \), and suppose \( x_i = [(x_{ij})] \). For each \( i \) choose \( w_i \) so that eventually \( d(x_{ij}, w_i) < 2^{-i} \). Given \( \epsilon \), for sufficiently large \( i_1, i_2 \), \( d(x_{i_1}, x_{i_2}) < \epsilon/4 \); and for any such, for sufficiently large \( j \), \( d(x_{i_1j}, x_{i_2j}) < \epsilon/2 \). Also, for sufficiently large \( i_1 \), and for sufficiently large \( j \) given such, \( d(x_{i_1j}, w_i) < \epsilon/4 \); and similarly for \( i_2 \). It follows that \( \langle w_i \rangle \) is a Cauchy sequence. Further, given \( \epsilon \), for \( i \) sufficiently large \( d(x_i, j[w_i]) < \epsilon/2 \) and \( d([w_i]), [j[w_i]]) < \epsilon/2 \), whence \( \langle x_i \rangle \) converges to \( [(w_i)] \). For part d, suppose \( f : X \to Y \) is uniformly continuous, where \( Y \) is complete. Given \( x \in X^{\text{cpl}} \), let \( \langle x_i \rangle \) be a sequence in \( X \) such that \( \langle j(x_i) \rangle \) converges to \( x \). Then \( \langle f(x_i) \rangle \) is a Cauchy sequence in \( Y \); let \( y \) be the limit. If \( \langle x'_i \rangle \) is another sequence converging to \( x \), its image converges to the same \( y \), because \( d(j(x'_i), y) \) converges to 0. Thus, a function \( g : X^{\text{cpl}} \to Y \) may be defined by letting \( g(x) = y \). By an argument similar to one given above, \( g \) is uniformly continuous; uniqueness follows since \( j[X] \) is dense. For part e, if \( f \) is an isometry so is \( g \).

198
In particular $X \mapsto X^{\text{cmp}}$ is the object map of a functor from $\text{Met}_U$ to $\text{CMet}_U$, which is the left adjoint to the forgetful functor. Observe that if $X' \subseteq X$ where $X$ is complete and $X'$ is dense in $X$, then $X$ is the completion of $X'$.

A topological space may have a variety of metrics defined on it. Two metrics are called equivalent if their metric topologies are the same. In this case topological properties are preserved by the identity map. Cauchy sequences however may not be. If the identity map is uniformly continuous in both directions the metrics are said to be uniformly equivalent and the Cauchy sequences are the same.

These comments are relevant to defining metrics on the product of metric spaces. In lemma 28 one was defined for countably many pseudo-metric or metric spaces. For finitely many spaces the metric on the product may be taken as $\sum_i d_i(x_i, y_i)$. Alternatively $\max_i d_i(x_i, y_i)$ may be used. Since for nonnegative real numbers,

$$\max(r_1, \ldots, r_n) \leq r_1 + \cdots + r_n \leq n \cdot \max(r_1, \ldots, r_n),$$

the two metrics are uniformly equivalent. Another common choice is $\sqrt{\sum_i (x_i - y_i)^2}$, which is also uniformly equivalent.

The product of finitely many complete metric spaces is complete. Using the max metric, it is readily verified that a sequence in the $n$-tuples is a Cauchy sequence iff it is a Cauchy sequence in each component. The componentwise limit of a Cauchy sequence exists, and is readily verified to be the limit in the product.

The notion of a metric space can be generalized, to “uniform spaces”; see [Kelley]. This category is better behaved than the metric spaces in some respects, in particular having arbitrary products.

9. Manifolds. From remarks in section 10.7, the Euclidean norm $|x| = \sqrt{\sum x_i^2}$ is a norm on the vector space $\mathbb{R}^n$ with the standard inner product. In general if $|x|$ is a norm on a real or complex vector space then the function $|x - y|$ is a metric, as is readily verified. It particular the Euclidean distance $\sqrt{\sum (x_i - y_i)^2}$ is a metric on $\mathbb{R}^n$; this normed real vector space is called $n$-dimensional Euclidean space.

Define a subset of metric space to be bounded if it is contained in some open ball. We leave it as exercise 12 to show that $\mathbb{R}^n$ has the following properties.

1. The metric topology equals the product topology.
2. Addition and scalar multiplication (as a function of two arguments) are continuous.
3. $\mathbb{R}^n$ is second countable and $T_4$.
4. $\mathbb{R}^n$ is complete.
5. A subset of $\mathbb{R}^n$ is closed iff it is complete.
6. A subset of $\mathbb{R}^n$ is bounded iff it is totally bounded.
7. A subset of $\mathbb{R}^n$ is compact iff it is closed and bounded (Heine-Borel theorem).
8. $\mathbb{R}^n$ is locally compact.
9. A nonempty subset of $\mathbb{R}$ is connected iff it is an interval.
10. Distinct points $x, y$ in a nonempty open connected subset $S \subseteq \mathbb{R}^n$ can be joined by a polygonal line whose segments are parallel to the coordinate axes.

Call a topological space $X$ locally $n$-Euclidean if every point has an open neighborhood which is homeomorphic to an open subset of $\mathbb{R}^n$. $X$ is clearly locally metrizable, locally compact, and locally path connected. We define an $n$-manifold to be a locally $n$-Euclidean space which is Hausdorff and second countable. It follows that $X$ is paracompact. Some authors require only paracompactness rather than second countability (these are equivalent requirements for a connected space). In either case, $X$ is metrizable. Infinite dimensional manifolds may also be considered; see chapter 24.

An open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$ (exercise 13); it follows that in a manifold every point has a neighborhood homeomorphic to $\mathbb{R}^n$. Sometimes this is given as the definition. Also, some of the requirements can be relaxed, but proving this requires nontrivial theorems from topology (some of which
isomorphism carries $X$ from $C$ to $Y$, maximal be enlarged. Since the collection of $C$ is in $A$. A topological pair is a pair $A$, $B$. $S$ may be empty, in which case the space is a manifold. Let $U$ be a class of continuous functions from open subsets of $R^n$ to $R^n$, which is closed under composition. A basic example is the differentiable functions, which are defined in chapter 24. Two charts are called $C$-compatible if either $U_1 \cap U_2 = \emptyset$, or both transition functions $\phi_2^{-1}\phi_1$ and $\phi_1^{-1}\phi_2$ are in $C$. A manifold is called $C$-smooth if it has an atlas where each pair of charts is $C$-compatible.

Call an atlas a $C$-atlas if each pair of charts is $C$-compatible. A maximal $C$-atlas is one which cannot be enlarged. Since the collection of $C$-atlases is readily verified to be inductive, any atlas is contained in a maximal $C$-atlas. It is a fact that a given manifold may have more than one maximal $C$-atlas. On the other hand, a given $C$-atlas is contained in exactly one maximal $C$-atlas $A$, namely that consisting of the charts $C$-compatible with every chart in $A$. This follows because if two charts are $C$-compatible with every chart in $A$ then they are $C$-compatible, as is easily seen.

A closed ball in $R^n$ is not a manifold; there is a more general definition which includes spaces such as this. A topological pair is a pair $\langle X, X' \rangle$ where $X'$ is a subspace of $X$. These form the objects of a category; the arrows from $\langle X, X' \rangle$ to $\langle Y, Y' \rangle$ are the continuous functions from $X$ to $Y$ such that $f[X] \subseteq Y'$. An isomorphism carries $X'$ bijectively to $Y'$, and is called a homeomorphism of pairs.

An $n$-manifold with boundary is a topological pair $\langle X, B \rangle$ such that
- $X$ is a second countable Hausdorff space,
- $B$ is closed in $X$,
- $X - B$ is an $n$-manifold, and
- if $x \in B$ then $x$ has an open neighborhood $V$ such that $\langle V, V \cap B \rangle$ is homeomorphic to $\langle R^{n-1} \times (0,1), R^{n-1} \times \{1\} \rangle$.

$B$ may be empty, in which case the space is a manifold.

**Exercises.**

1. Prove the facts stated in the chapter, whose proof was left as exercise 1. Hint: For the first claim of part g, if $f[X]$ is not connected then there are open subsets $U, V$ of $Y$ such that $f[X] \cap U$ and $f[X] \cap V$ are nonempty and disjoint.

2. Say that $x$ is an $\omega$-limit point of $S$ if every open subset containing $x$ contains infinitely many points of $S$. Show that if $X$ is Lindelof and every infinite subset has an $\omega$-limit point then $X$ is compact. Show that if $X$ is second countable then any limit point is an $\omega$-limit point.

3. Define the boundary of a set $S$ to be $S^{cl} - S^{int}$; write $S^{bd}$ for the boundary. Prove the following.
   a. $x$ is a limit point of $S$ iff $x$ is a limit point of $S^{int}$.
   b. $S^{int} = ((S^{c})^{cl})^{c}$.
   c. $x \in S^{bd}$ iff $x$ is a limit point of both $S$ and $S^{c}$.
   d. $S^{bd} \cap S^{int} = \emptyset$ and $S^{cl} = S^{bd} \cup S^{int}$.
   e. $S^{bd} = (S^{c})^{bd}$.
   f. $S^{bd}$ is closed.

4. Suppose that $f : X \mapsto Y$ is a function between topological spaces. For $x \in X$ say that $f$ is continuous at $x$ iff, for all open $V \subseteq Y$ with $f(x) \in V$, there is an open $U \subseteq X$ with $x \in U$ and $f[U] \subseteq V$. 

200
a. Show that \( f \) is continuous iff for all open \( V \subseteq Y \), and all \( x \in X \) with \( f(x) \in V \), there is an open \( U \subseteq X \) with \( x \in U \) and \( f[U] \subseteq V \).

b. Show that \( f \) is continuous at \( x \) iff for every open \( V \) containing \( f(x) \) there is an open \( U \) containing \( x \) with \( f[U] \subseteq V \).

c. Show that \( f \) is continuous iff it is continuous at each point of \( X \).

d. If \( X, Y \) are metric spaces show that \( f \) is continuous at \( x \) iff for each \( \epsilon > 0 \), there is a \( \delta > 0 \) such that \( d(x, y) < \delta \implies d(f(x), f(y)) < \epsilon \) for all \( x \in X \) and \( y \in X \) with \( y \neq x \).

e. Show that if \( f \) is continuous at \( x \), then if \( (x_n) \) converges to \( x \) then \( (f(x_n)) \) converges to \( f(x) \).

f. Show that the converse of d holds if there is a countable neighborhood base at \( x \). Hint: Choose \( V \) such that \( \forall U \exists y(y \in U \land f(y) \notin V) \), let \( U \) run over a countable neighborhood base, and choose a \( y \) from each \( U \).

5. Show that a compact totally disconnected space is homeomorphic to a subspace of the product of copies of the discrete two element space. Hint: Consider the product of a copy for each clopen set. Map \( x \) to the sequence which is 1 in position \( i \) if \( x \) is in the \( i \)th clopen set.

6. Show that in a Hausdorff space, if \( C \) is a collection of compact sets with the finite intersection property then \( \cap C \neq \emptyset \).

7. If \( X \) and \( Y \) are metric spaces a sequence of functions \( (f_n) \) where \( f_n : X \to Y \), is said to converge uniformly to a function \( f : X \to Y \) if, for each \( \epsilon > 0 \), there is an integer \( N \) such that \( |f(x) - f_n(x)| < \epsilon \) for all \( x \in X \) and \( n \geq N \). Show that if \( (f_n) \) is a sequence of continuous functions converging uniformly to \( f : X \to Y \), then \( f \) is continuous.

8. Complete the proof of corollary 9. Hint: Let \( g \) be \( \sum_i g_i \); then \( g : X \to [-1, +1] \), and using exercise 7 \( g \) is continuous.

9. In a field with an absolute value, an infinite series \( \sum_i x_i \) is said to converge absolutely if \( \sum_i |x_i| \) converges. Show that, if a series \( \sum_i a_i \) converges absolutely, any series obtained by rearranging the terms converges absolutely, and the sum is the same. Hint: Given \( n \) choose \( m \geq n \) so that the first \( n \) \( a_i \) occur among the first \( m \) terms of the rearranged series. If \( s_1 \) and \( s_2 \) are the two partial sums up to \( m \), \( |s_1 - s_2| \leq \sum_{i > n} |a_i| \).

10. Show the following.

a. \( \min(1, d(x, y)) \) is a pseudo-metric if \( d \) is.

b. \( d \) of lemma 28 is a pseudo-metric. Hint: For the triangle inequality, use exercise 9.

11. Show that if \( d \) is a pseudo-metric on a topological space \( X \) the identity map from \( X \) to \( (X, d) \) is continuous iff \( d \) is continuous.

12. Prove the properties of \( R^n \) stated in the text. Hint: For claim 4, use lemma 33 and remarks on products in section 8. For claim 5, use theorem 26. For claim 6, a totally bounded set is clearly bounded; for the converse, show that an open cell is totally bounded. For claim 7, use claims 5 and 6, and theorem 26. For claim 9, if \( S \) is not an interval take \( x < y < z \) with \( x, z \in S \) and \( y \notin S \). If \( S \) is not connected take \( A, B \) disjoint open sets with \( S \subseteq A \cup B \). Take \( x \in A, z \in B \) where we may assume \( x < z \). Let \( y = \sup \{w : x \leq w \leq z, w \in A \} \) and derive a contradiction. For claim 10, choose a point \( x \) in the subset \( X \) and consider all points of \( S \) reachable from \( x \) by such paths, and the set not reachable.

13. Show that an open ball in \( R^n \) is homeomorphic to \( R^n \). Hint: First show that \( x/(1 - x^2) \) is a homeomorphism from \( (-1, 1) \) to \( R \).

14. If \( X \) and \( Y \) are topological spaces, with \( \equiv_X \) \( \equiv_Y \) a partition on \( X \) \( (Y) \), define \( \equiv \) on \( X \times Y \) so that \( (x_1, y_1) \equiv (x_2, y_2) \) iff \( x_1 \equiv_X x_2 \) and \( y_1 \equiv_Y y_2 \). The elements of \( (X/\equiv_X) \times (Y/\equiv_Y) \) are of the form \( ([x], [y]) \). The elements of \( (X \times Y)/\equiv \) are of the form \( ([x], y) \), and this equals \( [x] \times [y] \). The maps \( ([x], [y]) \mapsto ([x], y) \)
and \( [(x, y)] \mapsto ([x], [y]) \) are well-defined and inverse to each other. Let \( T_1 \) be the subsets of \( X \times Y \) which are a union of sets \( U \times V \) where \( U \) (\( V \)) is saturated open in \( X \) (\( Y \)); ignoring some bijections these are the open sets of \( (X/\equiv_X) \times (Y/\equiv_Y) \). Let \( T_2 \) be the saturated subsets of \( X \times Y \) which are a union of sets \( U \times V \) where \( U \) (\( V \)) is open in \( X \) (\( Y \)); ignoring some bijections these are the open sets of \( (X \times Y)/\equiv \). Clearly \( T_1 \subseteq T_2 \). Suppose that for \( V \) saturated open in \( Y \), and \( x \in V \), there is a saturated open \( W \) with \( v \in W \subseteq V \) and \( W^{\text{cl}} \) compact. Prove that under this hypothesis \( T_1 = T_2 \). Hint: Let \( Q \) be a set in \( T_2 \), and suppose \( (u, v) \in Q \). Let \( V = \{ w : (u, w) \in Q \} \); \( V \) is saturated open. Choose \( W \) as in the hypothesis, and for each \( w \in W \) choose open \( U_w \) and \( V_w \) such that \( (u, w) \in U_w \times V_w \subseteq Q \). Choose a finite cover of \( W \) from among the \( V_w \), and let \( U \) be the intersection of the corresponding \( U_w \). Then \( U \times W \) is a subset of \( Q \); replace \( U \) by the largest such \( U \). Suppose \( u_1 \in U \) and \( u_2 \equiv_X u_1 \); conclude that \( u_2 \in U \).

15. A topological space \( X \) is said to be compactly generated if it is Hausdorff and has the topology coherent with its compact subspaces. Show the following.
   a. If \( X \) has the topology coinduced by a collection \( \{ A_j \} \) of subspaces, and \( Y \) is locally compact, then \( X \times Y \) has the topology coinduced by \( \{ A_j \times Y \} \).
   b. If \( X \) is compactly generated and \( Y \) is a locally compact Hausdorff space then \( X \times Y \) is compactly generated.

Hint: For part a, use theorem 4 and exercise 14. For part b, \( X \times Y \) has the topology coherent with \( \{ K \times Y : K \subseteq X \text{ is compact} \} \); and \( K \times Y \) has the topology coherent with \( \{ K \times L : L \subseteq Y \text{ is compact} \} \).
18. Tensor algebra.

1. Adjunction with a parameter. Recall the notation \( \text{Hom}^a \) and \( \text{Hom}_b \) introduced in section 13.7; and that these functors preserve limits and monics (in the contravariant case, mapping colimits to limits and epics to monics). For \( f : d \rightarrow d' \), \( \text{Hom}^a(f) \) is a map from \( \text{Hom}(a, d) \) to \( \text{Hom}(a, d') \), namely \( g \mapsto fg \).

A notation for this operation (composition with \( f \) on the left) is often introduced; we will use \( f^L \). (This conflicts with previous use for left inverse; which is meant will be clear from context.) It operates on a subclass of \( Ar \), namely \( \cup_a \text{Hom}(a, d) \). Similarly \( f^R \) will be used to denote the operation of composition on the right with \( f \).

We note the following useful fact about natural equivalences. If \( F \) and \( G \) are naturally equivalent and \( F(f)F(g) = F(h) \) then \( G(f)G(g) = G(h) \). It follows that if \( F \alpha \) is a cone from \( F(c) \) to \( FH \) then \( F \alpha \) is a cone from \( G(c) \) to \( GH \) (regardless of whether \( \alpha \) is a cone). If \( F \alpha \) is a limit cone then so is \( G \alpha \); a cone from \( c' \) to \( GH \) induces a cone to \( FH \), whence an arrow to \( F(c) \), whence an arrow to \( G(c) \), which is readily verified to be unique.

**Lemma 1.** If, for a family \( \alpha_i : c \rightarrow F(i) \), for each \( a \) \( \text{Hom}^a \) maps the family to a limit cone, then \( \langle c, \alpha \rangle \) is a limit for \( F \). Dually if for a family \( \alpha_i : F(i) \rightarrow c \), for each \( b \) \( \text{Hom}_b \) maps the family to a limit cone, then \( \langle c, \alpha \rangle \) is a colimit for \( F \).

**Proof:** Suppose \( \{\alpha_i : i \in J\} \) is a family and \( \mu : F(i) \rightarrow F(j) \). By hypothesis (in fact using only that \( \text{Hom}^c \) maps the family to a cone), \( \mu^L \alpha_j^L = \alpha_j^L \), so \( \mu \alpha_i = \mu^L \alpha_j^L(\iota_c) = \alpha_j^L(\iota_c) = \alpha_j \), and \( \alpha \) is a cone. Given any other cone \( \beta \) from \( d \rightarrow F \), since \( \text{Hom}^d \alpha \) is a limit cone there is a map \( \theta \) from \( \text{Hom}^d(d) \) to \( \text{Hom}^d(c) \) such that \( \alpha_i^L \theta = \beta_i^L \) for \( i \in J \); and \( \alpha_i \theta(\iota_d) = \beta_i \) for \( i \in J \). On the other hand if \( \alpha_i \eta = \beta_i \) for \( i \in J \) then \( \alpha_i^L \eta^L = \beta_i^L \) for \( i \in J \), so \( \eta^L = \theta \) and \( \eta = \eta^L(\iota_d) = \theta(\iota_d) \).

**Theorem 2.** Suppose \( F \) is a functor from \( A \times B \rightarrow C \), \( G \) is a functor from \( B^{op} \times C \rightarrow A \), and

\[
\text{Hom}_C(F(a, b), c) \cong \text{Hom}_A(a, G(b, c))
\]

is natural in all three variables. Then every \( F(a, -) \) is colimit preserving iff every \( G(-, c) \) is limit preserving.

**Proof:** Suppose \( \langle b, \alpha \rangle \) is a colimit in \( B \), and suppose \( G(-, c) \) is limit preserving for all \( c \). Then applying \( \text{Hom}(a, G(-, c)) \) yields a limit in \( \text{Set} \), for any \( a, c \). The natural equivalence yields that applying \( \text{Hom}(F(a, -), c) \) yields a limit in \( \text{Set} \), for any \( a, c \). Thus, by lemma 1 applying \( F(a, -) \) yields a colimit in \( B \), for all \( a \). For the converse, again suppose \( \langle b, \alpha \rangle \) is a colimit in \( B \), and argue similarly.

**Lemma 3.** Suppose \( (F_i, G_i, \psi_i), i = 1, 2 \), are adjunctions from \( A \rightarrow C \), and \( \alpha \) is a natural transformation from \( G_1 \rightarrow G_2 \). Then there is a unique natural transformation \( \beta \) from \( F_2 \rightarrow F_1 \) such that the diagram

\[
\begin{array}{ccc}
\text{Hom}_C(F_1(a), c) & \xrightarrow{\psi_{1ac}} & \text{Hom}_A(a, G_1(c)) \\
\downarrow{\beta'^{ac}_{\alpha}} & & \downarrow{\alpha^L_x} \\
\text{Hom}_C(F_2(a), c) & \xrightarrow{\psi_{2ac}} & \text{Hom}_A(a, G_2(c))
\end{array}
\]

commutes, for all \( a \in A, c \in C \). The dual statement holds.

**Proof:** The function

\[
\beta'^{ac}_{\alpha} = \psi^{-1}_{2ac}(\psi_{1ac})^L
\]

is the unique one making the diagram commutative. One verifies that the \( \beta'^{ac}_{\alpha} \) are the components of a natural transformation from \( \text{Hom}(F_1(a), c) \rightarrow \text{Hom}(F_1(a), c) \) (indeed, in each variable \( \beta' \) is the composition of three natural transformations). For fixed \( a \), by lemma 13.3 the transformation arises uniquely from \( \beta'^{ac}_{\alpha}(\iota_{F_1(a)}) \), which we denote \( \beta'_a \). From naturality of \( \beta'^{ac}_{\alpha} \) in \( a \), for \( f : a' \rightarrow a \), \( F_2(f)^R \beta'^{ac}_{\alpha} = \beta'^{ac}_{\alpha} F_1(f)^R \); applying both sides to \( \iota_{F_1(a)} \) yields \( \beta'_a F_2(f) = F_1(f) \beta'_{a'} \), which is to say \( \beta'_a \) is natural in \( a \).

203
THEOREM 4. Suppose $G$ is a functor from $B^{\text{op}} \times C$ to $A$, and each functor $G(b, -)$ has a left adjoint $F_b$. Then there is a unique functor $F : A \times B \to C$ agreeing with the $F_b$, such that the equivalence (1) holds. The dual statements holds.

PROOF: The object function of $F$ is clear; the arrow function must be specified. Suppose $f : b_2 \to b_1$, and let $F_1 = F_{b_1}$, $G_i = G(b_i, -)$, and $\alpha_c = G(f, \epsilon_c)$. By (1), (2) commutes; so by lemma 3 natural transformations $\beta_a$ are uniquely determined, where in fact $\beta_a = F(a, f)$. Thus, a functor $F(-, a)$ for each $a$ is uniquely determined; and this in turn uniquely determines a functor $F(-, -)$.

2. Cartesian closure. Adjunction with a parameter is used in developing the tensor product; it is also used in developing Cartesian closed categories, and these are presented here. A category with a specified terminal object, a specified product functor $X \times Y$, and a specified right adjoint to $- \times Y$ for each $Y$ is called Cartesian closed.

By theorem 13.4, Set is Cartesian closed. Another example is provided by Cat. The right adjoint to $- \times Y$ maps the category $Z$ to the functor category $Z^Y$. Let $\alpha$ denote the natural isomorphism, so that $\alpha_{X,Z} : \text{Hom}(X \times Y, Z) \to \text{Hom}(X, Z^Y)$. If $G \in \text{Hom}(X \times Y, Z)$ then $\bar{G} = \alpha_{X,Z}(G)$ maps $X$ to $Z^Y$. It maps the object $x$ of $X$ to the functor $\bar{G}(x)$, where $\bar{G}(x)(y) = G(a, b)$ (more properly $G((a, b)))$ and for $g : y \to y'$ $\bar{G}(x)(f) = G(x, f)$. It maps the arrow $f : x \to x'$ of $X$ to the natural transformation $\beta$, where $\beta_y = G(g, y_b)$. We leave the verification that this gives an adjunction to the reader.

By the dual of theorem 4, in a Cartesian closed category there is a unique functor $Z^Y$ from $C^{\text{op}} \times C$ to $C$, such that

$$\text{Hom}(X \times Y, Z) \cong \text{Hom}(X, Z^Y).$$

Note that in the exponential notation the arguments are written in reverse order; $Z^Y$ is called the exponentiation functor, and also the internal hom functor. In specific constructions, $Z^Y$ is often $\text{Hom}(Y, Z)$, equipped with additional structure to make it an object of the (concretely given) category $C$. Further, the component of the natural equivalence between $Z^{X \times Y}$ and $(Z^Y)^X$ is an isomorphism in the category $C$.

The notation $Z^Y$ for the exponentiation functor conflicts with the notation $Z^Y$ for the set of functions from $Y$ to $Z$. In Set they are the same; but not in other cases. Usually which is intended may be inferred from context; if not, the distinction will be made explicitly.

The counit of the adjunction of a Cartesian closed category is a system of arrows $\varepsilon_Y : Z^Y \times Y \to Z$; these are called the evaluation maps. Recalling that $\varepsilon_Y$ is the image of $\varepsilon_{Z^Y}$ under the adjunction map, one verifies that in Set, $\varepsilon(f, y) = y$. This is the case in various concrete Cartesian closed categories. The universality requirement of the counit is that given $h : W \times Y \to Z$, there is a map $\phi_h : W \to Z^Y$ such that $h = \varepsilon \circ (\phi_h \times \varepsilon_Y)$.

Suppose $C$ is a category with finite products. An object $Y$ in $C$ is called exponentiable if the functor $- \times Y$ has a right adjoint. This is so iff the counit naturality requirement is satisfied. $C$ is Cartesian closed if every object is exponentiable.

3. The tensor product. As seen in chapter 8, if $R$ is a commutative ring then the multilinear maps with pointwise operations form an $R$-module $\text{Mod}_R(A_1, \ldots, A_k; B)$. This yields a map from $(\text{Mod}_R)^k \times \text{Mod}_R$ to $\text{Mod}_R$ which is the object function of a functor. The image of a function $f : A_i \to A_i$ maps $g(a_1, \ldots, a_i, \ldots, a_k)$ to $g(a_1, \ldots, f(a_i), \ldots, a_k)$; and the image of a function $f : B \to B'$ maps $g(a_1, \ldots, a_k)$ to $f(g(a_1, \ldots, a_k))$. The equivalence between $\text{Hom}(A, \text{Hom}(B, C))$ and $\text{Hom}(A \times B, C)$ in Set induces an equivalence between $L(A; L(B; C))$ and $L(A, B; C)$. Indeed if $\bar{f} : A \to L(B; C)$ is linear then the map $f(a, b) = \bar{f}(a)(b)$ is bilinear; and if the bilinear map $f$ is given, each map $f_a(b) = f(a, b)$ is linear, and the map $\bar{f}(a) = f_a$ is also.
Although $L(A; B)$ is an $R$-module, $\text{Mod}_R$ is not Cartesian closed. What is true is that the Hom functor has a left adjoint, which is called the tensor product. To begin with, we seek an $R$-module $A \otimes B$ and a bilinear map $t : A \times B \rightarrow A \otimes B$ such that for any bilinear map $f : A \times B \rightarrow C$ there is a unique linear map $f' : A \otimes B \rightarrow C$ such that $f = f't$, as in the following diagram.

\[
\begin{array}{ccc}
A \times B & \xrightarrow{t} & A \otimes B \\
\downarrow{f} & & \downarrow{f'} \\
C & & 
\end{array}
\]

To construct the tensor product, note that the map from $A \times B$ to the free module $F$ generated by it (which we may consider to be inclusion) is not bilinear. However we can define a congruence relation on $F$ such that, composing with the canonical epimorphism, a bilinear map is obtained, in a necessary and sufficient manner. Namely, let $K$ be the submodule of $F$ generated by the elements

\[
\langle a_1 + a_2, b \rangle - \langle a_1, b \rangle - \langle a_2, b \rangle, \quad \langle a, b_2 + b_2 \rangle - \langle a, b_1 \rangle - \langle a, b_2 \rangle,
\]

\[
\langle ra, b \rangle - r \langle a, b \rangle, \quad \langle a, rb \rangle - r \langle a, b \rangle,
\]

for $a, a_i \in A$, $b, b_i \in B$, $r \in R$.

**Theorem 5.** $F/K$ is a tensor product for $A$ and $B$; the map $t$ is the composition $\eta \circ \subseteq$ where $\eta$ is the canonical epimorphism.

**Proof:** The generators of $K$ ensure that $t$ is bilinear. Given a bilinear map $f : A \times B \rightarrow C$ consider the unique map $f_1 : F \rightarrow C$. Since $f$ is bilinear $f_1$ is $0$ on $K$, so a map $f' : F/K \rightarrow C$ is induced, where $f' \eta = f_1$, so $f't = f$. Any such $f'$ induces a map $f_1$, so $f'$ is unique since $f_1$ is.

The notation $a \otimes b$ is used for $t(a, b)$; note that $\{a \otimes b : a \in A, b \in B\}$ generates $A \otimes B$. Given $f : A \rightarrow B$, $g : A' \rightarrow B'$, there is a unique map $f \otimes g : A \otimes B \rightarrow A' \otimes B'$ such that $(f \otimes g) \circ t = t \circ (f \times g)$, i.e., such that $(f \otimes g)(a \otimes b) = f(a) \otimes g(b)$. By uniqueness this makes $\otimes$ a functor from $\text{Mod}_R \times \text{Mod}_R$ to $\text{Mod}_R$.

Since $L(A, B; C)$ is in bijective correspondence with $L(A \otimes B; C)$, $L(A; L(B; C))$ is also. These bijective correspondences may be seen directly to be the components of a transformation which is natural in $A$ and $C$ (exercise 1). Using theorem 4 it is natural in $B$ also.

Since the functor $- \otimes B$ is left adjoint to $L(B; -)$, it preserves colimits. By theorem 2, the functor $A \otimes -$ also preserves colimits. In particular, tensor product distributes over coproduct, so that $(\oplus_i A_i) \otimes B \cong \oplus_i (A_i \otimes B)$ and $A \otimes (\oplus_i B_i) \cong \oplus_i (A \otimes B_i)$.

A tensor product of $R$-modules $A_1, \ldots, A_n$ is an $R$-module $P$ which yields a natural equivalence between $L(A_1, \ldots, A_n; B)$ and $L(P; B)$ for all $B$. The composition of equivalences

\[
L(A_1; L(A_2; L(A_3; B))) \cong L(A_1; L(A_2 \otimes A_3; B)) \cong L(A_1 \otimes (A_2 \otimes A_3); B)
\]

shows that $A_1 \otimes (A_2 \otimes A_3)$ is a tensor product for $A_1, A_2, A_3$. Generalizing, $A_1 \otimes (A_2 \otimes \cdots A_n)$ is a tensor product for $A_1, \ldots, A_n$; we will use $A_1 \otimes \cdots \otimes A_n$ to denote this.

Alternatively, a tensor product can be constructed as a quotient of the free module generated by the product of the modules, as in theorem 5. However, the first definition is more convenient in that standard machinery can be used, the second construction being more by way of an observation.

One basic fact is that there is a standard isomorphism from $(A_1 \otimes A_2) \otimes A_3$ to $A_1 \otimes (A_2 \otimes A_3)$, which is natural in all three arguments. The map $\langle a_1, a_2, a_3 \rangle \mapsto (a_1 \otimes a_2) \otimes a_3$ is multilinear, so factors uniquely through $A_1 \otimes (A_2 \otimes A_3)$, through the map $a_1 \otimes (a_2 \otimes a_3) \mapsto (a_1 \otimes a_2) \otimes a_3$. By an argument as above,
$L(A_1;L(A_2;L(A_3;B))) \cong L((A_1 \otimes A_2) \otimes A_3;B)$, so $(A_1 \otimes A_2) \otimes A_3$ is a tensor product, so there is a linear map with $a_1 \otimes (a_2 \otimes a_3) \mapsto (a_1 \otimes a_2) \otimes a_3$, which is clearly inverse to the first map. The verification that the system of isomorphisms is natural is routine and left to the reader.

A generalized associative law follows; details are left as exercise 3.

The map $(a,b) \mapsto b \otimes a$ is bilinear. By an argument similar to that just given, there is a standard natural equivalence from $B \otimes A$ to $A \otimes B$, which maps $b \otimes a$ to $a \otimes b$.

Some other useful facts are as follows.
- $R \otimes B \cong B$, via the map $r \otimes m \mapsto rm$, which is called the multiplication map; its inverse is $m \mapsto 1 \otimes m$.
- Likewise $A \otimes R \cong A$.
- $(f_1 \otimes g_1)(f_2 \otimes g_2) = (f_1 g_1) \otimes (f_2 g_2)$.
- $a \otimes 0 = 0 \otimes b = 0$.
- For $I$ finite
  \[
  \left( \sum_{i \in I} r_i a_i \right) \otimes b = \sum_{i \in I} (r_i a_i) \otimes b = \sum_{i \in I} r_i (a_i \otimes b),
  \]
  and similarly on the left.
- If $S \subseteq A$ generates $A$ and $T \subseteq B$ generates $B$ then $(a \otimes b: a \in S, b \in T)$ generates $A \otimes B$.

Since the tensor product preserves colimits, it is readily verified that the tensor product of two free $R$-modules, with bases $\{e_i: i \in I\}$ and $\{e_j: j \in J\}$, is the free $R$-module with basis $\{e_i \otimes e_j: i \in I, j \in J\}$. Indeed, $\otimes_i.j R \otimes R$ is a tensor product; it is generated by $\{e_i \otimes e_j\}$, and in fact is freely generated since the elements are in different summands. For example, the $m \times n$ matrices, with matrix addition and scalar multiplication, form a tensor product for $R^m$ and $R^n$. If $a \in R^m$, $b \in R^n$ then $a \otimes b$ is easily seen to be the matrix whose $i,j$ entry is $a_i b_j$; this matrix is known as the tensor product of the vectors, and by various other names.

It is also easy to compute the tensor product of two finite commutative groups $A, B$. This is the direct sum of the groups $Z_q \otimes Z_r$ where the $Z_q$, $Z_r$ are the summands in the decompositions of $A, B$ into prime power cyclic subgroups. Suppose $q$ is a power of $p_1$ and $r$ a power of $p_2$. If $p_1, p_2$ are distinct then $p_1(1 \otimes 1) = 0$ and $p_2(1 \otimes 1) = 0$, so $1 \otimes 1 = 0$ since gcd$(p_1, p_2) = 1$. If $p_1 = p_2$ then $Z_q \otimes Z_r = Z_s$, where $s = \min(q, r)$.

4. Non-commutative rings. If non-commutative rings are considered the tensor product may still be defined, but it becomes necessary to consider the “sidedness” of the modules. Suppose $R, S, T$ are rings, not necessarily commutative. For this section we write $R \text{Mod}_S$ for the left $R$, right $S$ modules. If $R$ or $S$ is absent it is assumed to be $Z$. If $R$ is commutative an $R$-module can be considered an object in $R \text{Mod}_R$ by defining the right action by $rx = rx$.

The foregoing notation will occasionally be used later in the text; but usually $\text{Mod}_R$ will denote the left $R$-modules.

Suppose
\[ A \in R \text{Mod}_T, \quad B \in S \text{Mod}_R, \quad C \in S \text{Mod}_T. \]

Hom$(B, C)$ in $S \text{Mod}$ becomes an Abelian group in the usual manner, and becomes an object in $R \text{Mod}_T$ with the actions
\[ (rf)(b) = f(br), \quad (ft)(b) = f(b)t. \]

Further, in this way Hom$(-, -)$ becomes the object map of a functor from $(S \text{Mod}_R) \times (S \text{Mod}_T)$ to $R \text{Mod}_T$, with the usual function maps $f \mapsto f^L, f^R$. For example,
\[ ((rh)f)(b) = (rh)(f(b)) = h(f(b)r) = h(f(br)) = (hf)(br) = (r(hf))(b), \]
i.e., $f^R(rh) = rf^R(h)$; we leave the rest to the reader.

206
Reversing the order for a reason to be noted shortly. the equivalence between \( \text{Hom}(A, \text{Hom}(B, C)) \) and \( \text{Hom}(B \times A, C) \) in Set induces an equivalence in Ab, provided we modify the second functor to consider only those functions \( f \) satisfying

\[
\begin{align*}
  f(b_1 + b_2, a) &= f(b_1, a) + f(b_2, a), \\
  f(b, a_1 + a_2) &= f(b, a_1) + f(b, a_2), \\
  f(sb, a) &= sf(b, a), \\
  f(b, at) &= f(b, a)t, \\
  f(br, a) &= f(b, ra).
\end{align*}
\]

(3)

A left adjoint \( B \otimes A \), also written \( B \otimes_R A \), to \( \text{Hom}(B, C) \) (where \( B \) is a parameter) may thus be constructed by taking the quotient \( F/K \) where \( F \) is the free left \( S \) right \( T \) module generated by \( B \times A \), and \( K \) enforces equations (3). Often \( sB_R \) is written to denote the ring actions, and with the order as given we have

\[ sB_R \otimes_R RA_T = sC_T. \]

One verifies that both \( B \otimes - \) and \( - \otimes A \) preserve colimits, and that the associative law

\[ (A \otimes B) \otimes C = A \otimes (B \otimes C) \]

holds. Also, the case where \( R \) is commutative and \( A, B \) are \( R \)-modules (so that \( S = R \) and \( T = Z \)) is easily seen to yield the same \( R \)-module as that given in the previous section. More generally, \( R, S, T \) can be considered as acting on \( q \text{Mod} \) where \( Q \) is some commutative ring; indeed further generalizations are possible.

We note that the multiplication maps \( R_R \otimes RA_T \rightarrow AT \) and \( sB_R \otimes R \rightarrow sB \) are isomorphisms, as in the commutative case.

Another important special case arises when \( h : R \rightarrow S \) is a ring homomorphism (and for simplicity \( T = Z \)). The map \( s \mapsto sh(r) \) gives a right action of \( R \) on \( S \), so that \( S \) becomes an object in \( s\text{Mod}_R \). The map \( S \otimes - \) is said to “change rings”; it converts an \( R \)-module to an \( S \)-module in a universal fashion. The functor \( \text{Hom}(S, C) \) is readily verified to be naturally equivalent to \( G(C) \), where \( G(C) \) is the \( R \)-module whose addition is the same as that of \( C \), and whose action is \( \langle r, c \rangle \mapsto h(r)c \). If \( R = Z \) \( G \) is the forgetful functor from \( s\text{Mod} \) to \( Ab \).

The module of fractions is a special case of changing rings. \( R \) is commutative and \( S \) is \( R_S \) for a multiplicative subset \( S \subseteq R^\times \). \( R_S \otimes M \) is isomorphic to the module \( M_S \) defined in section 8.6, via the map \( (r/s) \times m \mapsto (rm)/s \). Indeed, \( (r/s, m) \mapsto (rm)/s \) is readily verified to be well-defined and \( R \)-bilinear, so induces the desired \( R \)-linear map. The map \( m/s \rightarrow (1/s) \otimes m \) is also well defined; if \( t(m_1s_2 - m_2s_1) = 0 \) then \( (1/s_1) \otimes m_1 = (1/(ts_1s_2) \otimes tm_1s_2 = (1/(ts_1s_2) \otimes tm_2s_1/(1/s_2) \otimes m_2. \) Finally, the second map is inverse to the first.

The tensor product may in fact be generalized to a cocomplete Abelian category over a ring (see [Mitchell]).

5. Graded modules. The definition of a graded ring was mentioned in chapter 7. To repeat, if \( G \) is a monoid (with operation +, commonly commutative), a \( G \)-grading of a ring \( R \) is a direct sum decomposition \( R = \oplus_{i \in G} R_i \) of the additive group into subgroups, where \( R_iR_j \subseteq R_{i+j} \). Most commonly \( G \) is \( \mathbb{N} \) or \( \mathbb{Z} \).

If \( R \) is a \( G \)-graded ring, a \( G \)-grading of an \( R \)-module \( M \) is a direct sum decomposition \( M = \oplus_{i \in G} M_i \) of the additive group into subgroups, where \( R_iM_j \subseteq M_{i+j} \). More general situations occur; for example for \( R \) ungraded (or “concentrated at 0”) a direct sum decomposition of \( M \) as \( \oplus_i M_i \) satisfies \( RM_j \subseteq M_j \), and \( M \) is called graded. Facts stated below can be adapted to other situations as necessary.

Elements of \( R_i \) (or \( M_j \)) are called homogeneous; for example in a homogeneous polynomial all terms have the same degree. Any element in \( R \) (or \( M \)) can be written uniquely as a sum of homogeneous elements from distinct \( R_i \) (or \( M_j \)); this sum will be called the homogeneous decomposition.
Theorem 6. For a submodule $N$ of a graded module $M$, the following are equivalent.

a. $N = \oplus_i (N \cap M_i)$.

b. $N$ is generated by homogeneous elements.

c. If $m \in N$ and $m = m_1 + \cdots + m_r$ is its homogeneous decomposition then for each $s$, $m_s \in N$.

Proof: For $m \in M$ let $\sum_s m_s$ be its homogeneous decomposition. Suppose a holds, and $m \in N$; then $m_s$ must be in $N$, and b follows. Suppose b holds, and let $m = \sum_r r_t g_t$ is an expression in terms of homogeneous generators; then $m_s$ is the sum of those $r_t g_t$ which are in the same $M_i$ as $m_s$. This proves that $m_s \in N$. Suppose c holds and $m \in N$; then each $m_s$ is in $N$, proving a.

When the above conditions hold, the submodule is called homogeneous. Note that it is a graded module. A homogeneous ideal is a homogeneous submodule of $R$. For an example, a principal ideal in $pR[x]$ in the ring of polynomials is easily seen to be homogeneous if $p$ is homogeneous. Note that a finitely generated homogeneous submodule is finitely generated by homogeneous elements.

A homomorphism $h : M \to N$ of graded $R$-modules is called graded if $h[M_j] \subseteq N_j$. The graded $R$-modules with the graded homomorphisms form a category. Many constructions from $Mod_R$ carry over to this category, “gradewise”. The same is also true of the homogeneous submodules of a graded module. Examples include the following.

- The product of two graded $R$-modules may be graded, by letting $(M \times N)_j$ equal $M_j \times N_j$; each member of the product is in only finitely many $M_j$ and finitely many $N_j$, so only finitely many of these product submodules.

- A similar argument shows that any coproduct as modules is graded.

- The image of a homomorphism is homogeneous.

- The quotient $M/N$ of a graded module $M$ by a homogeneous submodule $N$ is graded. Writing $M_i/N$ for $\{x + N : x \in M_i\}$, $M_i/N$ is a subgroup of $M/N$, $R_j(M_i/N) \subseteq R_j M_i/N$, and $M/N$ is the direct sum of these (if $x_i + N = x_j + N$ where $i \neq j$ then $x_i - x_j \in N$, so $x_i, x_j \in N$). Further, $M_i/N$ and $M_i/(N \cap M_i)$ are isomorphic as $R_0$-modules.

- The family of homogeneous submodules is closed under arbitrary intersection (use theorem 6.c).

- The family of homogeneous submodules is closed under sum (use theorem 6.b).

The image $f[M]$ of a graded homomorphism $f : M \to N$ is a graded submodule of $N$. It is readily verified that $f[M]$ is generated by $\{f(m)\}$ for $m$ homogeneous; $f[M] = \sum_j \otimes_j h[M_j]$; and $h[M_j] = h[M] \cap N_j$. The coinage-image factorization as modules is a factorization as graded modules.

The kernel $\text{Ker}(f)$ of a graded homomorphism $f : M \to N$ is a graded submodule of $M$. It is readily verified that $\text{Ker}(f)$ is generated by $\{m : f(m) = 0\}$ for $m$ homogeneous; and $\text{Ker}(f) = \sum_j \text{Ker}(f \mid M_j)$. Also, if $N$ is a homogeneous submodule of $M$ then $M/N$ is generated by $\{m + N\}$ for $m$ homogeneous; and $M/N$ equals $\oplus_j (M_j + N)/N$, or $\oplus_j M_j/(M_j \cap N)$. To see that the sum of the $(M_j + N)/N$ is direct, note that if $m_1 + \cdots + m_r \in N$ then $m_i$ must be in $M_i \cap N$ where $m_i \in M_i$.

The preceding facts are readily adapted to the category of graded rings and graded homomorphisms. The image is a graded ring, and the kernel a homogeneous ideal.

Graded rings are sometimes constructed as follows. Let $R_i$ be an Abelian group, and suppose for each $i, j$ there is a map from $R_i \times R_j$ to $R_{i+j}$, such that for $r_i, r_i' \in R_i$, $r_j, r_j' \in R_j$, and $r_k \in R_k$ the following hold.

$$(r_i + r_i')r_j = r_ir_j + r_i'r_j, \quad r_i(r_j + r_j') = r_ir_j + r_ir_j', \quad r_i(r_j r_k) = (r_ir_j)r_k$$

Then there is a unique distributive associative operation on $R = \oplus_i R_i$ agreeing with the given maps; and it restricts to $R_0$. Provided there is an element in $R_0$ which acts as the identity in the given maps, $R$ becomes a ring. If $R_0$ is commutative $R$ is in fact an $R_0$-algebra.
6. The tensor algebra. Let $R$ be a commutative ring, and $M$ an $R$-module. The tensor product $M^k$ of $M$ with itself $k$ times is defined as in section 3, using right association of the binary tensor product. $M^{\otimes 0}$ may be defined to be $R$.

The natural equivalence of exercise 2 from $M^k \times M^l$ to $M^{(k+l)}$ may be considered a binary operation, and is bilinear. The notation $x \otimes y$ is commonly used for this operation. Given $a = a_1 \otimes \cdots \otimes a_k$ in $M^k$, $b = b_1 \otimes \cdots \otimes b_l$ in $M^l$, and $c = c_1 \otimes \cdots \otimes c_m$ in $M^m$, we have $(a \otimes b) \otimes c = a \otimes (b \otimes c)$. This then follows for any $a, b, c$ in the given modules.

As described in section 5, under circumstances such as this the direct sum $\oplus_k M^k$ may be given the structure of an $R$-algebra, where the ring is graded, with homogeneous submodules the $M^k$. This algebra is called the tensor algebra of the $R$-module $M$. We will use $M^\otimes$ to denote it. The tensor algebra has a variety of uses, especially in the case of vector spaces. There are two commonly encountered subalgebras, which will be defined shortly; these may be defined for any $R$-module.

For an $R$-module $M$, $M$ is included in $M^{\otimes}$ in an obvious manner. This inclusion is a universal arrow from $M$ to the forgetful functor from the $R$-algebras to the $R$-modules. To see this, suppose $f$ is an $R$-module homomorphism from $M$ to an $R$-algebra $A$. This has a unique extension to an $R$-algebra homomorphism from $M^{\otimes}$ to $A$, namely that where $f(m_1 \otimes \cdots \otimes m_k) = f(m_1) \cdots f(m_k)$. It follows that $M^{\otimes}$ is a functor, left adjoint to the forgetful functor. One readily verifies that for $f : M \to N$, the map $f^{\otimes} : M^{\otimes} \to N^{\otimes}$ is that taking $m_1 \otimes \cdots \otimes m_k$ to $f(m_1) \otimes \cdots \otimes f(m_k)$.

Let $\equiv$ be the least $R$-algebra congruence relation on $M^{\otimes}$ generated by the equations $x \otimes y = y \otimes x$. The quotient algebra is called the symmetric tensor algebra; we will use $M_S^{\otimes}$ to denote it. The congruence relation is clearly generated by the equations $a_{\pi(1)} \otimes \cdots \otimes a_{\pi(k)} = a_1 \otimes \cdots \otimes a_k$ where the right side is a generator of $M^k$ and $\pi$ is a permutation; indeed, $\pi$ may be restricted to a generating set of the symmetric group, such as the transpositions. It follows that $M_S^{\otimes}$ is graded, with the homogeneous submodule of degree $k$ being the quotient of $M^k$ by the restriction of $\equiv$.

$M_S^{\otimes}$ is clearly commutative. Also, $M$ is a submodule, since there is no collapsing in $M$ from the congruence relation. By an argument as above, the inclusion map is a universal arrow from $M$ to the forgetful functor from the commutative $R$-algebras to the $R$-modules; and $M_S^{\otimes}$ is a functor, left adjoint to the forgetful functor. It is also readily verified that for any $N$, the linear maps from $M_S^{\otimes}$ to $N$ are in bijective correspondence with the symmetric multilinear maps from $M^k$ to $N$.

The quotient of $M^{\otimes}$ by the least congruence relation containing the equations $x \otimes x = 0$ is called the alternating tensor algebra, or exterior algebra. We use $M^\wedge$ to denote it. From $(x+y) \otimes (x+y) = 0$, $x \otimes x = 0$, and $y \otimes y = 0$, it follows that $x \otimes y = -y \otimes x$. Using this, and the fact that for homogeneous $x$ and $y$, $x \otimes y$ and $y \otimes x$ are in the same homogeneous submodule, it follows that the congruence relation is generated by the equations $x \otimes x = 0$ for homogeneous $x$. Thus, $M^\wedge$ is graded. The image of $m_1 \otimes \cdots \otimes m_k$ under the canonical epimorphism is customarily denoted $m_1 \wedge \cdots \wedge m_k$, especially in the case of vector spaces.

Similarly to $M_S^{\otimes}$, $M$ is a submodule of $M^\wedge$, and the inclusion map is a universal arrow to the $R$-algebras such that $x^2 = 0$ for all $x$. $M^\wedge$ is a functor, left adjoint to the forgetful functor from such algebras to $R$-modules. For any $N$, the linear maps from $M^\wedge$ to $N$ are in bijective correspondence with the alternating multilinear maps from $M^k$ to $N$.

The tensor algebra over a vector space is of particular interest; for one thing a basis can be given explicitly, indeed for a free module. For an $n$-dimensional free $R$-module $M$, an element of $M^k$ equals

$$\sum_{i_1, \ldots, i_k} r_{i_1 \ldots i_k} e_{i_1} \otimes \cdots \otimes e_{i_k},$$

where $r_{i_1 \ldots i_k} \in R$ and $e_1, \ldots, e_n$ is a basis for $M$. If $e'_i$ is a different basis for $M$, and $e'_i = \sum_j a_{ij} e_j$, then by
multilinearity it follows that
\[ e'_{i_1} \otimes \cdots \otimes e'_{i_k} = \sum_{j_1, \ldots, j_k} a_{i_1,j_1} \cdots a_{i_k,j_k} e_{j_1} \otimes \cdots \otimes e_{j_k}. \]

Although some facts may be stated more generally, from hereon we consider the case of an \( n \)-dimensional vector space \( V \) over a field \( F \). Commonly \( F \) is the real numbers, but as usual in linear algebra the case of an arbitrary field is mostly identical. An element of the tensor algebra is called a tensor, and an element of \( V^\otimes k \) is said to be \( n \)-dimensional and rank \( k \). The coefficients in terms of the basis of \( V^\otimes k \) are called the components of the tensor, with respect to the basis for \( V \). A common definition of a real tensor is a system of \( n^k \) real numbers with respect to some basis, which transforms according to the law just given (further remarks are made below).

In some cases the components might be more meaningful in some bases than others. For example in Euclidean space the orthonormal bases are of special interest, and a particular such might be singled out as the “laboratory frame”. In other cases the bases are on an equal footing, for example the tangent space at a point of a manifold, to be considered further in chapter 24.

Operations on the tensor algebra may be viewed in terms of the components. In particular, addition and scalar multiplication are componentwise. Writing \( r \) for the tensor whose components are \( r_{i_1,\ldots,i_k} \), etc., the components of the product \( t = r \otimes s \) are
\[ t_{i_1,\ldots,i_k,j_1,\ldots,j_l} = r_{i_1,\ldots,i_k} s_{j_1,\ldots,j_l}, \]
as is readily verified. Operations may be described by specifying their effect on the components, provided that description is the same for all bases. For example, if \( r \) is a rank \( k \) tensor and \( \pi \) is a \( k \)-element permutation let \( r_{\pi} \) denote the tensor whose components are \( r_{\pi(i_1),\ldots,\pi(i_k)} \).

\( L(V^{k+l}, F) \) is readily verified to be a tensor product of \( L(V^k, F) \) and \( L(V^l, F) \). Thus, the tensor algebra over \( V^* \) may be given as the multilinear maps. Writing \( \chi_i \) for the dual basis element to \( e_i \), the basis element \( \chi_{i_1} \otimes \cdots \otimes \chi_{i_k} \) corresponds to the multilinear map \( \chi_{i_1,\ldots,i_k} \) which is 1 at \( (e_{i_1}, \ldots, e_{i_k}) \), and 0 at any other \( k \)-tuple of basis elements. Given multilinear maps \( r, s \in L(V^k, F) \) and \( t, \tau \in L(V^l, F) \), their tensor product is the multilinear map \( t \otimes \tau \in L(V^{k+l}, F) \) where \( t(x_1, \ldots, x_k, y_1, \ldots, y_l) = r(x_1, \ldots, x_k)s(y_1, \ldots, y_l) \). Note that the transformation law as stated above involves the \( a_{ij} \) where \( \chi'_i = \sum_j a_{ij} \chi_j \), and in defining the tensor algebra this way \( V \) plays a “phantom” role.

A tensor is called symmetric if \( r_{\pi} = r \) for any permutation \( \pi \) (this holds for the components in some basis iff it holds in any basis). The canonical epimorphism from \( V^\otimes k \) to \( S^\otimes k \) maps the tensor with components \( r_{i_1,\ldots,i_k} \) to the element of \( V^\otimes k \) whose coefficient at a multiset is the sum of the coefficients of the sequences which map to the multiset. From this, \( \eta(r_{\pi}) = \eta(r) \). Suppose \( F \) has characteristic 0; then \( \eta((1/k!) \sum_{\pi} r_{\pi}) = \eta(r) \). The tensor \( (1/k!) \sum_{\pi} r_{\pi} \) is called the symmetrization of \( r \). A tensor is symmetric iff it equals its symmetrization. The canonical epimorphism yields an \( R \)-module isomorphism between the symmetric tensors and the elements of the symmetric tensor algebra. A multilinear map to \( F \) is a symmetric tensor iff it is a symmetric multilinear map. The symmetric tensor algebra of an \( n \)-dimensional vector space is isomorphic as a graded algebra to the multivariate polynomials in \( n \) variables.

Similar remarks for antisymmetry. A tensor \( r \) is antisymmetric iff \( r_{\pi} = (-1)^{sg(\pi)} r \). In characteristic 0, the tensor \( (1/k!) \sum_{\pi} (-1)^{sg(\pi)} r_{\pi} \) is called the antisymmetrization of \( r \). A rank 2 tensor is the sum of its symmetrization and its antisymmetrization. The rank \( k \) submodule of the antisymmetric tensor algebra is 0 for \( k > n \), and has dimension \( \binom{n}{k} \) for \( k \leq n \).

In many applications, especially involving “physical” quantities, the notion of tensor considered above is too restrictive, and the transformation properties must be considered in more detail. Suppose \( e_i \) is a basis
for $V$, $e'_i$ is another basis, $e'_i = \sum_j a_{ij} e_j$, and $e_i = \sum_j \bar{a}_{ij} e'_j$. Substituting the second expression in the first and simplifying yields $\sum_k a_{ik} \bar{a}_{kj} = \delta_{ij}$; similarly $\sum_k \bar{a}_{ik} a_{kj} = \delta_{ij}$.

If $r = \sum_i r_i e_i$ then by substituting the expression for $e_i$, $r = \sum_j (\sum_i \bar{a}_{ij} r_i)e'_j$. The components of $r$ in the new basis become $\sum_i \bar{a}_{ij} r_i$; they transform “contravariantly” to the expression giving the elements of the new basis in terms of the old (in matrix terms, via the inverse transpose). An ordinary vector (say, given by its components in the laboratory system) is said to be a contravariant rank 1 tensor. Note that if the transformation is orthogonal, then the expressions become identical; it is only when “oblique” coordinate systems are allowed that the distinction becomes apparent.

Let $\chi_i$ be the dual basis elements for $V^*$, where $\chi_i(e_j) = \delta_{ij}$; and let $\chi'_i$ be the dual basis to $e'_i$. Then

$$\chi'_i(e_j) = \chi'_i(\sum_k \bar{a}_{jk} e'_k) = \sum_k \bar{a}_{jk} \delta_{ik} = \bar{a}_{ji};$$

and $(\sum_k \bar{a}_{ki} \chi_k)(e_j)$ also equals $\bar{a}_{ji}$. Thus, $\chi'_i = \sum_k \bar{a}_{ki} \chi'_k$. It follows that the components of a linear functional, expressed as a linear combination of the dual basis elements in either basis, transform by the same expression as the basis change expression, or “covariantly”.

To provide a mathematical setting for this phenomenon, the “mixed” tensor product $V^* \otimes V \otimes \cdots \otimes V$ is introduced. These vector spaces are graded by $\mathbb{N} \times \mathbb{N}$, and the product of

$$e_{i_1} \otimes \cdots \otimes e_{i_k} \otimes \chi_{j_1} \otimes \cdots \otimes \chi_{j_l}$$

and

$$e_{p_1} \otimes \cdots \otimes e_{p_r} \otimes \chi_{q_1} \otimes \cdots \otimes \chi_{q_s}$$

may be defined to be

$$e_{i_1} \otimes \cdots \otimes e_{i_k} \otimes e_{p_1} \otimes \cdots \otimes e_{p_r} \otimes \chi_{j_1} \otimes \cdots \otimes \chi_{j_l} \otimes \chi_{q_1} \otimes \cdots \otimes \chi_{q_s}.$$

For elements in $V^* \otimes V \otimes \cdots \otimes V$, the components transform contravariantly in the first $k$ indices, and covariantly in the last $l$. These tensors are said to be of type $(k, l)$.

The tensor products where the factor may be either $V$ or $V^*$ in any position are also considered, although the need for them arises sufficiently infrequently that the first formulation is more common. The homogeneous spaces may be taken as graded by the monoid of strings over \{0, 1\}, with concatenation as the operation.

The mixed tensors may be defined as multilinear forms; the tensors of type $(k, l)$ may be taken as $L(V^\otimes k, V^\otimes l; F)$. Initially, $V$ plays a phantom role, but it reappears as $V^{**}$, the contravariant vectors.

In applications tensors often arise as multilinear maps. For example when $F = \mathbb{R}$, for $f : V \mapsto \mathbb{R}$ the $k$th derivative $f^{(k)}$ is a multilinear map. Under suitable restrictions it is symmetric, and Taylor’s theorem with remainder

$$f(x + \Delta x) = \sum_{k=0}^l \frac{1}{k!} f^{(k)}(\Delta x_1, \ldots, \Delta x_l) + O(\Delta x^{k+1})$$

holds; see [Dieudonné]. Being a multilinear form on $V$, $f^{(k)}$ is a fully covariant type $(0, k)$ tensor. This may be seen directly in the case $k = 1$ as follows. The components of the gradient in the unprimed system are $\partial f / \partial x_i$; and in the primed system $\partial f / \partial x'_i$. The latter is the derivative along $e'_i$, which equals $\sum_k a_{ik} \partial f / \partial x_k$.

Rank 2 tensors also arise as linear maps. The natural equivalences

$$L(V; V) \cong L(V; V^{**}) \cong L(V; L(V^*, V)) \cong L(V, V^*; V) \cong L(V^*, V; V)$$

211
show that linear maps are type $(1,1)$. For an example, suppose $f : V \to V$ is a vector valued function (e.g., a function assigning to each point of Euclidean space in laboratory coordinates a vector in laboratory coordinates, a contravariant vector field on Euclidean space). The quantities $\partial f_i / \partial x_j$ are the components of a tensor of type $(1,1)$, contravariant in $i$ and covariant in $j$. As a linear map it gives $\Delta y$ as a linear function of $\Delta x$. Writing vectors as column vectors, the matrix of the map has components $\partial f_i / \partial x_j$. See section 24.12 for some relevant facts.

For simplicity we have written all tensor component indices as subscripts. A common device is to write contravariant indices as superscripts and covariant indices as subscripts. This may also be done with other quantities, such as basis elements $e_i$, dual basis elements $e^i$ (rather than $\chi_i$), and the Kronecker delta function $\delta^i_j$. In the “Einstein summation convention”, in a product of such factors, an identical index occurring in an upper and lower position is summed over. Examples include $r = r^i e_i$ for a contravariant vector, $r^i = a^j r_j$ for the contravariant vector transformation law, and $\delta^i_k \delta^j_l = \delta^i_j$ for the transformation coefficient inverse relationship.

A topic of recent interest is the tensor algebra of a $\mathbb{Z}_2$ graded, or “super”, vector space $V = V_0 \oplus V_1$. Each $V^\otimes k$ may be $\mathbb{Z}_2$ graded, with the even (odd) subspace being $\otimes_{i_1, \ldots, i_k} V_{i_1} \otimes \cdots \otimes V_{i_k}$ where $i_1 + \cdots + i_k = 0$ (1), addition being mod 2. This results in an $\mathcal{N} \times \mathbb{Z}_2$ grading of $V^\otimes$. Let $\equiv$ be the congruence relation generated the equations $v_{\pi(1)} \otimes \cdots \otimes v_{\pi(n)} = \pm v_1 \otimes \cdots \otimes v_n$, where $\pi$ is a transposition and the sign is negative if both transposed elements are in $V_1$. The ideal of $\equiv$ is homogeneous, and the quotient may be taken. We omit further discussion.

### 7. Tensor product of algebras.

Suppose $A$ and $B$ are $R$-algebras over a commutative ring $R$. The map $\langle a_1, b_1, a_2, b_2 \rangle \mapsto a_1 a_2 \otimes b_1 b_2$ from $A \times B \times A \times B$ to $A \otimes B$ is multilinear, and is readily seen to induce a bilinear map $\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle \mapsto a_1 a_2 \otimes b_1 b_2$ from $(A \otimes B) \times (A \otimes B)$ to $A \otimes B$. It is also readily seen that with this map as multiplication, $A \otimes B$ becomes an $R$-algebra; the multiplicative identity is $1 \otimes 1$. If $A$ and $B$ are commutative so is $A \otimes B$.

The map $a \mapsto a \otimes 1$ ($b \mapsto 1 \otimes b$) is an $R$-algebra homomorphism from $A$ ($B$) to $A \otimes B$. In the category of commutative $R$-algebras, $A \otimes B$ is a coproduct object, with these maps as injections. Indeed, given $R$-algebra homomorphisms $f : A \to C$ and $g : B \to C$, the map $\langle a, b \rangle \mapsto f(a)g(b)$ is bilinear, so induces an $R$-linear map $a \otimes b \mapsto f(a)g(b)$ from $A \otimes B$ to $C$. This map is an $R$-algebra homomorphism since $f(a_1 a_2)g(b_1 b_2) = (f(a_1)g(b_1))(f(a_2)g(b_2))$; remaining requirements for a coproduct are readily verified. For a universality property in the noncommutative case, see [Jacobson].

For commutative algebras $A \otimes B$ may also be characterized as the pushout in CRng of the diagram with maps $R \to A$ and $R \to B$, the ring homomorphisms corresponding to the algebras.

### Exercises.

1. Given $f \in L(A; L(B; C))$ let $\hat{f}$ denote the corresponding map in $L(A \otimes B; C)$, i.e., that where $\hat{f}(a \otimes b) = f(a)(b)$. Show that $f \mapsto \hat{f}$ is linear, and hence an isomorphism of $R$-modules. Show that this system of isomorphisms is natural in $A$ and $C$. Hint: For naturality in $A$ it must be shown that for $h : A \to A'$, $\hat{f} \circ (h \otimes \cdot) = \hat{f} \circ h$. This may be shown by applying both sides to $a \otimes b$. For naturality in $C$ suppose $h : C \to C'$, and given $f \in L(A; L(B; C))$ let $f'$ denote the corresponding map in $L(A; L(B; C'))$. Then $f'(a)(b) = h(f(a)(b))$, and $\hat{f} = h \hat{f}$ follows.

2. Show that there is a standard natural isomorphism between $(A_1 \otimes \cdots \otimes A_m) \otimes (A_{m+1} \otimes \cdots \otimes A_{m+n})$ and $A_1 \otimes \cdots \otimes A_{m+n}$. It maps $(a_1 \otimes \cdots \otimes a_m) \otimes (a_{m+1} \otimes \cdots \otimes a_{m+n})$ to $a_1 \otimes \cdots \otimes a_{m+n}$. Hint: Use induction on $m$.

3. An ordered binary tree with $n$ leaves specifies an associative grouping under a binary operation of the items at the leaves, from left to right. Considering the tensor product of $R$-modules, an $n$ argument functor $T$ is specified. Let $T_0$ be the right associated functor, chosen as “the” tensor product. Show that

212
there is a standard natural equivalence of \( T \) to \( T_0 \); it maps the specified grouping of \( a_1 \otimes \cdots \otimes a_n \) to the right associated grouping. Hint: Proceed by induction on the height of the tree and use exercise 2.

4. Prove directly that linear transformations are type \( \langle 1, 1 \rangle \) tensors. Hint: Let \( e_i \) and \( e'_j \) be the unprimed and primed bases, with \( e'_j = \sum_j a_{ij} e_j \). Let \( \lambda \) be the map, with unprimed and primed components \( \lambda_{ij} \) and \( \lambda'_{ij} \), so that \( \lambda(e_i) = \sum_j \lambda_{ij} e_j \) and \( \lambda(e'_j) = \sum_j \lambda'_{ij} e'_j \). Then

\[
\lambda(e'_j) = \sum_t \delta_{it} \lambda(e'_t) = \sum_{t,r} a_{ir} \tilde{a}_{rt} \lambda(e'_t) = \sum_r a_{ir} \lambda(e_r) = \sum_{r,s} a_{ir} \lambda_{rs} e_s = \sum_j \sum_{r,s} \tilde{a}_{sj} a_{ir} \lambda_{rs} e'_j.
\]

1. Limits in functor categories. Suppose $C$ is a standard category, and $J, K \in \text{Cat}$. As when $C \in \text{Cat}$ (see chapter 18) there is a natural equivalence (now in $J$ and $K$ only) $C^{J \times K} \cong (C^K)^J$. Writing $\bar{F}$ for the image of $F : J \times K \to C$, for $j \in J, k \in K$, $\beta : j \mapsto j'$, and $\gamma : k \mapsto k'$, $\bar{F}(j)(k) = F(j, k)$, $\bar{F}(j)(\gamma) = F(j, \gamma)$, and $\bar{F}(\beta)$ is a natural transformation whose $k$ component is $F(\beta, k)$. There is also a natural equivalence $C^{J \times K} \cong (C^J)^K$; writing $\bar{F}^*$ for the image of $F : J \times K \to C$, $\bar{F}^*(j)(k) = F(j, k)$, etc.

A functor from $K$ to $C^J$ is an element of $(C^J)^K$. It equals $\bar{F}^*$ for some $F \in C^{J \times K}$. For the following, for $j \in J$ let $F_j$ denote $\bar{F}(j)$, an object of $C^K$; and for $k \in K$ let $F_k$ denote $\bar{F}^*(k)$, an object of $C^J$.

**Theorem 1.** With notation as above, suppose $F_j$ has a limit $\langle d_j, \alpha_j \rangle$ for each $j \in J$; let $\alpha^*_k(j) = \alpha_j(k)$. There is a unique functor $D \in C^J$, where $D(j) = d_j$, such that $k \mapsto \alpha^*_k$ is a cone from $D$ to $\bar{F}^*$ in $C^J$. Further, this is a limit cone. If $D$ has a limit $d$ then $d$ is a limit for $F$, with the obvious cone. The dual statements hold.

**Proof:** Suppose $\beta : j \mapsto j'$; by the universality of $d_j$, there is a unique arrow $D(\beta) : d_j \to d_j'$ such that for all $k F_k(\beta)\alpha_j(k) = \alpha_j(k)D(\beta)$. That $D$ is a functor follows by uniqueness. That $k \mapsto \alpha^*_k$ is a cone follows because, for $\gamma : k \mapsto k'$

$$F_j(\gamma)\alpha^*_k(j) = F_j(\gamma)\alpha_j(k) = \alpha_j(k') = \alpha^*_{k'}(j),$$

since $F_j(\gamma)$ is the $j$ component of the natural transformation from $F_k$ to $F_{k'}$ induced by $\gamma$. Let $\langle D' , \alpha'^* \rangle$ be any other cone to $\bar{F}^*$, and let $\alpha'_j(k) = \alpha'^*_k(j)$. Then $\alpha'_j$ is a cone from $d'_j$ to $F_j$, so there is a unique arrow $\theta_j : d'_j \to d_j$ such that $\alpha_j(k)\theta_j = \alpha'_j(k)$, all $k \in K$. The $\theta_j$ are the components of a natural transformation, that is, $\theta_j D'(\beta) = D(\beta)\theta_j$, because both are arrows $\psi : d'_j \to d_j$, satisfying $\alpha_j \psi = \alpha'_j D'(\beta)$ for all $j \in J$, as we leave to the reader to verify. Suppose $d, \delta$ be a limit for $D$; write $\alpha_{jk}$ for $\alpha_j(k)$, and let $\epsilon_{jk} = \alpha_{jk}\delta_j$. This is a cone to $F$; for $\beta : j \mapsto j'$,

$$F_k(\beta)\alpha_{jk}\delta_j = \alpha_{j'k}D(\beta)\delta_j = \alpha_{j'k}\delta_{j'},$$

and for $\gamma : k \mapsto k' F_j(\gamma)\alpha_{jk}\delta_j = \alpha_{j'k}\delta_j$. If $\langle d', \epsilon' \rangle$ is any other cone to $F$, $\epsilon'_j$ factor through $d_j$, and these maps in turn factor through $d$; remaining details are left to the reader.

It is also true that (with notation as above) if $d$ is given as a limit for $F$ then when $D$ exists $d$ is a limit for $D$. The cone is obtained by factoring $\epsilon_j$ through $d_j$. Write $D^*$ for a limit of $\bar{F}^*$; assuming $D, D^*$ exist the limits of $D, D^*, F$ exist together and are isomorphic.

In sum, limits in $C^J$ may be taken “pointwise” when possible. By virtue of this fact, $C^J$ “inherits” properties of $C$. If $C$ is complete then $C^J$ is, and if $C$ has finite limits then $C^J$ does; and dually. Also, the functor $\text{Lim}_J : C^J \to C$ preserves limits; and dually.

Morphisms in $C^J$ which are pointwise monic are monic; indeed, if $fh = gh$ then $f_jg_j = f_jh_j$ for all $j$, whence $g_j = h_j$ for all $j$, whence $g = h$. Similarly pointwise epis (isomorphisms) are epis (isomorphisms). If $C$ is exact then a monic $f$ in $C^J$ is pointwise monic; this follows since $\text{Ker}(f) = 0$. Similarly epis are pointwise epic.

Any functor from $C$ to $D$ yields a functor from $C^J$ to $D^J$, by composition on the right. It acts “pointwise”.

If $C$ is Abelian then $C^J$ is. It suffices to show that monics are normal and epis are conormal. Suppose the arrow $\eta$ of $C^J$ is monic; then $\text{Ker}(\eta) = 0$, so $\text{Ker}(\eta_j) = 0$ for all $j \in J$. If $\theta = \text{Coker}(\eta)$ then $\theta_j = \text{Coker}(\eta_j)$, all $j \in J$. By normality of $C$ $\text{Ker}(\theta_j) = \eta_j$ for all $J$, whence $\text{Ker}(\theta) = \eta$ and $\eta$ is normal.

The proof that epis are conormal is dual. If $C$ is Abelian, it is easily seen that monics in $C^J$ are pointwise monic, and similarly for epis and isomorphisms.
Being Abelian, \( C^J \) has an addition defined on its Hom sets. This may be described explicitly as follows. If \( \alpha \) and \( \beta \) are natural transformations from \( F \) to \( G \), \( (\alpha + \beta)_j = \alpha_j + \beta_j \), where the addition on the right is that of \( \text{Hom}_A(F(j), G(j)) \).

It is also true that if \( C \) is exact then \( C^J \) is exact (exercise 1).

There is an evaluation functor \( \varepsilon \) from \( C^J \times J \) to \( C \). For \( F : J \rightarrow C \) and \( j \in J \), \( \varepsilon(F, j) = F(j) \). For a natural transformation \( \eta \) from \( F \) to \( G \), and \( \beta : j \mapsto j' \) in \( J \), \( \varepsilon(\eta, \beta) = \eta_j F(\beta) = G(\beta) \eta_j \). This may be verified directly to be a functor. For each \( j \in J \) there is a “pointwise” evaluation functor \( \varepsilon_j \) from \( C^J \) to \( C \), which maps \( F \) to \( F(j) \) and \( \eta \) to \( \eta_j \).

Exercise 2 gives another property of limits and colimits in the setting of theorem 1.

2. Diagram lemmas. It is readily verified that in a category with finite products and equalizers, for maps \( f_i : b_i \rightarrow c \), \( i = 1, 2 \), the equalizer of \( f_1 p_1 \) and \( f_2 p_2 \) is the pullback of \( f_1 \) and \( f_2 \), where \( p_1, p_2 \) are the projections from \( b_1 \times b_2 \). As noted in chapter 13, in an Abelian category the equalizer of \( f_1 \) and \( f_2 \) is the kernel of \( f_1 - f_2 \). For the rest of the section suppose the category is Abelian. If \( f : a \rightarrow b \), write \( f^K : a^K \rightarrow a \) for the kernel, and \( f^C : b \rightarrow b^C \) for the cokernel.

Lemma 2. If \( g_i : a \rightarrow b_i \) is a pullback for \( f_i : b_i \rightarrow c \) and \( f_1 \) is epic, then \( g_2 \) is epic and \( f_1^K = g_1 g_2^K \). Dually if \( g_i : b_i \rightarrow a \) is a pushout for \( f_i : c \rightarrow b_i \) and \( f_1 \) is monic, then \( g_2 \) is monic, and \( f_1^K = g_1^C g_2^C \).

Proof: Let \( p_i \) and \( m_i \) be the projections and injections for \( b_1 \times b_2 \). Let \( \phi = f_1 p_1 - f_2 p_2 \), and \( k = \text{Ker}(\phi) \), so that \( g_i \) may be taken as \( p_i k \). Now, \( \phi m_1 = f_1 \), so \( \phi \) is epic since \( f_1 \) is. It follows that \( \phi = \text{Coker}(k) \). Suppose \( h p_2 k = 0 \); then \( h p_2 = h' \phi \) for some \( h' \), and \( 0 = h p_2 m_1 = h' \phi m_1 = h' f_1 \). Since \( f_1 \) is epic \( h' = 0 \), so \( h p_2 = 0 \), so \( h = 0 \). For the second claim, \( f_1 g_1 g_2^K = f_1 g_2 g_2^K = 0 \). If \( f_1 x = 0 \) then \( \phi m_1 x = 0 \), so \( m_1 h = ky \) for some \( y \). Also \( g_2 y = p_2 k y = p_2 m_1 x = 0 \), so \( y = g_2^K z \) for some \( z \); thus, \( x = p_1 m_1 x = p_1 k y = g_1 g_2^K z \). Conversely if \( \phi m_1 = 0 \) then \( m_1 h = ky \) for some \( k \) and \( y = g_2^K z \) for some \( z \); thus, \( x = p_1 m_1 x = p_1 k y = g_1 g_2^K z \).

Lemma 3. \( \text{Im}(f) \equiv \text{Ker}(g) \) iff \( g f = 0 \), and if \( g x = 0 \) then \( x \varepsilon = f t \) for some \( t \) and epic \( \varepsilon \). Dually \( \text{Coin}(f) \equiv \text{Coker}(g) \) iff \( f g = 0 \), and if \( x g = 0 \) then \( x \mu = t f \) for some \( t \) and monic \( \mu \).

Proof: Write \( f \) as a coimage-image pair \( m e \). If \( \text{Ker}(g) = m \) then \( g m = 0 \), so \( g f = 0 \); and if \( g x = 0 \) then \( x = m \theta \) for some \( \theta \), so by pullback \( x e = m \theta e = m e t = f t \) for some \( t \) and epic \( \varepsilon \). Conversely if \( g m e = g f = 0 \) then \( g m = 0 \); and letting \( x = \text{Ker}(g) \), \( x \varepsilon = m e t \) for some \( t \) and epic \( \varepsilon \). The left side is a coimage-image pair for the right side, so \( x = m \theta \) for some \( \theta \). It follows that \( m \equiv \text{Ker}(g) \).

![Figure 1](image-url)
Figure 1 is an important diagram in homology. Supposing that the middle two rows and the columns are exact, the remaining arrows are determined, and the remaining rows are exact. Also, if \( m \) is monic then \( m_t \) is; and if \( e' \) is epic then \( e_b \) is. Indeed \( gm f^K = m' f f^K = 0 \), so \( m_i \) is uniquely determined; the other three arrow are similarly determined. Further \( h^K e t m_l = e m f^K = 0 \), so \( e_t m_l = 0 \). Suppose \( e_t x = 0 \); then \( e g^K x = h^K e_t x = 0 \), so \( \exists w, e m w = g^K x e \). Then \( m' f w = g m w = g g^K x e = 0 \), so \( f w = 0 \), so \( \exists u (w = f^K u) \); and \( g^K x e = m w = m f^K u = g^K m_l u \), so \( x e = m t u \). This shows \( \text{Im}(m_l) \equiv \text{Ker}(e_t) \); if \( m \) is monic we may take \( e = t \), showing \( m_l t \equiv \text{Ker}(e_t) \). The claims for the last row follow dually.

Suppose that two diagrams \( D_1 \) and \( D_2 \) as in figure 1 are given, with arrows \( \alpha, \beta \) from the objects \( x \) of the middle two rows of \( D_1 \) to the corresponding objects in \( D_2 \), such that the combined diagram is commutative in the middle two rows. The same argument used to show uniqueness of \( m_t \), etc., shows that arrows may be added uniquely from the remaining objects, such that commutativity holds for “vertical” squares. It follows also for horizontal squares. For example, \( m_i 2 \alpha_\alpha = \alpha_b x m_i \); because the equation left multiplied by with \( g^2 \) holds.

![Diagram](image)

**Figure 2**

**THEOREM 4 (SNAKE LEMMA).** Given a commutative diagram as in figure 1, there is an arrow \( \delta : e^K \mapsto a^C \) such that \( \text{Im}(e_t) \equiv \text{Ker}(\delta) \) and \( \text{Coim}(m_b) \equiv \text{Coker}(\delta) \).

**PROOF:** From the diagram of figure 1, construct the diagram of figure 2 as follows. Let \( b_0 \) be the pullback object; by lemma 2, \( e_0 \) is epic and \( g_0 e^K_0 = e^K \). Since \( e^K \equiv \text{Im}(m) \), let \( e'_0 = e^K_0 \text{Coim}(m) \). Dually, \( b_1 \) is the pushout object, \( m_1 \) is monic, and \( m'_1 = \text{Im}(e') m^C_1 \). Let \( \delta_0 = g_1 g g_0 \). Then \( \delta_0 e'_0 = m_1 f^C e f = 0 \), so \( \delta_0 e^K_0 = 0 \), so \( \delta_0 = \delta_1 e_0 \) for some \( \delta_1 \). Further \( m'_1 \delta_1 e_0 = m'_1 \delta_0 = h h^K e_0 = 0 \), so \( m'_1 \delta_1 = 0 \), so \( m^K_1 \delta_1 = 0 \), so \( \delta_1 = m_1 \delta \) for some \( \delta \). Now, since \( e g^K = h^K e_t \) there is an \( \alpha \) with \( e_t = e_0 \alpha \) and \( g^K = g_0 \alpha \), and

\[
m_1 \delta e_t = m_1 \delta e_0 \alpha = \delta \alpha = g_1 g g_0 \alpha = g_1 g g^K = 0,
\]

so \( \delta e_t = 0 \). If \( \delta x = 0 \), by pullback \( \exists t_1, e_1 (e_0 t = x e_1) \) where \( e_1 \) is epic; then

\[
g_1 g g_0 t_1 = m_1 \delta e_0 t_1 = m_1 \delta x e_1 = 0.
\]

Then \( e' g g_0 t_1 = h e g_0 t_1 = h h^K e_0 t_1 = 0 \), so by lemma 3 \( \exists t_2, e_2 (g g_0 t_1 e_2 = m' t_2) \). Then

\[
m_1 f^C t_2 = g_1 m' t_2 = g_1 g g_0 t_1 e_2 = 0,
\]

216
so \( f^C t_2 = 0 \), so \( \exists t_3(t_2 = ft_3) \), and

\[
gmt_3 = m' ft_3 = m' t_2 = g g_0 t_1 \varepsilon_2.
\]

Thus \( g(m t_3 - g_0 t_1 \varepsilon_2) = 0 \), whence \( \exists t_4(gK t_4 = g_0 t_1 \varepsilon_2 - m t_3) \). Thus

\[
h^K x \varepsilon_1 \varepsilon_2 = h^K e_0 t_1 \varepsilon_2 = e g_0 t_1 \varepsilon_2 = e(gK t_4 + m t_3) = e gK t_4 = h^K e_4 t_4,
\]

whence \( x \varepsilon_1 \varepsilon_2 = e_4 t_2 \). By lemma 3 we have shown \( \text{Im}(e_4) \equiv \text{Ker}(\delta) \); the proof that \( \text{Coim}(m_0) \equiv \text{Coker}(\delta) \) is dual.

This proof becomes more transparent if it is compared to the proof in Ab along similar lines. In this case,

- \( b_0 = \{ \langle x, u \rangle \in C \times b : h^K(x) = e(u) \}, e_0(\langle x, u \rangle) = x, g_0(\langle x, u \rangle) = u. \)
- \( b_1 = b' \times a'' C/S \) where \( S = \{ \langle m'(v), -f^C(v) \rangle : v \in a' \}, m_1(y) = \langle 0, y \rangle + S, g_1(w) = \langle w, 0 \rangle + S. \)
- \( \text{Ker}(e_0) = \{ \langle 0, u \rangle : e(u) = 0 \} = \{ \langle 0, m(t) : t \in a' \} \subseteq \text{Ker}(\delta_0) \) since \( \langle gm(t), 0 \rangle = \langle m' f(t), -f^C f(t) \rangle \in S. \)
- If \( h^K(x) = e(u) \) then \( e' g(u) = h e(u) = hh^K(x) = 0, \) so \( g(u) \in \text{Ker}(e') = \text{Im}(m') \), so

\[
\langle g(u), 0 \rangle + S = \langle m'(v), 0 \rangle + S = g_1(m'(v)) = m_1(f^C(v))
\]

for some \( v. \)

Thus \( \delta_0 \) factors as \( m_1 \delta e_0 \) where \( \delta(x) = y \) iff \( \exists u, v(h^K(x) = e(u) \land m'(v) = g(u) \land f^C(v) = y). \)

We then have

\[
\delta(x) = 0 \iff \exists u, v(h^K(x) = e(u) \land m'(v) = g(u) \land f^C(v) = 0)
\]

\[
\iff \exists u, t_2(h^K(x) = e(u) \land gm(t_2) = m' f(t_2) = g(u))
\]

\[
\iff \exists u, t_2, t_3(h^K(x) = e(u) \land gK(t_3) = u - m(t_2))
\]

\[
\iff \exists t_2, t_3(h^K(x) = e(gK(t_3) + m(t_2)) = e gK(t_3)) = h^K e_4(t_3)
\]

\[
\iff \exists t_3(x = e_4(t_3)).
\]

The arrow \( \delta \) of the snake lemma is called the connecting morphism, for reasons which will become apparent in section 3. It is functorial, in the sense that if two diagrams \( D_1 \) and \( D_2 \) as in figure 1 are given, with arrows \( \alpha_x \) from the objects \( x \) \( D_1 \) to \( D_2 \), such that the combined diagram is commutative, then the diagram remains commutative if both connecting morphisms are added. Indeed, this is certainly true when (in the notation of the proof) \( \delta_1 \) is added, because \( \delta_0 = \delta_1 \varepsilon_0 \) where \( \varepsilon_0 \) is epic. The claim for \( \delta \) follows dually from this.

Other well known “diagram lemmas” will be stated below, and proved in the exercises for Abelian categories. The proofs are similar to the foregoing, using lemma 2 to adapt a suitable proof in Ab. More general methods use “embedding theorems”; further comments will be made following theorem 22 below.

Stating facts of homological algebra in the generality of Abelian categories is useful, since there are important Abelian categories in more advanced areas, which are not \( R \)-modules. Also, a typical benefit of using category theory is derived, that the arguments are given in their most basic and general form. Specific Abelian categories may satisfy additional restrictions, which may be imposed on an Abelian category to prove facts in general; examples may be found in [Mitchell]. In this text we are interested only in providing an introduction to the subject. Further facts will be stated for Abelian categories, but if this becomes excessively complicated they will be given for \( \text{Mod}_R \).

A commutative diagram as in the following figure where the rows are exact is called a morphism of short exact sequences. Since \( e \) is the cokernel of \( m \), if \( f \) and \( g \) are given then there is a unique \( h \) making the
diagram commutative; and dually if \(g\) and \(h\) are given there is a unique \(f\). The short exact sequences in a given Abelian category, with these morphisms, form a category. Some properties of this category are given in [Mitchell], exercise I.18.

\[
\begin{array}{ccccccccc}
0 & \rightarrow & a & \overset{m}{\rightarrow} & b & \overset{c}{\rightarrow} & c & \rightarrow & 0 \\
\downarrow f & & \downarrow g & & \downarrow h & & & & \\
0 & \rightarrow & a' & \overset{m'}{\rightarrow} & b' & \overset{e'}{\rightarrow} & c' & \rightarrow & 0
\end{array}
\]

Figure 3

**Theorem 5 (Short five lemma).** In a morphism of short exact sequences as in figure 3, if \(f, h\) are monic then \(g\) is; and dually.

**Proof:** Exercise 3.

**Theorem 6 (Five lemma).** Suppose in the diagram in an Abelian category

\[
\begin{array}{cccccccc}
a_1 & \overset{a_1}{\rightarrow} & a_2 & \overset{a_2}{\rightarrow} & a_3 & \overset{a_3}{\rightarrow} & a_4 & \overset{a_4}{\rightarrow} & a_5 \\
\downarrow f_1 & & \downarrow f_2 & & \downarrow f_3 & & \downarrow f_4 & & \downarrow f_5 \\
b_1 & \overset{b_1}{\rightarrow} & b_2 & \overset{b_2}{\rightarrow} & b_3 & \overset{b_3}{\rightarrow} & b_4 & \overset{b_4}{\rightarrow} & b_5
\end{array}
\]

the top row is exact at \(a_3\), the bottom row is exact at \(b_3\), \(f_1\) is epic, and \(f_2\) and \(f_4\) are monic; then \(f_3\) is monic. Dually if the top row is exact at \(a_4\), the bottom row is exact at \(b_3\), \(f_5\) is monic, and \(f_2\) and \(f_4\) are epic, then \(f_3\) is epic. If both sets of hypotheses hold then \(f_3\) is an isomorphism.

**Proof:** Exercise 4.

**Theorem 7 (Nine lemma).** Given a diagram in an Abelian category

\[
\begin{array}{cccccccc}
0 & \rightarrow & 0 & \rightarrow & 0 & \rightarrow & 0 \\
\downarrow a_0 & & \downarrow b_0 & & \downarrow c_0 & & \downarrow c_0 \\
0 & \rightarrow & a & \overset{m}{\rightarrow} & b & \overset{e}{\rightarrow} & c & \rightarrow & 0 \\
\downarrow f_1 & & \downarrow g_0 & & \downarrow h_0 & & \downarrow h_0 \\
0 & \rightarrow & a_1 & \overset{m_1}{\rightarrow} & b_1 & \overset{e_1}{\rightarrow} & c_1 & \rightarrow & 0 \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
0 & \rightarrow & 0 & \rightarrow & 0 & \rightarrow & 0
\end{array}
\]

with exact columns and bottom two rows, there exist unique \(m_0\) and \(e_0\) making the diagram commutative; further the top row extends to a short exact sequence. The claim follows for the third row, given exact sequences in the top two rows.

**Proof:** Exercise 5.

**Theorem 8 (Middle nine lemma).** Suppose that, in the diagram of theorem 7 short exact sequences are given in the first and third rows, and arrows \(m, e\) in the middle row. Then the middle row extends to a short exact sequence.

**Proof:** Exercise 6.
An important consequence of the nine lemma is the first Noether isomorphism, which states that given monics \( C \hookrightarrow B \) and \( B \twoheadrightarrow A \), there is a morphism of short exact sequences

\[
\begin{array}{c}
0 \rightarrow B \rightarrow A \rightarrow A/B \rightarrow 0 \\
0 \rightarrow B/C \rightarrow A/C \rightarrow A/B \rightarrow 0.
\end{array}
\]

This follows by taking \( C, C, 0 \) in the first row of the nine lemma.

3. Chain complexes. For an Abelian category \( A \), the chain complexes in \( A \) are a full subcategory of the functors from \( \mathbb{Z}^{\text{op}} \) to \( A \) where \( \mathbb{Z} \) is considered as the linear order. Writing \( C_n \) for the image of \( n \), and \( \partial_n \) for the image of the arrow \( n \mapsto n-1 \), a chain complex must satisfy \( \partial_{n-1} \partial_n = 0 \). We will use \( A\text{-Ch} \) to denote this category; arrows are called chain maps.

In many contexts \( \partial \) may be written rather than \( \partial_n \). Indeed, for a chain \( C \), let \( C^R \) be the chain where \( C^R_n = C_{n-1} \). The morphisms \( \partial_n : C_n \rightarrow C_{n-1} \) comprise the pointwise components of a morphism from \( C \) to \( C^R \). Also, in \( \text{Mod}_R \) (or any cocomplete Abelian category) \( \partial \) may be considered as an operation on the graded \( R \)-module \( \oplus C_n \). The requirement for a chain complex may be written as \( \partial \partial = 0 \), or \( \partial^2 = 0 \). An \( R \)-module with a map \( \partial \) where \( \partial^2 = 0 \) is called a differential module, and \( \partial \) a differential map. The reader should not confuse this use of the term with other uses; \( \partial \) is also called a boundary map.

If \( C \) is a limit in \( A^{\text{op}} \) of chain complexes, then \( C \) is a chain complex. This may be seen from the proof of theorem 1. The 0 arrow from \( C_n \) to \( C_{n-2} \) satisfies the required diagram property, so equals \( \partial^2 \) in \( C \) by uniqueness. Dually a colimit of chain complexes is a chain complex. By arguments as above, \( A\text{-Ch} \) is an Abelian category.

As noted in section 1, a functor \( F : A \rightarrow B \) yields a functor from \( A^{\text{op}} \) to \( B^{\text{op}} \). This restricts to a functor from \( A\text{-Ch} \) to \( B\text{-Ch} \).

The object of the kernel of the chain map \( \partial \) from \( C \) to \( C^R \) is called the cycles of \( C \), and denoted \( Z \). \( Z \) is used indifferently to denote the kernel chain, or the functor \( C \mapsto Z \) from \( A\text{-Ch} \) to \( A\text{-Ch} \). Likewise, \( Z_n \) denotes the object at position \( n \), or the functor from \( A\text{-Ch} \) to \( A \). Since the kernel is taken pointwise, the kernel map \( Z_n \mapsto C_n \) is monic. In a concrete category, the kernel map may be taken as inclusion, and the map \( Z_n \mapsto Z_{n-1} \) is seen to be the restriction of \( \partial \).

Similarly let \( C^L \) be the chain where \( C^L_n = C_{n+1} \); then \( \partial \) is a morphism from \( C^L \) to \( C \). The object of the image of this chain map is called the boundaries of \( C \), and denoted \( B \). \( B_n \) denotes the object in position \( n \), etc. In a concrete category, the image map may be taken as inclusion, and the map \( B_n \mapsto B_{n-1} \) is seen to be the restriction of \( \partial \).

In an Abelian category, if \( fg = 0 \) then a monic from the image object of \( g \) to the kernel object of \( f \) is determined. Indeed, let \( me \) be a coinage-image factorization of \( g \), and \( j \) a kernel arrow for \( g \). Then \( fme = 0 \), so \( fm = 0 \), so \( m \leq c m \), so \( m = kj \) where \( j \) is unique and monic (see section 13.9).

In particular, from \( \partial_n \partial_{n+1} = 0 \) it follows that a monic from \( B_n \) to \( Z_n \) is determined. In the case of a concrete category, this may be taken is inclusion. There is thus a short exact sequence

\[
0 \rightarrow B \rightarrow Z \rightarrow Z/B \rightarrow 0.
\]

The chain \( Z/B \) is denoted \( H \), and called the homology chain. \( H_n \) denotes the object in position \( n \), etc.; note that \( H_n = Z_n/B_n \).

The requirement \( B_n \subseteq Z_n \) is weaker than the exactness condition \( B_n = Z_n \); it is called subexactness. \( H_n \) measures the difference from exactness of a chain complex at \( C_n \). In particular, the complex is exact at \( C_n \) iff \( H_n = 0 \). A chain complex whose homology is 0 is called acyclic.
Various facts of interest in applications of homology can be proved in the general setting established so far. Let \( m_n, e_n \) be a coinage-image factorization of \( \partial_n \); and let \( k_n = \text{Ker}(\partial_n) \). We have already observed that there is a unique \( j_n \) with \( m_{n+1} = k_n j_n \). Since \( e_n = \text{Coker}(k_n) \), \( B_{n-1} \) is isomorphic to \( C_n / Z_n \). \( H \) may be seen to be a functor on \( A \)-Ch, since there is a functor whose value at \( C \) is the functor in \( C' \) for the appropriate functor category \( J \), whose image is \( B \rightarrow Z \); \( H \) is the cokernel object of the latter. If \( f : C \rightarrow C' \), \( H(f) \) is the unique arrow from \( H(C) \) to \( H(C') \) giving a morphism of short exact sequences with other arrows \( B(f) \) and \( Z(f) \).

**Theorem 9.** Given a chain complex \( C \), for each \( n \) there is an exact sequence

\[
0 \rightarrow H_n \rightarrow C_n / B_n \rightarrow Z_{n-1} \rightarrow H_{n-1} \rightarrow 0.
\]

**Proof:** Using the notation above, \( m_{n+1} e_n = 0 \). Thus, the canonical epimorphism \( C_n \twoheadrightarrow C_n / B_n \) factors through \( e_n \), yielding an arrow \( C_n / B_n \rightarrow B_{n-1} \), whence an arrow \( C_n / B_n \rightarrow C_n / Z_n \). By the first Noether isomorphism theorem, the kernel of \( C_n / B_n \rightarrow C_n / Z_n \) is \( Z_n / B_n = H_n \). The kernel of the composed arrow \( C_n / B_n \rightarrow B_{n-1} \rightarrow Z_{n-1} \) is \( H_n \) also, since \( B_{n-1} \rightarrow Z_{n-1} \) is monic. \( C_n / B_n \rightarrow B_{n-1} \) is epic since \( e_n \) is, so the cokernel of \( C_n / B_n \rightarrow Z_{n-1} \) is \( Z_{n-1} \rightarrow H_{n-1} \).

Suppose

\[
0 \rightarrow C' \rightarrow C \rightarrow C'' \rightarrow 0
\]

is an exact sequence of chain complexes, so that \( C'' = C / C' \). \( C' \) preserves colimits, in particular is right exact; and \( \text{Ker} \) is left exact. Thus, the diagram of figure 1 arises, with vertical rows the sequences of theorem 9. By the snake lemma, for each \( n \) there is a connecting morphism \( \delta_n : H''_n \rightarrow H'_{n-1} \). There arises an exact sequence

\[
\cdots \rightarrow H'_n \rightarrow H_n \rightarrow H''_n \delta_n \rightarrow H'_{n-1} \rightarrow H_{n-1} \rightarrow H''_{n-1} \rightarrow \cdots,
\]

called the long exact sequence. Further, it arises via a functor from the exact sequences in \( A \)-Ch to \( A \)-Ch.

\( H'' = H(C'') = H(C / C') \) is called the relative homology. In the case \( A = \text{Mod}_R \) for a ring \( R \), the relative homology module \( H''_n = Z''_n / B''_n \) can be viewed as a quotient of submodules of \( C_n \), using theorem 4.11. Letting \( c_n \) denote an element of \( C_n \), \( c_n + C'_n \) is in \( \text{Ker}(\partial'_n) \) iff \( \partial_n(c_n) \in C'_{n-1} \) iff \( c_n \in \partial'_n^{-1}[C'_{n-1}] \); let \( Z''_n \) denote this submodule. Also, \( c_n + C'_n \) is in \( \text{Im}(\partial'_n) \) iff there exists \( c_{n+1} \) with \( \partial_{n+1}(c_{n+1}) = c_n + C'_n \), iff \( c_n \in B_n + C'_n \); let \( B''_n \) denote this submodule. Then \( H''_n \) is isomorphic to \( Z''_n / B''_n \).

In applications the chains are obtained by a functor \( C \) from some category \( T \) to \( A \)-Ch. For an object \( X \) of \( T, H(C(X)) \) is denoted \( H(X) \) (the homology objects are obtained by composing with the homology functor). An example is given in the next section. The chain functor usually preserves monics. In this case, if \( i : Y \rightarrow X \) in \( T \) is monic then \( C(i) : C(Y) \rightarrow C(X) \) is monic, giving rise to a short exact sequence

\[
0 \rightarrow C(Y) \rightarrow C(X) \rightarrow C(X) / C(Y) \rightarrow 0.
\]

The relative homology object \( H(C(X) / C(Y)) \) is denoted \( H(X,Y) \). \( H(X) \) is usually \( H(X,I) \) where \( I \) is an initial object in \( T \) (for example the empty set), and is called the absolute homology. An obvious question is whether there is a construction in \( T \) to which \( C(X) / C(Y) \) corresponds. The answer is “sometimes”; obtaining the exact sequence of chains “algebraically” avoids considering the question when this is not necessary.

For another example of a general fact, if the objects of a chain \( C \) are in \( \text{Mod}_R \), and \( X \) is an \( R \)-module, then the chain complex \( X \otimes C \) can be considered. In a common case \( R = \mathbb{Z} \) and \( X \) is a ring, changing the “coefficient ring” of the chains. Viewing the situation thus allows making use of facts concerning changing rings. Also, as mentioned earlier the tensor product can be considered in more general categories.
Two chain maps $f_1, f_2 : C \to C'$ are said to be chain homotopic if there is a system of arrows $h_n : C_n \to C'_{n+1}$ such that $f_1 n - f_2 n = \partial'_{n+1} h_n + h_{n-1} \partial_n$ for all $n$. The origin of the term will be seen in lemma 24.40; chain homotopy is a homological brand of homotopy, and is an important tool in homology. As usual, we let $h$ denote the system of $h_n$; note that it is not in general a chain map.

**Theorem 10.**

a. The relation of being chain homotopic is an equivalence relation on the chain maps.
b. Composition respects the equivalence relation.
c. For chain homotopic chain maps $f_1$ and $f_2$, $H(f_1) = H(f_2)$ for all $n$.

**Proof:** Let $\sim$ denote the chain homotopy relation. Letting $h = 0$ shows that $f \sim f$; if $h$ shows that $f_1 \sim f_2$ then $-h$ (i.e., the system of negations) shows that $f_2 \sim f_1$; and if $h$ shows that $f_1 \sim f_2$ and $h'$ shows that $f_2 \sim f_3$ then $h + h'$ shows that $f_1 \sim f_3$. This proves part a. Suppose that for $i = 1, 2$, $f'_i : C \to C'$ and $f''_i : C' \to C''$. Suppose $h$ shows that $f_1 \sim f_2$, and $h'$ shows that $f'_1 \sim f'_2$. Let $\bar{h} = f'_1 h + h' f_2$, that is, $\bar{h}_n = f'_{1,n+1} h_n + h'_{n,2n}$; note that $\bar{h}_n : C_n \to C''_{n+1}$. Using $f'_1 f_1 - f'_2 f_2 = f'_1(f_1 - f_2) + (f'_1 - f'_2) f_2$, and the fact that $f_1, f'_i$ are chain maps, one verifies that $\bar{h}$ shows that $f'_2 f_2 \sim f'_1 f_1$. For part c, let $\eta_n : Z_n \to H_n$ be the canonical epimorphism, let $k_n : Z_n \to C_n$ be the inclusion, and similarly for $\eta'_n, k'_n$. For an arrow $s : C_n \to C'_n$, $H_n(s) \eta_n = \eta'_n Z_n(s)$, so to show that $H_n(s) = 0$ it suffices to show that $\eta'_n Z_n(s) = 0$. If $s = h_{n-1} \partial_n$ then $k'_n Z_n(s) = sk_n = h_{n-1} \partial_n k_n = 0$, so $Z_n(s) = 0$. If $s = \partial'_{n+1} h_n$, let $m'_{n+1} e'_{n+1}$ be a coinage-image factorization of $\partial'_{n+1}$. Let $j'_n$ be the induced arrow from $B'_n$ to $Z'_n$. Then $k'_n Z_n(s) = sk_n = \partial'_{n+1} h_n k_n = m'_{n+1} e'_{n+1} h_n k_n = k'_n j'_n e'_{n+1} h_n k_n$. Thus, $Z_n(s) = j'_n e'_{n+1} h_n k_n$, and $\eta'_n Z_n(s) = 0$.

The proof of part c in Mod$_R$ is as follows. To show that $H_n(s) = 0$ it suffices to show that if $c_n \in Z_n$ then $s(c_n) \in B'_n$. When $s = \partial'_{n+1} h_n + h_{n-1} \partial_n$, $h_{n-1} \partial_n(c_n) = 0$, and $\partial'_{n+1} h_n(c_n) \in B'_n$.

A chain complex is said to be chain contractible if the identity arrow is chain homotopic to the zero arrow. It follows from the theorem that if this is so, then the chain complex is acyclic (because $H(\epsilon) = \epsilon$ and $H(0) = 0$). Chain contractibility is a stronger requirement in general, although for free chain complexes the two notions are equivalent; see [Spanier], section 4.2.

In any Abelian category, it follows from the long exact sequence that $H_n$ is half exact, and so preserves biproducts. In Mod$_R$ much more is true, a fact of interest in algebraic topology. We first note that in a category of two-sided modules (two-sided modules will be used in section 8), one verifies by standard componentwise arguments the following.

- Homomorphisms $f_i : M_i \to N_i$ are monic (epic) for all $i$ iff $\times_i f_i : \times_i M_i \to \times_i N_i$ is monic (epic); and similarly for the coproduct.
- If homomorphisms $f_i : M_i \to N_i$ are monic for all $i$, an isomorphism $\times_i (M_i/ N_i) \to (\times_i M_i)/(\times_i N_i)$ is induced; and similarly for the coproduct.
- Given homomorphisms $f_i : M_i \to N_i$, let $f$ denote $\times_i f_i$ (so that $f(\langle x_i \rangle) = \langle f_i(x_i) \rangle$). The kernel (object) of $f$ is the product of the kernels of the $f_i$; and similarly for coproducts.
- With $f$ as above, the image (object) of $f$ is the product of the images of the $f_i$; and similarly for coproducts.

In particular, $Z_n, B_n$, and $H_n$ preserve products and coproducts, whence $Z, B$, and $H$ do. It is worth noting that in any Abelian category, the product functor on any index set preserves monics, and the coproduct epics, so the special fact about a category of modules is that the product functor preserves epics and the coproduct functor preserves monics; in particular they are exact. An exact functor between Abelian categories $F : A \to B$ preserves homology on the chain complexes, in that $F(Z_n)/F(B_n)$ is isomorphic to $F(H_n)$. Thus, when $A$ is a category of two-sided modules, the product functor from $A^I$ to $A$ preserves homology; and also the coproduct functor.
In a category of two-sided modules, for any directed index set \( J \), the direct limit functor on \( J \) preserves monics, hence is exact, hence preserves homology on the chain complexes. To prove this, suppose \( f \) is a morphism from the direct system with objects \( A_i \) to that with objects \( A'_i \). Let \( a_i \) denote an element of \( A_i, \phi_{ij} \) an arrow of the direct system, and similarly for the primed system. The direct limit of the \( A_i \) is \( \oplus_i A_i / \equiv \), where \( \equiv \) is generated by the equivalences \( a_i \equiv a_j \) when for some \( k, \phi_{ik}(a_i) = \phi_{jk}(a_j) \). Suppose each component \( f_i \) of \( f \) is injective. Suppose \( \phi_{ik}(f_i(a_i)) = \phi_{jk}(f_j(a_j)) \); then \( f_k(\phi_{ik}(a_i)) = f_k(\phi_{jk}(a_j)) \), whence \( \phi_{ik}(a_i) = \phi_{jk}(a_j) \).

\[ H_0 \] does not preserve inverse limits, so inverse limit is not exact; see [Spanier], section 4.1, for an example.

Complexes where the boundary operator goes up rather than down are also considered. These are functors from \( Z \) to \( A \), where \( \partial_0 \partial_{n-1} = 0 \). Formally there is little distinction, indeed they are clearly the same as the chain complexes, composed with the “reversal” functor from \( Z \) to \( Z^{\text{op}} \). In applications, however, the direction of the differential map (to lower or higher degree) arises from the nature of the map.

Complexes where the boundary map goes up are called cochain complexes. The object at position \( n \) is often denoted \( C^n \), although other notations are used. Similarly, “\( \text{co} \)” is prepended to the terms cycle, boundary, and homology, and the symbols \( Z, B, \) and \( H \) superscripted; \( Z^n \) is the kernel of \( \partial_n \), \( B^n \) the image of \( \partial_{n-1} \), and \( H^n \) the quotient.

In many applications \( C_0 = 0 \) for \( n < 0 \); such complexes are called positive, although nonnegative might be better terminology. For such complexes, \( Z_0 = C_0 \). For cochain complexes where \( C^n = 0 \) for \( n < 0 \), \( B^0 = 0 \).

In \( \text{Ab} \) (in fact in Mod\(_R \) where \( R \) is a principal ideal domain), an important identity regarding the ranks of the chain groups can be proved. Using theorem 8.9, the value \( \rho(M) \) for a finitely generated \( R \)-module may be defined as the dimension of the free \( R \)-module \( M/\text{Tor}(M) \). Further, if \( N \subseteq M \) is a subgroup, then \( N \) and \( M/\text{Tor}(M) \) are finitely generated, and \( \rho(M) = \rho(N) + \rho(M/N) \) (exercise 8).

**Theorem 11.** Suppose \( C \) is a chain complex in \( \text{Ab} \) such that each \( C_n \) is finitely generated, and only finitely many are nonzero. Then \( \sum_n (-1)^n \rho(C_n) = \sum_n (-1)^n \rho(H_n) \).

**Proof:** We have \( \rho(Z_n) = \rho(B_n) + \rho(H_n) \), and since as observed above \( B_{n-1} \) is isomorphic to \( C_n/Z_n \), \( \rho(C_n) = \rho(Z_n) + \rho(B_{n-1}) = \rho(B_n) + \rho(H_n) + \rho(B_{n-1}) \). Multiplying by \( (-1)^n \) and summing yields the claim.

The value in the theorem is called the Euler characteristic of the chain complex. The numbers \( \rho(H_n) \) are called the Betti numbers.

**4. Abstract simplicial complexes.** In this section one of the most important chain functors will be given. It is defined “combinatorially”. In chapter 24 it will be seen to have fundamental applications in topology. It also has applications in combinatorics, and has become of computational interest.

An abstract simplicial complex is a collection \( K \) of nonempty finite subsets of a set \( V \) of “vertices”, with the following properties.

1. If \( S \in K \) and \( T \subseteq S \) is nonempty then \( T \in K \).
2. \( V = \cup K \).

The sets \( S \in K \) are called simplexes; a simplex of size \( k + 1 \) is called a \( k \)-simplex. If \( T \subseteq S \) \( T \) is called a subsimplex or face of \( S \). By an abuse of terminology a 1-simplex may be called a vertex.

A finite set \( S = \{v_0, \ldots, v_k\} \) has \((k+1)!\) orderings \( \langle v_{i_0}, \ldots, v_{i_k} \rangle \). Say that two orderings \( \langle v_{i_0}, \ldots, v_{i_k} \rangle \) and \( \langle v_{j_0}, \ldots, v_{j_k} \rangle \) are equivalent if the map \( \pi \) where \( \pi(i_s) = j_s \) is an even permutation. An oriented \( k \)-simplex is an equivalence class of orderings of a \( k \)-simplex; we use \( [v_0, \ldots, v_k] \) to denote such. In some contexts symbols \([v_0, \ldots, v_k] \) arise where the \( v_i \) are not all distinct. These may be considered to denote \( 0 \).

The group \( C_k(K) \) of \( k \)-chains is defined to be the Abelian group generated by the oriented \( k \)-simplexes, subject to the relations \([v_0, v_1, \ldots, v_k] = -[v_1, v_0, \ldots, v_k] \) for \( k \geq 2 \). For \( k \geq 0 \) \( C_k \) is a free group of dimension
equal to the number of \( k \)-simplexes, which may be 0; defining \( C_k \) with two generators per simplex is done to facilitate the definition of the boundary operator. \( C_k(K) \) is defined for \( k < 0 \), to be the 0-group.

The group homomorphism \( \partial_k : C_k \rightarrow C_{k-1} \) is defined by its action on the generators, namely,

\[
\partial_k([v_0, \ldots, v_k]) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k],
\]

where \([v_0, \ldots, \hat{v}_i, \ldots, v_k]\) obtained from \([v_0, \ldots, v_k]\) by deleting \( v_i \) (a standard notation).

**Theorem 12.** \( \partial_{k-1} \partial_k = 0 \).

**Proof:** This is trivial if \( k < 1 \). For \( k \geq 1 \),

\[
\partial \partial([v_0, \ldots, v_k]) = \sum_i (-1)^i \left( \sum_{j<i} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, v_k] + \sum_{j>i} (-1)^{j-1} [v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_k] \right).
\]

This is readily seen to equal 0.

Rather than the Abelian group \( C_k \), the \( R \)-module generated by the oriented \( k \)-simplexes can be considered, for any commutative ring \( R \). This is easily seen to equal \( R \otimes C_k \), so that this is an example of taking the tensor product of the chain functor with a fixed object, mentioned earlier. The case of \( R \) being the finite field of order 2 is commonly used in topology; in this case the oriented \( k \)-simplexes may be replaced by the sets of size \( k + 1 \).

A function \( f : K \rightarrow L \) between abstract simplicial complexes is said to be a simplicial map if it is induced by a function on the vertices, that is, if there is a function \( f_0 : V_K \rightarrow V_L \) such that \( f(S) = f_0[S] \) for \( S \in K \). It is readily verified that the abstract simplicial complexes together with the simplicial maps form a category, which we denote \( \text{ASC} \).

A simplicial map \( f : K \rightarrow L \) induces a map \( f' \) from the oriented \( k \)-simplexes of \( K \) to the \( k \)-chains of \( L \) by letting \( f'([v_0, \ldots, v_k]) = [f(v_0), \ldots, f(v_k)] \). By the convention given above, \( f'([v_0, \ldots, v_k]) = 0 \) if the \( f(v_i) \) are not distinct.

It is routine to check that a functor from \( \text{ASC} \) to \( \text{Ab-Ch} \) has been defined, the simplicial chain functor.

These notions have turned out to be useful in homological algebra. Projective modules were first defined in the late 1950’s. It will be seen below that free modules are projective; various properties of interest in fact hold for projective modules. Projectives can be defined in any Abelian category, and the definition dualizes. In this section we will consider some equivalent formulations; some sufficient conditions for an Abelian category to have “enough” projectives or injectives; and effects of restrictions on the ring \( R \) on projectives and injectives in \( \text{Mod}_R \).
In an Abelian category define \( b \) to be a summand of \( a \) of \( a \equiv b \oplus c \) for some \( c \). By lemma 13.23 this is so if \( m : b \rightarrow a \) is split monic, or \( p : a \rightarrow b \) is split epic. Note that in a short exact sequence \( 0 \rightarrow a \rightarrow b \rightarrow c \rightarrow 0 \), the monic is split iff the epic is split iff \( b \equiv a \oplus c \); in this case the short exact sequence is said to split, or to be split. In \( \text{Mod}_R \), \( \equiv \) is replaced by \( = \), a definition previously given.

**Theorem 13.**

a. \( a \) is projective iff \( \text{Hom}^a \) is exact; dually \( a \) is injective iff \( \text{Hom}_a \) is exact.

b. \( a \) is projective iff any epic to \( a \) is split; dually \( a \) is injective iff any monic from \( a \) is split.

c. If \( a \) is projective and \( m : b \rightarrow a \) is split monic then \( b \) is projective; dually if \( a \) is injective and \( p : a \rightarrow b \) is split epic then \( b \) is injective.

d. A summand of a projective is projective, and a summand of an injective is injective.

e. If \( a = \oplus_j a_j \) then \( a \) is projective iff every \( a_j \) is; dually if \( a = \times_j a_j \) then \( a \) is injective iff every \( a_j \) is.

**Proof:** For part a, \( \text{Hom}^a \) is left exact, and a functor between Abelian categories is right exact (preserves cokernels) iff it preserves epics. For part b, suppose \( a \) is projective, and \( f \) is an epimorphism to \( a \); then there is a \( g \) such that \( fg = t_a \). Conversely suppose every epic to \( a \) is split. Let \( f : c \rightarrow b \) be epic, and \( h : a \rightarrow b \). In the pullback \( f', h' \) of \( f, h \) \( f' \) is an epic to \( a \), so has a right inverse \( m \), and \( f(h'm) = h'fm' = h \). For part c, let \( p \) be a left inverse to \( m \), and suppose \( f : d \rightarrow c \) is epic and \( h : b \rightarrow c \). Then there is an \( h' \) with \( fh' = hp \), and \( f(h'm) = h \). Part d follows by part c. For part e, the injections into a coproduct in a category with 0 are split, so if the coproduct is projective then the \( a_j \) are by part c. Conversely given \( f : c \rightarrow b \) epic and \( h : a \rightarrow b \) there are arrows \( h'_j : a_j \rightarrow c \) with \( fh'_j = hm_j \), and so an arrow \( h' : a \rightarrow c \) with \( fh' = h \); then \( fh'_j = hm_j \) for all \( j \) and so \( fh' = h \).

An Abelian category \( A \) is said to have enough projectives if for every object \( a \) there is a projective \( b \) and an epic \( f : b \rightarrow a \). Suppose \( A \) is cocomplete, and say that an object \( r \in A \) is a generator if for every \( a \in A \) there is an index set \( J \) and an epic \( f : \oplus_{j \in J} r \rightarrow a \).

Dually an object \( s \) in a complete Abelian category \( A \) is called a cogenerator if for every \( a \in A \) there is an index set \( J \) and a monic \( f : a \rightarrow \times_{j \in J} s \). \( A \) is said to have enough injectives if for every object \( a \) there is an injective \( b \) and a monic \( f : a \rightarrow b \).

**Lemma 14.** In a cocomplete Abelian category \( A \), an object \( r \) is a generator iff whenever \( f : a \rightarrow b \) is nonzero there is a \( \sigma : r \rightarrow a \) such that \( f \sigma \) is nonzero. In particular, if \( a \) is a nonzero object there is a nonzero arrow from \( r \) to \( a \). If \( r \) is projective, this condition suffices for \( r \) to be a generator. The dual statements hold in a complete Abelian category.

**Proof:** If \( r \) is a generator, let \( h : \oplus_{j \in J} r \rightarrow a \) be epic; if \( f \) is nonzero then for some \( j \), \( f(hm_j) \neq 0 \). Conversely let \( J = \text{Hom}(r, a) \), and let \( h : \oplus_{j \in J} r \rightarrow a \) be the arrow where \( hm_j = j \). Then \( h \) is epic; if \( fh = 0 \) then \( fj = 0 \) for all \( j \in J \), so by hypothesis \( f = 0 \). The second claim follows by considering \( t_a \). For the third claim, let \( f_1f_2 \) be a coinage-image pair for \( f \), so that \( f_2 : a \rightarrow c \) is epic. By hypothesis there is a nonzero arrow \( \tau : r \rightarrow c \). If \( r \) is projective, there is an arrow \( \sigma : r \rightarrow a \) with \( f_2\sigma = \tau \). Then \( f\sigma = f_1f_2\sigma = f_1\tau \) is nonzero, since \( \tau \) is and \( f_1 \) is monic.

The requirement that \( \sigma \) exist for nonzero \( f_2 \) is simply the requirement that \( \text{Hom}(M, -) \) be faithful. For later use we note the following. Say that a functor \( F \) between Abelian categories “reflects exactness” if exactness of \( 0 \rightarrow F(L) \rightarrow F(M)M \rightarrow F(N) \rightarrow 0 \) implies exactness of \( 0 \rightarrow L \rightarrow M \rightarrow N \rightarrow 0 \); then if \( F \) is faithful it reflects exactness. The proof is left to exercise 10. See [Mitchell], section II.7, for further facts concerning reflection properties of functors.

One easily sees that if an Abelian category \( A \) has a projective generator then it has enough projectives; and dually if \( A \) has an injective cogenerator then it has enough injectives. For the remainder of the section we consider the category \( \text{Mod}_R \) for a ring \( R \).
Theorem 15. In $\text{Mod}_R$ the following hold.

a. A free module is projective.

b. Any module is the epimorphic image of a free module.

c. $R$ is a projective generator.

d. Any summand of a free module is projective.

e. Any projective module is a summand of a free module.

f. If $R$ is a principal ideal domain then a projective module is free.

g. If $R$ is a principal ideal domain then a projective module is free.

Proof: For part a, if the module $A$ has basis $X$, $f : B' \rightarrow B$ is an epimorphism, and $h : A \rightarrow B$ is linear, the restriction of $h$ to $X$ may be lifted to a map $h_1 : X \rightarrow B'$; namely $h_1(x)$ is any $y$ such that $f(y) = h(x)$. The unique linear map $h' : A \rightarrow B'$ extending $h_1$ satisfies $fh' = h$. For part b, if $A$ is a module let $F$ be the free module with basis $A$ (considered as a set); the homomorphism from $F$ to $A$ induced by the identity on $A$ is clearly surjective. Part c follows, because $R$ is free and by part b is a generator, Part d follows by part a and theorem 13.d. Part e follows by part b and theorem 13.b and the remark preceding theorem 13. Part f follows by part e and theorem 8.6.

In particular $\text{Mod}_R$ has enough projectives. Note that theorems 13.b and 15.a provide another proof of lemma 8.7. Not every projective module is free; $\mathbb{Z}_2$ as a $\mathbb{Z}_2 \times \mathbb{Z}_2$ module is projective, being a summand of the free module $\mathbb{Z}_2 \times \mathbb{Z}_2$; but it is not free. Another example will be given in section 20.11.

A relatively recent result, called the Quillen-Suslin theorem after the two mathematicians who settled it independently in 1976, states that if $R$ is the ring $F[x_1, \ldots, x_k]$ of $n$ variable polynomials over a field $F$, then a projective $R$-module is free. See [Lam] for an excellent treatment of this, and various other circumstances under which projective modules are free. More recently, as a result of advances in computing in polynomial rings, effective versions of the Quillen-Suslin theorem have been proved; see [LogSturm].

Lemma 16. If given a left ideal $I \subseteq R$ and a linear map $f : I \rightarrow A$, there is a linear map $f' : R \rightarrow A$ extending $f$, then $A$ is injective.

Proof: Suppose $C$ is a submodule of $B$, and $f : C \rightarrow A$. By a typical application of Zorn’s lemma (see for example theorem 15.4) it suffices to show that, for a submodule $D$ with $C \subseteq D$, and $x \in B - D$, a function $g : D \rightarrow A$ can be extended to the submodule generated by $D$ and $x$. Let $I = \{r \in R : rx \in D\}$; clearly $I$ is a left ideal. The linear map $f(r) = g(rx)$ has an extension $f' : R \rightarrow A$. The map $g'((y + rx) = g(y) + f'(r))$, $y \in D$, $r \in R$, is well defined; if $y + rx = 0$ then $r \in I$ so $f'(r) = g(rx) = -g(y)$. It is also readily verified to be linear.

An $R$-module $A$ is said to be divisible if whenever $r \in R$ is not a zero divisor and $a \in A$ there is a $b \in A$ with $rb = a$.

Lemma 17. If an $R$-module $A$ is injective it is divisible. The converse holds if $R$ has no zero divisors and every left ideal is principal.

Proof: Consider the linear map $f : Rr \rightarrow A$ given by $f(sr) = sa$; $f$ is well defined because $r$ is not a zero divisor. Since $A$ is injective $f$ has an extension to a linear map $f' : R \rightarrow A$; then $a = f'(r) = rf'(1)$. For the second claim, if $I = Rr$ is an ideal and $f : I \rightarrow A$ then there is a $b \in A$ with $rb = f(r)$, and the map $f' : R \rightarrow A$ given by $f'(s) = sb$ extends $f$. The claim thus follows by lemma 16.

Dedekind domains will be defined in the next chapter. An integral domain is a Dedekind domain iff every divisible module is injective; see [Rotman], theorem 4.19.

If $M$ is a divisible $R$-module and $N$ is a submodule then $M/N$ is divisible; indeed if $rb = a$ then $rb + N = a + N$. It is also readily verified that a direct sum of divisible modules is divisible.

225
Clearly, \( \mathbb{Q}/\mathbb{Z} \) is a divisible commutative group. Thus, \( \mathbb{Q}/\mathbb{Z} \) is a divisible commutative group, so by lemma 17 it is an injective commutative group. It is readily verified that there is a nonzero homomorphism from any cyclic group to \( \mathbb{Q}/\mathbb{Z} \). Since \( \mathbb{Q}/\mathbb{Z} \) is injective there is a nonzero homomorphism from any commutative group. Thus, by lemma 14 \( \mathbb{Q}/\mathbb{Z} \) is an injective cogenerator in \( \text{Ab} \).

One may prove directly, without the use of cogenerators, that \( \text{Ab} \) has enough injectives. Given \( M \in \text{Ab} \) write it as \( (\sum_{j \in J} \mathbb{Z})/K \) for some index set \( J \) and some subgroup \( K \) of the free group. This can be embedded in \( (\sum_{j \in J} \mathbb{Q})/K \). By remarks above, the latter is divisible.

Let \( G : \text{Mod}_R \hookrightarrow \text{Ab} \) be the forgetful functor, and suppose \( L \in \text{Mod}_R \) and \( N \in \text{Ab} \). The Abelian group \( \text{Hom}(G(R), N) \) may be considered a left \( R \)-module, with the action \((rf)(r') = f(r'r)\). There is a natural equivalence

\[
\text{Hom}(L, \text{Hom}(G(R), N)) \cong \text{Hom}(G(L), N)
\]

in \( \text{Ab} \), which maps \( a \mapsto \psi_a \) to \( a \mapsto \psi_a(1) \), as is readily verified. If \( N \) is injective in \( \text{Ab} \), the right side of (1) is exact, as a functor of \( L \). It follows that the left side is, whence \( \text{Hom}(G(R), N) \) is injective in \( \text{Mod}_R \). If \( C \) is an injective cogenerator then the right side is nonzero for nonzero \( L \), and the left side is, and so \( \text{Hom}(G(R), N) \) is an injective cogenerator. In particular, \( \text{Hom}(G(R), \mathbb{Q}/\mathbb{Z}) \) is an injective cogenerator for \( \text{Mod}_R \).

The preceding construction of an injective cogenerator is readily adapted to the right or two-sided \( R \)-modules, by considering the action \((fr)(r') = f(rr')\).

To see directly that \( \text{Mod}_R \) has enough injectives, note that if \( m : G(L) \hookrightarrow N \) is monic where \( N \in \text{Ab} \) is injective then \( L \) can be embedded in \( \text{Hom}(G(R), L) \), via \( a \mapsto (r \mapsto ra) \), and \( \text{Hom}(G(R), L) \) can be embedded in \( \text{Hom}(G(R), N) \), via \( g \mapsto mg \).

**Theorem 18.** The following are equivalent for a ring \( R \).

a. \( R \) is semisimple.

b. Every \( R \)-module is projective.

c. Every \( R \)-module is injective.

d. Every left ideal of \( R \) is injective.

**Proof:** For a \( \Rightarrow \) b, if \( M \) is an \( R \)-module then \( M = F/K \) where \( F \) is free. Since \( R \) is semisimple \( M \) is, and so \( K \) is a summand of \( F \), whence \( M \) is a summand of \( F \) and \( M \) is projective. For b \( \Rightarrow \) c, if \( M \) is an \( R \)-module then \( M \subseteq K \) where \( K \) is injective. Since \( K/M \) is projective it is a summand of \( K \), so \( M \) is a summand of \( K \) and \( M \) is injective. That c \( \Rightarrow \) d is trivial. For d \( \Rightarrow \) a, if \( L \subseteq R \) is a left ideal then \( L \) is injective and so is a summand of \( R \); thus, \( R \) is semisimple.

A ring \( R \) where every left ideal is projective is called left hereditary. A generalization of theorem 8.6 states that a submodule of a free module over such an \( R \) is a direct sum of left ideals; see [Rotman], theorem 4.4. This implies theorem 8.6 because every nonzero ideal of a principal ideal domain is isomorphic to the domain. It also implies that every submodule of a projective \( R \)-module is projective. In fact ([Rotman, theorem 4.10]) \( R \) is left hereditary iff every submodule of a projective module is projective iff every quotient of an injective module is injective. A ring where every finitely generated left ideal is projective is called left semihereditary. This will be so if every finitely generated left ideal is principal. We have seen in theorem 16.19 that regular rings have this latter property; valuation rings do also, as we will see in chapter 20.

**6. Derived functors.** Derived functors are a tool of homological algebra which is used in algebra, algebraic geometry, and algebraic topology. Basic properties can be considered abstractly, and this constitutes a branch of homological algebra.

Given an object \( M \) in an Abelian category \( A \), a projective resolution of \( M \) is an exact sequence

\[
\cdots \longrightarrow P_1 \longrightarrow P_0 \longrightarrow M \longrightarrow 0,
\]

226
where each \( P_n \) is projective. If \( A \) has enough projectives than any object \( M \) has a projective resolution. Let \( P_0 \) be a projective, with an epimorphism to \( M \). Inductively, let \( K_n \) be the kernel object of the arrow out of \( P_n \), and let \( P_{n+1} \) be a projective, with an epimorphism to \( K_n \), composed with the monomorphism \( K_n \hookrightarrow P_n \).

We remark that, to avoid class choice in this argument, the kernel and projective may be chosen from among a set of such, namely within the set of least set-theoretic rank containing such.

In \( \text{Mod}_R \), a free resolution is defined as an exact sequence as above where each \( P_n \) is free. Any \( R \)-module has a free resolution, by the same argument. A free resolution is a projective resolution.

We introduce some standard and convenient terminology regarding chain complexes. Given an object \( M \) in \( A \), a chain complex over \( M \) is a pair \( \langle C, \epsilon \rangle \) where \( C \) is a positive chain complex, \( \epsilon : C_0 \to M \), and

\[
\cdots \to C_1 \to C_0 \xrightarrow{\epsilon} M \to 0
\]

is also a chain complex, that is, \( \epsilon \partial_1 = 0 \). The complex with \( M \) is called the augmented complex, and \( \epsilon \) the augmentation arrow or simply augmentation; and the complex without \( M \) is called the deleted complex. Often \( \epsilon \) is epic, and some authors require it to be.

If \( f : M \to M' \), a morphism of chain complexes over \( f \) is a chain map \( \bar{f} : C \to C' \) where \( \langle C, \epsilon \rangle \) is a chain complex over \( M \), \( \langle C', \epsilon' \rangle \) is a chain complex over \( M' \), and \( \epsilon' f_0 = f \epsilon \). Thus, it is a chain map which when extended to the augmented complexes remains a chain map.

Dually, a complex under \( M \) is a pair \( \langle C, \mu \rangle \) where \( C \) is a positive cochain complex, \( \mu : M \to C_0 \) is monic, and \( \partial^0 \mu = 0 \). In a chain map under \( f \), \( \mu' f = \hat{f}_0 \mu \).

In a projective resolution, the deleted complex is a complex of projective objects \( P_n \), exact at \( P_n \) for \( n > 0 \), and with \( \epsilon \) a cokernel for \( \partial_1 \). We will say that \( \langle P, \epsilon \rangle \) is a projective resolution of \( M \).

The dual of a projective resolution, called an injective resolution (or coresolution), is an exact sequence

\[
0 \to M \to J^0 \to J^1 \to \cdots,
\]

where \( J^n \) is injective. If \( A \) has enough injectives, any object \( M \) has an injective resolution. Deleting \( M \) yields a cochain complex where \( \mu : M \to J^0 \) is a kernel for \( \partial^0 \).

**Theorem 19.** Given a diagram

\[
\begin{array}{ccccccccc}
\cdots & \partial_2 & P_1 & \partial_1 & P_0 & \epsilon & M & 0 \\
\downarrow & & \downarrow f & & \downarrow f & & \downarrow & \\
\cdots & d_2 & X_1 & d_1 & X_0 & \eta & N & 0
\end{array}
\]

with subexact top row, exact bottom row, and projective \( P_n \), there are arrows \( \hat{f}_n : P_n \to X_n \) yielding a commutative diagram. Further, given two such systems of arrows, the induced chain maps on the deleted complex are chain homotopic. The dual statement holds for an injective resolution, where \( f : N \to M \).

**Proof:** Since \( P_0 \) is projective, \( f \epsilon : P_0 \to N \), and \( \eta \) is epic, \( \hat{f}_0 \) exists. Inductively, for \( n > 0 \) (writing \( \partial_0 \) for \( \epsilon \), etc.), let \( d_n = me \) be a coimage-factorization. By hypothesis \( m = \text{Ker}(d_{n-1}) \), and \( d_{n-1} f_{n-1} \partial_n = \hat{f}_{n-1} \partial_n = 0 \). Hence \( \hat{f}_{n-1} \partial_n = mg \) for some \( g \). Hence there is a \( \hat{f}_n \) with \( g = \epsilon \hat{f}_n \). Then \( d_n \hat{f}_n = me \hat{f}_n = mg = \hat{f} \partial_n \). Given two chain maps \( \hat{f}, \hat{f} \) on the deleted complex, let \( h_{-1} = 0 \) (of necessity since \( P_{-1} = 0 \)). Inductively, for \( n \geq 0 \), let \( q = \hat{f} - \hat{f}_{n-1} - h_{n-1} \partial_n \), and let \( d_{n+1} = me \) be a coimage-factorization. Then \( d_n q = d_n (\hat{f}_{n-1} - \hat{f}_n) - d_n h_{n-1} \partial_n \). The second term equals \( (\hat{f}_{n-1} - \hat{f}_n) - h_{n-1} \partial_n \), or \( (\hat{f}_{n-1} - \hat{f}_n) \partial_n \). Since \( \hat{f}_n \) and \( \hat{f} \) are chain maps, \( d_n q = 0 \). Since \( m = \text{Ker}(d_n) \), \( q = mg \) for some \( g \). Since \( P_n \) is projective, \( g = \epsilon h_n \) for some \( h_n \). Then \( d_{n+1} h_n = me h_n = mg = q \).
Suppose $F : A \mapsto B$ is a functor between Abelian categories, and $\langle P, e \rangle$ is a projective resolution of $M$. Then $F(P)$ is a chain complex; if $F$ is exact, $F(P)$ will be exact at every $P_n$ with $n > 0$. In general it may not be, and the homology chain yields useful information. From hereon we assume $F$ is additive.

Suppose $A$ has enough projectives, a projective resolution $P$ is chosen for each object $M$ in $A$, and an induced chain map $\tilde{f} : P \mapsto P'$ is chosen for each arrow $f : M \mapsto M'$ (an $\tilde{f}$ exists by theorem 19). The $n$th left derived functor of $F$, with respect to these choices, maps $M$ to $H_n(F(P))$, and $f : M \mapsto M'$ to $H_n(F(\tilde{f}))$. We will show that this is a functor, and up to natural equivalence does not depend on the choices.

By theorem 19, additivity of $F$, and theorem 10.c, $H_n(F(\tilde{f}))$ is independent of the choice of $\tilde{f}$. Given projective resolutions $P_1$ and $P_2$ of $M$, let $j : P_1 \mapsto P_2$ and $k : P_2 \mapsto P_1$ be the chain maps induced by $\iota_M$. By theorem 10.c, $jk$ and $kj$ are both chain homotopic to $\iota$, since the latter is a chain map for $\iota_M$. It follows that $H_n(F(j))$ is an isomorphism from $H_n(F(P_1))$ to $H_n(F(P_2))$.

Next, the diagram

$$
\begin{array}{ccc}
H_n(F(P_1)) & \xrightarrow{H_n(F(\tilde{f}))} & H_n(F(P'_1)) \\
\downarrow H_n(F(j)) & & \downarrow H_n(F(j')) \\
H_n(F(P_2)) & \xrightarrow{H_n(F(\tilde{f}))} & H_n(F(P'_2))
\end{array}
$$

is commutative. This follows as usual from the fact that $\tilde{f}_2 j$ and $j' \tilde{f}_1$ are chain homotopic. This in turn follows because they are induced by $f \iota_M$ and $\iota_M' f$. Finally, since composition of chain maps is a chain map, by the usual argument $H_n(F(\tilde{f} g)) = H_n(F(\tilde{f})) H_n(F(\tilde{g}))$.

Other derived functors may be defined by duality. Suppose $A$ has enough projectives and enough injectives. Let $P$ denote a chosen projective resolution, $J$ a chosen injective resolution, $F$ a covariant functor, and $\tilde{F}$ a contravariant functor. Let

- $\mathcal{L}_n F(M) = H_n(F(P))$,
- $\mathcal{R}^n F(M) = H^n(F(J))$,
- $\mathcal{L}_n \tilde{F}(M) = H_n(\tilde{F}(J))$, and
- $\mathcal{R}^n \tilde{F}(M) = H^n(\tilde{F}(P))$.

Some authors, for example [Osborne], vary the notation; also, [Rotman] does not bother defining $\mathcal{L}_n \tilde{F}$. The value of any of these functors on the arrow $f$ is obtained by using $\tilde{f}_n$ on the right of the above expressions. $\mathcal{L}_n F$ and $\mathcal{R}^n F$ are covariant or contravariant, as $F$ is.

As a general principle, any fact proved for, say, $\mathcal{L}_n F$ may be stated and proved for the other functors by duality. We will give the dual statements in some cases, and leave some to the reader.

**Lemma 20.** Suppose $\epsilon : M \mapsto M'$ is an epimorphism of short exact sequences, and $f : P \mapsto M$ is a morphism, where

$$
0 \longrightarrow P_1 \xrightarrow{m_1} P_1 \oplus P_2 \xrightarrow{e_2} P_2 \longrightarrow 0,
$$

and $P_i$ is projective for $i = 1, 2$. Then there is a morphism $f : P \mapsto M'$ such that $\epsilon f' = f$.

**Proof:** Write $M$ for the short exact sequence

$$
0 \longrightarrow M_1 \xrightarrow{m} M_2 \xrightarrow{e} M_3 \longrightarrow 0,
$$

e tc. Write $m_2$ and $e_1$ for the other canonical maps of the biproduct. Let $f'_1$ be such that $\epsilon_1 f'_1 = f_1$. Let $\gamma$ be such that $\epsilon_2 \gamma = f_2 m_2$. Let $f'_2 = \gamma_2 + m' f_1 e_1$. Let $f'_3 = \epsilon' \gamma$. One verifies that $\epsilon_2 f'_2 = f_2$, $\epsilon_3 f'_3 = f_3$, $f'_2 m_1 = m' f_1$, and $f'_3 e_2 = e' f_2$. We prove the first, and leave the rest to the reader; $\epsilon_2 f'_2 = \epsilon_2 \gamma e_2 + \epsilon_2 m' f_1 e_1 = f_2 m_2 e_2 + m f_1 e_1 = f_2 m_2 e_2 + m f_1 e_1 = f_2 (m_2 e_2 + f_2 m_1 e_1) = f_2$.}

228
Theorem 21. Given a short exact sequence

\[ 0 \rightarrow M' \rightarrow M \rightarrow M'' \rightarrow 0 \]  

(2)

of objects of \( A \), and projective resolutions \((P', e')\) of \( M' \) and \((P'', e'')\) of \( M'' \), there is a projective resolution of \( M \) with \( P_n = P'_n \oplus P''_n \). The canonical monomorphisms \( m_n \) from \( P'_n \) yield a chain map \( \bar{m} : P' \mapsto P \) over \( m \), and the canonical epimorphisms \( e_n \) to \( P''_n \) yield a chain map \( \bar{e} : P \mapsto P'' \) over \( e \). The sequence

\[ 0 \rightarrow P' \xrightarrow{\bar{m}} P \xrightarrow{\bar{e}} P'' \rightarrow 0 \]  

(3)

is exact. Further, this map from the short exact sequences to the chain complexes may be extended to a functor. The dual statements hold for injective resolutions.

Proof: By theorem 13 \( P \) is projective; and \( \bar{m} \) and \( \bar{e} \) are pointwise exact. We construct \( e \) \((\partial_0)\) and \( \partial_n \) for \( n > 0 \) to make \( \bar{m} \) and \( \bar{e} \) chain maps over \( m \) and \( e \). Since \( P''_n \) is projective and \( \partial_{n-1} \) is epic there is an arrow \( \gamma_n : P''_n \mapsto P_{n-1} \) such that \( e_{n-1}\gamma_n = \partial''_n \). Let \( e'_n : P_n \mapsto P'_n \) be the canonical arrow, and let \( \partial_n = \gamma_n e_n + m_{n-1}\partial_n e'_n \). Then \( e_{n-1}\partial_n = e_{n-1}\gamma_n e_n = \partial'_n e_n \) and \( \partial_n m_n = m_{n-1}\partial_n e'_n m_n = m_{n-1}\partial'_n \). By theorem 5, \( \epsilon \) is epic. Exactness of \( P \) follows from exactness of \( P' \) and \( P'' \) by long exact sequence applied to the short exact sequence just constructed. Given a morphism of short exact sequences with arrows \( f' : M'_1 \mapsto M'_2, f : M_1 \mapsto M_2, \text{ and } f'' : M''_1 \mapsto M''_2; \) and projective resolutions \( P'_i, P_i, \text{ and } P''_i \) of each short exact sequence as in the theorem, chain maps \( \bar{f}' : P'_1 \mapsto P'_2 \), \( \bar{f} : P_1 \mapsto P_2 \), \text{ and } \bar{f}'' : P''_1 \mapsto P''_2 \) over \( f' \), \( f \), \text{ and } f'' \), which further pointwise comprise a morphism of short exact sequences, may be constructed using lemma 20 and an appropriate version of theorem 19.

Given a short exact sequence (2), and an additive functor \( F \) on \( A, F \) may be applied to the short exact sequence (3). Since (3) is split pointwise, a short exact sequence of chain complexes is obtained. Applying \( H \) yields a long exact sequence; further the sequence for \( M \) may be replaced by the one chosen for the derived functors, changing the long exact sequence only up to isomorphism. That is, there is a long exact sequence

\[ \cdots \rightarrow \mathcal{L}_1 F(M') \rightarrow \mathcal{L}_1 F(M) \rightarrow \mathcal{L}_1 F(M'') \rightarrow \mathcal{L}_0 F(M') \rightarrow \mathcal{L}_0 F(M) \rightarrow \mathcal{L}_0 F(M'') \rightarrow 0. \]

This chain complex is functorial in the short exact sequence, meaning that the object map can be extended to a functor.

The terminology “left derived functor” arises from the long exact sequence, which extends to the left. The same is true for contravariant \( F \), recalling that in this case one uses an injective resolution. For right derived functors (covariant and injective, contravariant and projective), the long exact sequences extend to the right. Note that \( \mathcal{L}_n F \) is right exact, for \( F \) either covariant or contravariant; \( \mathcal{R}^n F \) is left exact, for \( F \) either covariant or contravariant; and for \( n > 0 \) all derived functors are half exact.

By the foregoing, if \( F \) is naturally equivalent to \( \mathcal{L}_0 F \) then \( F \) is right exact. The converse is also true, provided \( \mathcal{L}_0 F(M) \) is defined appropriately; it may be defined as the codomain of \( \eta = \text{Coker}(F(\partial_1)) \). Since \( F(\epsilon)F(\partial_1) = 0 \), there is a unique \( \nu : \mathcal{L}_0 F(M) \rightarrow F(M) \) such that \( F(\epsilon) = \nu \eta \). The system of such \( \nu \) is a natural transformation from \( \mathcal{L}_0 F \) to \( F \); indeed, for \( f : M \rightarrow M' \), \( F(f)\nu \eta = \nu \mathcal{L}_0 F(f)\eta \) (exercise), so \( F(f)\nu = \nu \mathcal{L}_0 F(f) \). If \( F \) is right exact then \( \text{Coker}(F(\partial_1)) \equiv \text{Coim}(F(\epsilon)) \equiv F(\epsilon) \), and it follows that \( \nu \) is an isomorphism. The dual statements are left to the reader.

We noted above that if \( F \) is exact then \( \mathcal{L}_n F(M) = 0 \) for all \( n > 0 \) and \( M \). The dual statements hold.

If \( M \) is projective consider the resolution where \( \epsilon = \iota_m \), and \( P_n = 0 \) for \( n > 0 \). One concludes that \( \mathcal{L}_n F(M) = 0 \) for all \( n > 0 \), \( \mathcal{L}_0 F(M) \equiv M \). The dual statements hold (for projectives for \( \mathcal{L}_n F \) and \( \mathcal{R}^n F \), and injectives for \( \mathcal{L}^n F \) and \( \mathcal{R}_n F \)).
We prove the first equivalence; the proof of the second is a trivial variation. Let tools of homology.

As usual, many important applications of Ext are for the Ext functor on \( \text{Mod}_R \) for a ring \( R \), although the definition in an arbitrary Abelian category is a typical generalization. Note that in \( \text{Mod}_R \) for \( R \) commutative, \( \text{Ext}^n(M, N) \) may be enriched to an \( R \)-module; this follows because \( \text{Hom}(M, N) \) may be.

For this section, let \( \text{ext}^n(-, N) \) denote the \( n \)-th right derived functor of the contravariant functor \( \text{Hom}(-, N) \). By duality \( \text{ext}^n(M, N) \) is a bifunctor. It is in fact naturally equivalent to \( \text{Ext}^n(M, N) \) (assuming \( A \) has both enough projectives and enough injectives); proving this requires developing some more tools of homology.

A double complex in \( A \) is a family \( M^{mn} \) of objects together with arrows \( \partial^{mn} : M^{mn} \to M^{m+1,n} \) and \( \partial^{mn} : M^{mn} \to M^{m,n+1} \), such that \( \partial^2 = 0 \), \( \partial^2 = 0 \), and \( \partial^{m,n+1} \partial^{mn} = \partial^{m+1,n} \partial^{mn} \). That is, the diagram is a cochain in each row and column, and the squares are commutative. In particular, each system of maps from a column (row) to the next is a cochain map; a cochain map is the special case where \( \partial^{mn} \) is a family of cochains in each row and column, and the squares are commutative. It is common usage to use cochains for double complexes. There is no loss of generality, since a chain can be converted to a cochain by reversing it. A positive chain becomes a negative cochain. A double complex \( M \) (columns) are exact at \( n > 0 \).

One readily verifies by direct computation that if \( m + n = l \) then \( \partial_T^{l+1} h^{mn} = 0 \) (using the definition of \( h^{mn}, \partial_T^{l+1} m^{m+1,n} = h^{m+1,n}, \) etc.). There is a unique arrow \( g \) such that \( g m^{mn} = \partial_T^{l+1} h^{mn} \) for all \( m, n \) with \( m + n = l \). It follows that \( \partial_T^{l+1} \partial_T^l = 0 \), so the total complex is in fact a cochain complex.

For the following theorem let \( M^m \) denote the \( m \)-th “column” cochain and \( \partial^m \) the chain map from \( M^m \) to \( M^{m+1} \); and similarly for \( \bar{M}n \) and \( \bar{\partial}^n \). Let \( k^{mn} : Z^{mn} \to M^{mn} \) be the kernel of \( \partial^m \), and \( \bar{k}^{mn} : \bar{Z}^{mn} \to M^{mn} \) the kernel of \( \bar{\partial}^n \); let \( Z^n \) and \( \bar{Z}^n \) denote the column and row cochains. Let \( \cong \) denote natural equivalence.

**Theorem 22.** Suppose \( M^{mn} \) is a double complex in the first quadrant. Suppose also that the rows (columns) are exact at \( M^{mn} \) for \( m > 0 \). Then

\[
H^n(M_T) \cong H^n(Z^0) \cong H^n(\bar{Z}^0).
\]

**Proof:** We prove the first equivalence; the proof of the second is a trivial variation. Let \( k^n : Z^n_T \to Z^n_{0} \) be the kernel of the induced arrow \( \partial^n_Z : Z^n_{0} \to Z^{n+1}_{0} \). Direct computation shows that \( \partial^n_Z m^n x^n k^n 0 = 0 \), yielding an induced arrow \( j^n : Z^n_2 \to Z^n_2 \). We give the rest of the proof only for \( \text{Mod}_R \) for a ring \( R \). In this case \( j^n(x_0, \ldots, x_n) = (\partial_Z^n x_0, \partial_Z^n(x_0) - \bar{\partial}^{n-1}(x_1), \ldots, \partial_Z^n(x_n)) \); and \( j^n(x) \) equals \( (x_0, \ldots, 0) \). One sees that \( x \in B^n_T \) iff \( j^n(x) \in B^n_T \), whence \( j^n \) induces an injection on the homology modules. Suppose \( x \in Z^n_T \) where \( x = (x_0, \ldots, x_k, 0, \ldots, 0) \) for some \( k > 0 \). Then \( \partial^n x)^{-k}(x_k) = 0 \), so by the exactness hypothesis \( x_k = \partial^{k-1,n-k}(w) \) for some \( w \). Letting \( x' = (x_0, \ldots, x_{k-1} - w, 0, \ldots, 0) \), one verifies that \( x - x' \in B^n_T \).
Proceeding inductively, any \( x \in Z^n_R \) is homologous to an element of \( j^n[Z_2^n] \), and the induced map on homology modules is an isomorphism. Now, the double complexes form a category, a full subcategory of a functor category. The map \( M \mapsto M_T \) from this category to the cochain complexes is functorial. Given \( f : M_1 \mapsto M_2 \) (i.e., maps \( f^{mn} : M_1^{mn} \mapsto M_2^{mn} \) making the combined diagram commutative), the induced map (call it \( f_T \)) is a collection of maps \( f_T^n : M_1^n \mapsto M_2^n \). In \( \text{Mod}_R \), \( f_T^n \) acts componentwise, and clearly \( j_T^n(f^{mn}(x)) = f_T^n(j_T^n(x)) \). Naturality of the isomorphisms follows.

The above theorem has been proved only in \( \text{Mod}_R \). It is true in any Abelian category; a direct proof is lengthy and unnecessary. The result for Abelian categories follows from that for \( \text{Mod}_R \) for all \( R \), by the “full embedding theorem”, which states that “every small Abelian category admits a full, exact, covariant embedding into \( \text{Mod}_R \) for some \( R \)” ([Mitchell], theorem VI.7.2).

To prove a theorem about Abelian categories using this theorem one proceeds as follows. Take a finite diagram containing the objects and arrows of interest. Enlarge it to a small full Abelian subcategory. Embed this in some \( \text{Mod}_R \). Prove the theorem in the image subcategory of \( \text{Mod}_R \). Conclude that it follows in the small category, and hence in the category. The method applies to the preceding theorem, noting that for each \( n \) it is a statement about a finite diagram.

The reader might wonder why Abelian category theory isn’t done this way, for example homology theory. Indeed in [Mitchell] the snake lemma is proved using an embedding argument. The proof of the embedding theorem is fairly involved, and giving direct proofs of basic theorems increases the reader’s understanding of Abelian categories. However, the line must be drawn somewhere, and the proof of the above theorem is best given using the embedding theorem, so the full proof is omitted here.

Let \( \langle P, e \rangle \) be a projective resolution of \( M \), and let \( \langle J, \mu \rangle \) be an injective resolution of \( N \). Consider the diagram below, where the arrows are obtained via the Hom functors. This satisfies the hypotheses of theorem 22. Since Hom is left exact in either argument, \( Z^0 \cong \text{Hom}(P^+, N) \) where \( P^+ \) is the augmented complex, and \( Z^0 \cong \text{Hom}(M, J^+) \). It follows that \( \text{Ext}^n(M, N) \) is isomorphic to \( \text{ext}^n(M, N) \) (note that 1 must be added to \( n \)).

To prove the naturality of the isomorphism, a further argument is required. Suppose \( f : M \mapsto M' \); as usual let \( f : P \mapsto P' \) be the chosen chain map. There are two diagrams as above, and considering the columns yields maps from \( \text{Hom}(P^+_n, J^{+m}) \), yielding a diagram which is commutative in the columns; it is readily verified to be commutative in the rows also. Arguing in \( \text{Mod}_R \), naturality follows as in the proof of theorem 22.

**Theorem 23.** If \( \text{Ext}^1(N, M) = 0 \) then for any \( L \), the short exact sequence \( 0 \mapsto M \mapsto L \mapsto N \mapsto 0 \) splits.

**Proof:** \( \text{Hom}(N, M) \) is left exact in either argument, so \( \text{ext}^0(N, M) \cong \text{Hom}(N, M) \). If \( \text{Ext}^1(N, M) = 0 \) then for any \( L \), by the long exact sequence for \( \text{Hom}(-, M) \),

\[
\begin{align*}
0 & \longrightarrow \text{Hom}(N, M) \longrightarrow \text{Hom}(L, M) \longrightarrow \text{Hom}(M, M) \longrightarrow 0
\end{align*}
\]
is exact. Thus there is a \( g \in \text{Hom}(L, M) \) such that \( gm = \iota_M \), and the short exact sequence is split.

Facts such as this are the origin of the name “Ext” for this functor; a short exact sequence as in the theorem is called an extension of \( M \) by \( N \). Although we will not prove it, the converse of the theorem holds. Indeed, for an integer \( n \geq 0 \) and objects \( M, N \) let \( S^n(M, N) \) be the exact sequences

\[
\begin{array}{ccccccccc}
0 & \rightarrow & M & \rightarrow & L_1 & \rightarrow & \cdots & \rightarrow & L_n & \rightarrow & N & \rightarrow & 0
\end{array}
\]

(this is a proper class). For \( s_1, s_2 \in S^n(M, N) \) define \( R_1(s_1, s_2) \) to hold if there is a chain map from \( s_1 \) to \( s_2 \), which is the identity on \( M \) and \( N \) (this is also). Define \( \equiv \) to be the symmetric transitive closure. It can be shown (appropriate care being taken regarding the metamathematical issues) that there is a natural equivalence from \( S^n(M, N)/\equiv \) to \( \text{Ext}^n(M, N) \); see [Mitchell], VII.7. In the case \( n = 1 \) the split short exact sequences are the 0 element.

If \( J \) is injective then \( \text{Hom}(-, J) \) is exact, so by general facts about derived functors, \( \text{Ext}^n(M, J) = 0 \) for all \( n > 0 \) and all \( M \). Conversely, if \( \text{Ext}^1(M, J) = 0 \) for all \( M \), then by theorem 23 and theorem 13 \( J \) is injective. Dually (and using the natural equivalence of Ext and ext), \( P \) is projective iff \( \text{Ext}^n(P, N) = 0 \) for all \( n > 0 \) and all \( N \), iff \( \text{Ext}^1(P, N) = 0 \) for all \( N \).

Next we will show that \( \text{Ext}^n(M, -) \) preserve products; the proof will be given in a complete Abelian category with enough injectives. By suitable dualization, \( \text{Ext}^n(-, N) \) preserves products (maps coproducts to products) in a cocomplete Abelian category with enough projectives. The proof is of a type known as dimension shifting. The claim follows for \( n = 0 \) since it holds for \( \text{Hom}(M, -) \).

Given a product \( \times_i N_i \), for each \( i \) choose an exact sequence

\[
\begin{array}{ccccccccc}
0 & \rightarrow & N_i & \rightarrow & J_i & \rightarrow & Q_i & \rightarrow & 0
\end{array}
\]

where each \( J_i \) is injective. By remarks above,

\[
\begin{array}{ccccccccc}
0 & \rightarrow & \times_i N_i & \rightarrow & \times_i J_i & \rightarrow & \times_i Q_i & \rightarrow & 0
\end{array}
\]

is exact; also \( \times_i J_i \) is injective. Taking the long exact sequence for Ext, the objects \( \text{Ext}^n(M, \times_i J_i) \) for \( n > 0 \) vanish, yielding a series of finite exact sequences.

The first subsequence is

\[
\begin{array}{ccccccccc}
0 & \rightarrow & \text{Hom}(M, \times_i N_i) & \rightarrow & \text{Hom}(M, \times_i J_i) & \rightarrow & \text{Hom}(M, \times_i Q_i) & \rightarrow & \text{Ext}^1(M, \times_i N_i) & \rightarrow & 0
\end{array}
\]

There is also an exact sequence

\[
\begin{array}{ccccccccc}
0 & \rightarrow & \times_i \text{Hom}(M, N_i) & \rightarrow & \times_i \text{Hom}(M, J_i) & \rightarrow & \times_i \text{Hom}(M, Q_i) & \rightarrow & \times_i \text{Ext}^1(M, N_i) & \rightarrow & 0
\end{array}
\]

In the first three positions, the objects are isomorphic. In the fourth position, there is an arrow from the object in the first sequence to the object in the second; by theorem 6 this arrow is an isomorphism. This proves the case \( n = 1 \).

For \( n > 1 \) there is a subsequence

\[
\begin{array}{ccccccccc}
0 & \rightarrow & \text{Ext}^{n-1}(M, \times_i Q_i) & \rightarrow & \text{Ext}^n(M, \times_i N_i) & \rightarrow & 0;
\end{array}
\]

and also an exact sequence

\[
\begin{array}{ccccccccc}
0 & \rightarrow & \times_i \text{Ext}^{n-1}(M, Q_i) & \rightarrow & \times_i \text{Ext}^n(M, N_i) & \rightarrow & 0.
\end{array}
\]

232
The arrows are isomorphisms, and by induction the objects in the first position are isomorphic.

8. Flat modules and Tor. Flatness and Tor involve the tensor product, so in this section $A$ will be the category of $R$-modules for a ring $R$, although as mentioned previously a tensor product may be defined in other Abelian categories. Also, for the noncommutative case right $R$-modules must be considered, so in this section the notation of section 18.4 is used, with $_R\text{Mod}$ denote the left $R$-modules and $\text{Mod}_R$ the right $R$-modules. Recall that $\otimes_R$, or simply $\otimes$, is a bifunctor from $\text{Mod}_R \times _R\text{Mod}$ to Ab.

A right $R$-module $M$ is said to be flat if $M \otimes -$ preserves monics (if $m : N_1 \hookrightarrow N_2$ is monic then $\iota_M \otimes m : M \otimes N_1 \hookrightarrow M \otimes N_2$ is monic). Similarly to theorem 13.a, $M$ is flat iff $M \otimes -$ is exact ($M \otimes -$ is right exact because it preserves colimits, and a functor between Abelian categories preserves monics iff it preserves kernels).

A left $R$-module is flat if $- \otimes M$ preserves monics, iff $- \otimes M$ is exact. When $R$ is commutative, $M$ is simply an $R$-module; since $M \otimes N$ is naturally equivalent to $N \otimes M$, $M$ is flat iff $M \otimes -$ preserves monics iff $- \otimes M$ preserves monics. One readily verifies that for commutative $R$, the tensor product of flat $R$-modules is flat.

Given a map $f : N_1 \hookrightarrow N_2$ and modules $M_i$, $(\oplus_i M_i) \otimes N_j$ is naturally isomorphic to $\oplus_i (M_i \otimes N_j)$ for $j = 1, 2$, whence $(\oplus_i M_i) \otimes -$ preserves monics iff each $M_i \otimes -$ does. Thus, $\oplus_i M_i$ is flat iff every $M_i$ is flat (cf. theorem 13.c). A fortiori a summand of a flat module is flat.

$R$ (considered to be a right $R$-module) is flat, since $R \otimes N$ is naturally isomorphic to $N$. Thus, any free right $R$-module is flat, and thus any projective right $R$-module is flat (being a summand of a free module). We will see below that the converse does not hold. A flat resolution of an $R$-module $M$ is an exact sequence as in a projective resolution, where each $P_n$ is flat. These clearly exist, and sometimes it is of interest to note that one suffices.

As observed above, in $_R\text{Mod}$ the direct limit functor for any index set is exact. Also, tensor product preserves direct limits (since it is the left adjoint). It follows that the direct limit of flat modules is flat. It is easily seen that a module is the direct limit of its finitely generated submodules (in fact it is the union of them), and so a module is flat if all its finitely generated submodules are.

In the natural isomorphism for the general tensor product, let $S = T = Z$. Then for $L \in _R\text{Mod}$, $M \in \text{Mod}_R$, and $N \in \text{Ab}$, $\text{Hom}(M, N)$ in $\text{Ab}$ is a left $R$-module with the action $(rf)(m) = f(mr)$; $M \otimes L \in \text{Ab}$; and in $\text{Ab}$ $\text{Hom}(M \times L, N) \cong \text{Hom}(L, \text{Hom}(M, N))$.

Theorem 24. With notation as in the preceding paragraph, if $N$ is an injective cogenerator then $M$ is flat iff $\text{Hom}(M, N)$ is injective.

Proof: $\text{Hom}(M, N)$ is injective iff $\text{Hom}(-, \text{Hom}(M, N))$ is exact iff $\text{Hom}(M \times -, N)$ is exact. Since $N$ is injective, $\text{Hom}(-, N)$ is exact. If $M$ is flat then $M \otimes -$ is exact, whence, since the composition of exact functors is exact, $\text{Hom}(M \times -, N)$ is exact. Suppose $\text{Hom}(M \times -, N)$ is exact. If $0 \to L_1 \to L_2 \to L_3 \to 0$ is an exact sequence in $_R\text{Mod}$ then $0 \to \text{Hom}(M \times L_1, N) \to \text{Hom}(M \times L_2, N) \to \text{Hom}(M \times L_3, N) \to 0$ is an exact sequence in $\text{Ab}$. Since $N$ is a cogenerator, $0 \to M \times L_1 \to M \times L_2 \to M \times L_3 \to 0$ is an exact sequence in $\text{Ab}$. This shows that $M$ is flat.

Theorem 25. Suppose $M \in \text{Mod}_R$, and whenever $I \subseteq R$ is a finitely generated ideal the induced map $M \otimes I \hookrightarrow M \otimes R$ is monic. Then $M$ is flat.

Proof: Let $N$ be an injective cogenerator in $\text{Ab}$. Since any ideal $I$ is the union of its finitely generated submodules, and since $M \otimes -$ preserves colimits, $0 \to M \otimes I \to M \otimes R$ is exact for any ideal $I$, whence $\text{Hom}(M \otimes R, N) \to \text{Hom}(M \otimes I, N) \to 0$ is exact for any ideal $I$, whence $\text{Hom}(R, \text{Hom}(M, N)) \to \text{Hom}(I, \text{Hom}(M, N)) \to 0$ is exact for any ideal $I$. By lemma 16 $\text{Hom}(M, N)$ is injective. By theorem 24 $M$ is flat.

233
Recall that if $\phi : R \to S$ is a homomorphism of commutative rings then $S$ is an $R$-module (in fact $R$-algebra) with the action $rs = \phi(r)s$. The homomorphism is called flat if $S$ is a flat $R$-module. One readily verifies the following.

- If $\phi$ is a flat homomorphism and $M$ is a flat $S$-module then $M$ is a flat $R$-module.
- The composition of flat homomorphisms is flat.

The canonical map $R \to R_S$ where $R$ is commutative and $S \subseteq R^\times$ is a multiplicative subset is a flat homomorphism. This follows by the isomorphism between $R_S \otimes M$ and $M_S$ given in section 18.4, and the observation in section 8.6 that if $N \subseteq M$ then $N_S$ may be considered a submodule of $M_S$.

Another example of a flat homomorphism is the inclusion of a commutative ring $R$ in the ring $R[x]$ of multinomials. Indeed, the latter is a free $R$-module.

It follows from theorem 25 that $Q$ is a flat commutative group. It is not projective; we prove this by proving a fact of interest in itself, if an integral domain $R$ is not a field, then the only $R$-linear map from $R_{R^\times}$ to $R$ is the zero map. Given this, the identity map in $R_{R^\times}$ does not factor through the inclusion of $R$, and since the inclusion is epic, $R_{R^\times}$ is not a projective $R$-module.

To prove the fact, suppose $R$ an integral domain and $f : R_{R^\times} \to R$ is $R$-linear; then $1 \cdot f(1) = f(1 \cdot 1) = f(1)$, so if $f(1) \neq 0$ then $f(1) = 1$. Also $sf(1/s) = f(1)$, and it follows that $f$ is the identity and $R$ is a field.

For $M \in \text{Mod}_R$ $M \otimes -$ is a functor from $\mathbf{R}_R$ to $\mathbf{Ab}$. The $n$th left derived functor is denoted $\text{Tor}_n(M, -)$. As for $\text{Ext}^n$, $\text{Tor}_n(M, N)$ is readily verified to be a bifunctor from $\text{Mod}_R \times \text{Mod}_R$ to $\mathbf{Ab}$; given an arrow $f : M \to M'$, an object $N$, and a projective resolution of $N$, let $C$ denote the augmented complex. Applying the bifunctor $\otimes$ yields a chain map from $M \otimes C$ to $M' \otimes C$. This in turn yields a map from $\text{Tor}_n(M, N)$ to $\text{Tor}_n(M', N)$.

Tor could be defined more generally on $\mathbf{S}_{\text{Mod}_R} \times \mathbf{R}_{\text{Mod}_R}$, but the definition given is the common case. If $R$ is commutative $M \otimes N$ may be considered an $R$-module, and $\text{Tor}_n$ may also.

As for $\text{ext}$, temporarily let $\text{tor}_n(-, N)$ denote the $n$th left derived functor of the contravariant functor $- \otimes N$. Similarly to $\text{ext}$ and $\text{Ext}$, the bifunctor $\text{tor}$ is naturally equivalent to $\text{Tor}$; we indicate the changes required to the argument. Projective resolutions $P$ of $M$ and $Q$ of $N$ may be considered cochains by reversal. This results in a double complex in the third quadrant. Replace $Z^0$ and $\hat{Z}^0$ by $C^0$ and $\hat{C}^0$, the cokernels of the arrows $\partial^0n$ and $\hat{\partial}^0m$. Theorem 22 holds with these modifications. It may be applied to the double complex with entries $P^+_m \otimes Q^+_n$. Since $\otimes$ is right exact in either argument, $C^0 \cong P^+ \otimes N$ and $\hat{C}^0 \cong M \otimes P^+$. The above proof that $\text{tor} \cong \text{Tor}$ is unchanged if $P$ or $Q$ is replaced by a flat resolution. Thus, a flat resolution may be used in determining $\text{Tor}_n(M, N)$, in either argument.

Recalling the opposite ring defined in section 16.5, a right (left) $R$-module $M$ yields a left (right) $R^{\text{op}}$-

module $M'$ in an obvious manner. One verifies that $N' \otimes M' \cong M \otimes N$, and $P'$ is projective if $P$ is; and that $\text{Tor}_n(N', M') \cong \text{Tor}_n(M, N)$. For $R$ commutative, $\text{Tor}_n(N, M) \cong \text{Tor}_n(M, N)$.

Other facts which follow by typical arguments include the following.

- $\text{Tor}_0(M, N) \cong M \otimes N$. This follows since $\otimes$ is right exact.
- If $M$ is flat then $\text{Tor}_n(M, N) = 0$ for $n > 0$ and any $N$. This follows by exactness of $M \otimes -$ and general facts about derived functors.
- For any $M \in \text{Mod}_R$, the map $M \otimes I \to M \otimes R$ is monic iff $\text{Tor}_1(M, R/I) = 0$. This follows by considering the long exact sequence arising from the short exact sequence $0 \to I \to R \to R/I \to 0$.
- If $\text{Tor}_1(M, R/I) = 0$ for any finitely generated left ideal $I \subseteq R$ then $M$ is flat. This follows by the preceding and and theorem 25.

$\text{Tor}_n$ preserves direct limits. The proof is similar to the proof that $\text{Ext}^n(M, -)$ preserves products (dimension shifting), with the following differences. The case $n = 0$ follows since tensor product preserves direct limits. An exact sequence $0 \to K_i \to P_i \to N_i \to 0$ is used, where $P_i$ is projective (it suffices that $P_i$ be flat). The fact that the direct limit functor is exact in $\mathbf{R}_R$ is used. The long exact sequence extends
to the left, so the sequence for \( n = 1 \) is “reversed”, and a different case of theorem 6 is used. The sequences for \( n > 1 \) are also reversed.

For a left ideal \( I \subset R \), the composition \( M \otimes I \to M \otimes R \to M \) of the multiplication map with the induced map is always a surjection to \([MI]\) (the submodule generated by \( MI \)). When \( M \) is flat it is an isomorphism.

Suppose \( M \) is flat, \( r \in R \), and the map \( \psi_r (r') = r'r \) from \( R \) to \( R \) is injective (\( r \) is not a right zero divisor). The map \( M \to M \otimes R \to M \otimes R \to M \), where the middle map is that induced by \( \psi_r \), applied to \( m \), yields \( m \to m \otimes 1 \to m \otimes r \to mr \). Thus, the map \( \phi_r (m) = nr \) is injective.

In particular, if \( R \) is an integral domain and \( M \) is flat then \( M \) is torsion-free.

With \( r \) as above the exact sequence \( 0 \to R \to R \to R/Rr \to 0 \) is in fact a free resolution of \( R/Rr \).

Tensoring with \( M \) and using the multiplication map, it follows that
- \( \text{Tor}_0(M, R/Rr) = M/ Mr \), the cokernel of \( \phi_r \);
- \( \text{Tor}_1(M, R/Rr) = M_r \), the kernel of \( \phi_r \); and
- \( \text{Tor}_n(M, R/Rr) = 0 \) for \( n \geq 2 \).

Suppose \( R \) is a principal ideal domain and \( M \) is torsion-free. Then by the preceding paragraph \( \text{Tor}_1(M, R/I) = 0 \) for every nonzero ideal \( I \). \( \text{Tor}_1(M, R) \) always equals 0 since \( R \) is projective. Thus, \( M \) is flat.

An example of a torsion-free module over an integral domain which is not flat may be found in [Eisenbud], exercise 6.6.

The following theorem illustrates the origin of the notation \( \text{Tor}_n \). Note that this is the object map of a functor, where the image of \( f \) is the restriction / corestriction.

**Theorem 26.** Suppose \( R \) is an integral domain, and \( M \) an \( R \)-module. Let \( Q \) denote \( R/R \neq \), and let \( J \) denote \( Q/R \). Then \( \text{Tor}(M) \cong \text{Tor}_1(M, J) \).

**Proof:** \( R \) and \( Q \) are both flat, so from the long exact sequence for \( 0 \to R \to Q \to J \to 0 \), \( \text{Tor}_n(M, J) = 0 \) for \( n \geq 2 \). If \( M \) is torsion then \( M \otimes Q = 0 \) (if \( rm = 0 \) for \( r \neq 0 \) then \( m \otimes r' = (rm) \otimes (r/r') = 0 \)), so \( \text{Tor}_1(M, J) \cong M \otimes R \cong M \). If \( M \) is torsion-free, as observed in section 8.6 \( M \to M \otimes Q \) is an embedding; also, \( M \otimes Q \) is a vector space over \( Q \). Thus there is an exact sequence \( 0 \to M \to V \to N \to 0 \) where \( V \) is a flat \( R \)-module. From the long exact sequence, \( \text{Tor}(M, J) = 0 \). From the long exact sequence for \( 0 \to \text{Tor}(M) \to M \to M/\text{Tor}(M) \to 0 \), the sequence

\[
\text{Tor}_2(M/\text{Tor}(M), J) \to \text{Tor}_1(\text{Tor}(M), J) \to \text{Tor}_1(M, J) \to \text{Tor}_{1}(M/\text{Tor}(M), J)
\]

is exact. The first term is 0, and as already noted the second term is isomorphic to \( \text{Tor}(M) \) and the last term is 0. Naturality is left to the reader.

Finally, we mention that a right \( R \)-module \( M \) is said to be faithfully flat if \( M \otimes - \) is exact and faithful. This may be seen to be the case iff \( M \) is flat, and for any \( N \in R\text{Mod} \), if \( N \neq 0 \) then \( M \otimes N \neq 0 \) (exercise 11). Various additional facts concerning flatness and Tor may be found in standard references, including [Eisenbud] and [Matsumura].

9. **Cohomology of groups.** Homological algebra was developed partly because of applications of homology in algebra. A basic example is a cohomology theory which arises in the study of group extensions. We will give a brief introduction.

An exact sequence of groups \( 1 \to N \to G \to Q \to 1 \) is called an extension of \( N \) by \( Q \). Two extensions are said to be equivalent if there is a morphism of short exact sequences between them, which is the identity at \( N \) and \( Q \). In the following lemma, let \( \gamma \) denote the remaining map.

**Lemma 27.** Given an equivalence of extensions, \( \gamma \) is an isomorphism.
Then the Abelian group of functions $f$ may be taken to be inclusion, by “transferring” the group multiplications via bijections. $G$ is also said to be a semidirect product of an extension. If an extension is split, there is an extension system where $\Phi$. Further $\Phi$ forms an Abelian group under this operation. At least one semidirect product, namely the direct product, where $\Phi$ is also true. For such an extension system, the map $\alpha \in \{1, n\}$, $\alpha_\Phi$ is said to comprise an extension system. Given an extension system, the operation on $\alpha \in \{1, n\}$, $\alpha_\Phi$ is said to be equivalent if there are elements $n_q \in N$ for $q \in Q$, and factors $\alpha_\Phi, r \in N$ for $q, r \in Q$ satisfying 1-3, are said to comprise an extension system. Given an extension system, the operation on $Q \times N$ defined according to 0 (that is, $\langle g_1, n_1 \rangle \langle g_2, n_2 \rangle = \langle g_1, n_1 \rangle q(g_1, n_2) \rangle$ is a group operation on $Q \times N$. Ignoring the map $\alpha \in \{1, n\}, N \in Q \times N$; the map $\langle q, n \rangle \mapsto q$ is an epimorphism to $Q$ with kernel $N$.

Two extension systems are said to be equivalent if there are elements $n_q \in N$ for $q \in Q$, such that $1. \alpha^2_q(n) = n^{-1}_q \alpha^1_q(n)n_q; \alpha^2_q(n)\alpha^2_q(n)$. Given $n \in N$, $\alpha_q$ of $N$ for $q \in Q$, and factors $\alpha_\Phi, r \in N$ for $q, r \in Q$ satisfying 1-3, are said to comprise an extension system. Given an extension system, the operation on $Q \times N$ defined according to 0 (that is, $\langle g_1, n_1 \rangle \langle g_2, n_2 \rangle = \langle g_1, n_1 \rangle q(g_1, n_2) \rangle$ is a group operation on $Q \times N$. Ignoring the map $\alpha \in \{1, n\}, N \in Q \times N$; the map $\langle q, n \rangle \mapsto q$ is an epimorphism to $Q$ with kernel $N$.

Two extension systems are equivalent if the corresponding extensions are (the embeddings of the extensions may be taken to be inclusion, by “transferring” the group multiplications via bijections).

An extension $N \in G$ is said to be split if there is a homomorphism $q \mapsto \bar{q}$ which is a right inverse to $e$; $G$ is also said to be a semidirect product of $N$ by $Q$. An extension isomorphic to a split extension is a split extension. If an extension is split, there is an extension system where $\alpha_\Phi, r \in N$ for all $q, r$; the converse is also true. For such an extension system, the map $\alpha \mapsto \bar{q}$ is a group homomorphism. There is always at least one semidirect product, namely the direct product, where $\alpha_q = 1$ for all $q$.

If $Q$ is Abelian various simplifications may be made. The map $q \mapsto \alpha_q$ is a homomorphism, and is the same for equivalent extension systems. Let $\Phi$ be the factor sets with respect to such a homomorphism (i.e., those satisfying requirement 2 for an extension system). Given $\alpha^1, \alpha^2 \in \Phi$, their pointwise sum is again in $\Phi$. Further $\Phi$ forms an Abelian group under this operation.

Continuing to suppose $N$ to be Abelian, the map $Q \mapsto Aut(N)$ is an action of $Q$ on $N$, as defined in section 16.1; we say that $N$ is a $Q$-module, and write $q \cdot n$ for $\alpha_q(n)$. Given a $Q$-module $N$, let $C^n$ be the Abelian group of functions $f : Q^n \mapsto N$. For example, a factor set is an element of $C^2$.

Let $\partial_n : C^n \mapsto C^{n+1}$ be defined as follows.

$$
\partial f(q_0, \ldots, q_n) = \sum_{i=1}^{n} (-1)^i f(q_0, \ldots, q_i, q_i+1, \ldots, q_n) + (-1)^n f(q_0, \ldots, q_{n-1})
$$

Then $\partial^2$ may be verified to be zero, so that with these maps the modules $C^n$ form a cochain complex. In particular,

$$
\partial^0 n(q_0) = q_0 \cdot n - n

\partial^1 f(q_0, q_1) = q_0 \cdot f(q_1) - f(q_0q_1) + f(q_0)

\partial^2 f(q_0, q_1, q_2) = q_0 \cdot f(q_1, q_2) - f(q_0q_1, q_2) + f(q_0, q_1q_2) - f(q_0, q_1)
$$

236
One verifies the following.
- Since \( B^0 = 0 \), \( H^0 = Z^0 \), the elements of \( N \) which are fixed by every automorphism \( \alpha_q \).
- \( Z^2 \) is the factor sets for the given \( Q \)-module (provided requirement 3 is ignored).
- \( B^2 \) is the factor sets corresponding to splitting extensions.
- Two factor sets are equivalent iff they differ by a 2-boundary.
- Every extension of \( N \) by \( Q \) with the given automorphisms is isomorphic to one determined by an element of \( H^2 \), and distinct elements determine non-isomorphic extensions. Alternatively, one may consider a representative set of extensions, such as those where the group multiplications is defined on the set \( N \times Q \), and establish a bijection of quotients of equivalence relations.

\( Z^1 \) is known as the crossed homomorphisms, and \( B^1 \) as the principal ones, of the \( Q \)-module \( N \).

The cohomology groups defined above are in fact those of Ext functors. Let \( Z[Q] \) be the free \( Z \)-algebra over \( Q \), that is, the expressions \( \sum k_i q_i \) where the \( q_i \) are distinct and \( k_i \in Z \), with the obvious addition and multiplication. The \( Q \)-modules are in obvious bijective correspondence with the \( Z[Q] \)-modules: \( \sum k_i q_i \) applied to \( n \in N \) yields \( \sum k_i (q_i \cdot n) \), an element of \( N \) (a similar extension of a group action to a ring action was used in section 15.7).

Let \( P_n \) be the free \( Z \)-module \( Z[Q]^{\otimes (n+1)} \); the elements \( g_0 \otimes \cdots \otimes g_n \) are a basis. \( P_n \) becomes a \( Z[Q] \)-module by defining \( q \cdot (g_0 \otimes \cdots \otimes g_n) \) to be \( (q \cdot g_0) \otimes \cdots \otimes g_n \). These modules comprise a free resolution in the \( Z[Q] \)-modules of \( Z \); applying the \( \text{Hom} \) functor yields the cochain complex given above. Exactness is proved by proving chain contractibility. See [Jacobson] for details.

**Exercises.**

1. Show that if \( C \) is an exact sequence category then so is \( C^J \). Hint: First, if \( \text{Ker}(f) = 0 \) then this holds pointwise, whence \( f \) is pointwise monic, whence monic. Second, images may be defined pointwise. Suppose \( f : a \mapsto b \), and \( \eta \) is a natural transformation from \( F_1 \) to \( F_2 \). Let \( m_n : c_n \mapsto F_2(a) \) be an image for \( \eta_n \), and \( m_b : c_b \mapsto F_2(b) \) for \( \eta_b \). Then \( F_2(f)m_n \) is monic, so factors uniquely through \( c_b \). This yields the arrow map of the functor \( H \), where \( H(a) \) is the chosen image object \( c_n \) (the composition axiom following as usual by uniqueness). The system of maps \( m_n \) is clearly a natural transformation from \( H \) to \( F_2 \). The pointwise maps \( a \mapsto H(a) \) also comprise a natural transformation. Any other monic in \( C_J \) to \( F_2 \) factors pointwise through its domain, and the claim follows by arguments similar to those already given. In fact, a pointwise coimage-image factorization yields a coimage-image factorization.

2. In the notation of theorem 1, suppose that a colimit \( D \) of the \( F_k \), and a limit \( D^* \) of the \( F_k \), exist; write \( \alpha_{jk} : F(j, k) \mapsto D(j) \), \( \alpha^*_{jk} : D^*(k) \mapsto F(j, k) \). Suppose \( \langle d, \delta \rangle \) is a colimit for \( D \) and \( \langle d^*, \delta^* \rangle \) is a limit for \( D^* \). Then for fixed \( k \) the arrows \( \alpha_{jk} \alpha^*_{jk} \) form a cone from \( d^*_k \) to \( D \), and hence factor through \( d \), as \( \delta \zeta_k \) for some \( \zeta_k \). The \( \zeta_k \) form a cone from \( D^* \) to \( d \), so factor through \( d^* \), as \( \eta \delta^*_k \) for some \( \eta : d^* \mapsto d \). Show that if \( C \) is Set, \( J \) is finite, and \( K \) is a directed preorder, then \( \eta \) is an isomorphism. Show that this in fact holds when \( C \) is \( \text{Mdl}_T \) for \( T \) a set of equations. Hint: Elements of \( d^* \) are of the form \( \langle x_1, \ldots, x_J \rangle \) where \( x_j \in F_j \) for some \( k \) and the equivalence relation is on the disjoint union of the \( d^*_k \). Elements of \( d \) are of the form \( \langle \langle x_1 \rangle, \ldots, \langle x_J \rangle \rangle \) where \( x_j \in F_j \) for some \( k \) and the \( \delta \) equivalence relation is on the disjoint union of the \( F_j \). The map \( \eta \) takes \( \langle \langle x_1, \ldots, x_J \rangle \rangle \) to \( \langle \langle x_1 \rangle, \ldots, \langle x_J \rangle \rangle \). If \( \langle \langle x_1, \ldots, x_J \rangle \rangle \) is consistent then we can choose \( k \) and the \( x_j \) so that \( \langle x_1, \ldots, x_J \rangle \) is consistent; thus, \( \eta \) is surjective. If \( \langle x_j \rangle = \langle x'_j \rangle \), all \( j \in J \), we can choose \( k \) so that \( x_j, x'_j \) map to the same element, all \( j \in J \), and so \( \eta \) is injective. If \( L \) has no relations then \( \eta \) is an isomorphism since it is a bijective homomorphism. One verifies directly that \( \eta \) preserves a relation \( R \).

3. Prove Theorem 5. Hint: Show \( egK = 0 \), so \( \exists g' (yK = mg') \); but \( g' = 0 \) because \( n'f g' \) does.

4. Prove Theorem 6. Hint: From \( f_3 x = 0 \) deduce \( g_3 x = 0 \), so \( g_2 y = x c_2 \); then \( h_2 f_2 y = 0 \), so \( h_1 y' = f_2 y c_1 \); then \( y c_0 = f_1 z \). Conclude \( f_2 g_1 z = f_2 y c_1 c_0 \), so \( g_1 z = y c_1 c_0 \); whence conclude \( x c_2 y c_1 c_0 = 0 \).

5. Prove Theorem 7. Hint: All claims follow as for a morphism of short exact sequences, except that
e_0 : b_0 \mapsto c_0 is epic. This follows by the snake lemma. For a direct proof, show that h_1 \equiv \text{Coker}(e_0) (by dualizing a proof that m_0 \equiv \text{Ker}(e_0), which is similar to one already given). Then \text{Coim}(h_1) \equiv \text{Coker}(h_0e_0), so h_0 \equiv \text{Ker}(h_1) \equiv \text{Im}(h_0e_0) and e_0 \equiv \text{Coim}(h_0e_0) is epic. The proof for the third row is dual.


7. Suppose M is an R-module for a commutative ring R, and N is a submodule. N is said to be a pure submodule if rN = N \cap rM for any r \in R. Show the following.
   a. N is pure iff, for r \in R, n \in N, if rm = n for some m \in M then rn_1 = n for some n_1 \in N.
   b. M/N is torsion-free iff, for r \in R, r \neq 0, m \in M, if rm \in N then m \in N. In particular, if M/N is torsion-free then N is pure.
   c. If M is torsion-free then N is pure iff M/N is torsion-free.
   d. If N is divisible it is pure.
   e. If M is divisible and N is pure then N is divisible.
   f. If N is a summand then it is pure.
   g. If R is a principal ideal domain and M = \mathbb{R}^n for some n then N is pure iff it is a summand.

8. Suppose M is a finitely generated R-module, and N \subseteq M is a submodule. Show that N and M/N are finitely generated, and \rho(M) = \rho(N) + \rho(M/N). Hint: First, \rho(M_1 + M_2) = \rho(M_1) + \rho(M_2). Write M = M_f \oplus M_t where M_f is free and M_t is the torsion submodule. Then N = N_f \oplus N_t where N_f \subseteq M_f and N_t \subseteq M_t. Also, M/N is isomorphic to (M_f/N_f) \oplus (M_t/N_t). Observe that \rho(M_f/N_f) = 0. By the proof of theorem 8.6, M has a basis consisting of two parts S_1 and S_2, where a basis for N_f is obtained as linear combinations of S_1; let M_f = M_1 \oplus M_2 be the corresponding decomposition. M_1/N is torsion (by induction on the construction of S_1, for any x_i \in S_1 some multiple is in N; this can be seen also using Hermite normal form, covered in chapter 23). M_2/N is isomorphic to M_2 (the canonical epimorphism is injective).

9. Show that in an Abelian category, \text{Im}(f) \equiv gK iff gf = 0 and f^CgK = 0 (where the notation of section 2 is used). Hint: Use the fact that \text{Im}(f) = f^{CK}. If gf = 0 then f^{CK} = gK_{j_1} for some j_1; and \text{Im}(f) \equiv gK iff j_1 is an isomorphism. If f^CgK = 0 then gK = f^{CK}j_2 for some j_2.

10. Show that a functor F between Abelian categories is faithful iff it reflects exactness. Hint: Suppose

\[ 0 \longrightarrow F(L) \xrightarrow{F(f)} F(M) \xrightarrow{F(g)} F(N) \longrightarrow 0 \]

is exact. F reflects monics, epics, and 0 objects. Since F(g)F(f) = 0, gf = 0. Since F(g)F(g^K) = 0, F(g^K) = (F(g))^K_{j_1} for some j_1. Since F(f^{CK})F(f) = 0, F(f^{CK}) = j_2(F(f))^{CK} for some j_2. It follows that F(f^{CK}g^K) = 0, whence f^{CK}g^K = 0. By exercise 9 \text{Im}(f) \equiv \text{Ker}(g). For the converse, use the fact that f : M \mapsto N is 0 iff

\[ M \xrightarrow{f} N \]

is exact.

11. Show that a flat right R-module M is faithfully flat iff, for any N \in RMod, if N \neq 0 then M \otimes N \neq 0. Hint: If M is faithfully flat, and M \otimes N = 0, then 0 \mapsto M \otimes N \mapsto 0 is exact, whence 0 \mapsto N \mapsto 0 is. For the converse, suppose

\[ N_1 \xrightarrow{f} N_2 \xrightarrow{g} N_3 \]

is exact after tensoring with M. Letting f' denote the tensor’ed map, the image object of gf' is 0, whence the image object of gf is, whence gf = 0. The homology module after tensoring is 0, so the homology module is 0, so the sequence is exact.
20. Further topics in rings and fields.

1. Algebraic closure. Given a field $F$ let $R$ be the ring of polynomials with coefficients in $F$ and variables from some set of variables. By ignoring a canonical embedding, $F$ may be considered a subring of $R$, namely the constant polynomials. An ideal in $R$ is trivial iff it contains a constant polynomial; if $I \subseteq R$ is a proper ideal, again $F$ may be considered a subring of $R/I$.

Recall that a field $F$ is algebraically closed if any polynomial $p \in F[x]$ has a root (and hence all its roots) in $F$. Also recall that given $p$, there is a field in which $p$ has a root, namely $F/I$ where $I$ is the principal ideal generated by an irreducible factor of $p$ of which $a$ is a root. In view of this latter fact it seems reasonable to suppose that any field $F$ has an algebraic extension which is algebraically closed; such a field is called an algebraic closure of $F$.

**Theorem 1.** Any field $F$ has an algebraic closure.

**Proof:** The idea is that a maximal algebraic extension of $F$ is algebraically closed; but these must be given as subfields of a fixed extension. First we show that there is an extension $F' \supseteq F$ in which every $p \in F[x]$ has a root. This can be proved using transfinite recursion, after well-ordering $F[x]$ (yielding an algebraic extension in fact). One way of avoiding this, attributed to E. Artin, is as follows. For each $p \in F[x]$ introduce the variable $x_p$, and consider the ring $R$ of polynomials over $F$ involving such variables. Let $I$ be the ideal generated by $(p(x_p) : p \in F[x])$. Suppose $I$ is not proper, and that $\sum_i q_ip_i(x_{p_i}) = 1$. Let $E \supseteq F$ be an algebraic extension in which each $p_i$ has a root, say $\alpha_i$. Then in the polynomial ring in the $x_p$ with coefficients in $E$, $1 = \sum_i q_ip_i = 0$ where $p' = p$ with $a_i$ substituted for $x_{p_i}$. This is a contradiction, so $I$ is proper. Let $J$ be a maximal ideal containing $I$. In $R/J$ each $p \in F[x]$ has a root, namely $x_p + J$. Let $F_0 = F$, and for $i > 0$ let $F_i = F_{i-1}$. Let $E$ be the subfield of $\bigcup_i F_i$ of elements which are algebraic over $F$. Then $E$ is an algebraic closure of $F$.

**Theorem 2.** Suppose $L$ is an algebraically closed field, $\sigma : F \hookrightarrow L$ is an embedding, and $E \supseteq F$ is an algebraic extension. Then there is an embedding $\tau : E \hookrightarrow L$ extending $\sigma$.

**Proof:** The pairs $\langle K, \psi \rangle$ where $F \subseteq K \subseteq E$, $\psi$ is a embedding mapping $F$ to $L$, and $\psi$ extends $\sigma$, may be partially ordered, with $(K_1, \psi_1) \leq (K_2, \psi_2)$ if $K_1 \subseteq K_2$ and $\psi_2$ extends $\psi_1$. This partial order is inductive; given a chain $(K_i, \psi_i)$ let $K = \bigcup_i K_i$ and $\psi(x) = \psi_i(x)$ for any $i$ where $x \in K_i$. By Zorn’s lemma there is a maximal element $\langle K, \tau \rangle$. We claim that $K = E$; this follows because if $a \in E - K$, the embedding could be extended to $K[a]$, as observed in chapter 9.

**Corollary 3.** Any two algebraic closures of a field $F$ are isomorphic, by an isomorphism fixing $F$.

**Proof:** Let $K_1$ and $K_2$ be two algebraic closures. By the theorem there is an embedding $\sigma : K_1 \hookrightarrow K_2$ fixing $F$. Further $\sigma[K_1]$ is algebraically closed and $K_2$ is algebraic over it, so $\sigma[K_1] = K_2$.

2. Transcendence bases. Suppose $E \supseteq F$ is an extension of fields. Generalizing earlier notation, if $S \subseteq E$ let $F[S]$ be the ring generated by $F$ and $S$, namely the values of multinomials in elements of $S$ with coefficients in $F$. Let $F(S)$ be the generated field, namely the quotient field of $F[S]$. Note that $F(a, b) = F(a)(b)$. Indeed $F(a, b)$ contains $F(a)$, and so $F(a)(b)$; and $F(a)(b)$ contains $a$ and $b$.

A subset $S \subseteq E$ is said to be algebraically independent if no nontrivial multinomial over $F$ in its elements equals 0. The theory of algebraically independent subsets of $E$ closely parallels the theory of linearly independent subsets of a vector space $V$ over $F$, as described in section 8.7. In particular, they form an inductive family, so there are maximal ones. A maximal algebraically independent subset of $E$ is called a transcendence base.

An algebraically independent set $S$ is maximal iff $E$ is algebraic over $F(S)$. Indeed, if $S$ is maximal, for any element $x \in E$ there is a polynomial with coefficients in $F(S)$ (indeed $F[S]$) of which $x$ is a root; thus,
$E$ is algebraic over $F(S)$. Conversely if $S$ is algebraically independent and $E$ is algebraic over $F(S)$ then $S$ cannot be enlarged, since for any nonzero $x$ there is a nontrivial polynomial over $F(S)$ in $x$ equaling 0; and this yields an algebraic dependence by clearing denominators.

The analog of lemma 8.3 holds, but its proof is slightly more involved. First note that $a$ is algebraic over $F(b)$ iff $\{a, b\}$ is dependent, iff $b$ is algebraic over $F(a)$.

**Lemma 4.** Suppose $E \supseteq F$ is an extension of fields. Let $S \subseteq E$ be an algebraically independent subset of size $k > 0$, and let $T$ be a subset of size $k + 1$. If $S \cup \{t\}$ is algebraically dependent for each $t \in T$ then $T$ is algebraically dependent.

**Proof:** Suppose $\{1, \ldots, k\}$ is the disjoint union of $I_s$ and $I_t$; we claim that every $s_l$ and $t_i$ is algebraic over $F(\{s_1 : i \in I_s\} \cup \{s_l : i \in I_t\})$. The claim is proved by induction on $l = |I_s|$, the basis $l = 0$ following by hypothesis. For the induction step, the claim for $s_l$ for $l + 1$ follows from the claim for $t_i$ for $l$; the claim for $t_j$ for $l + 1$ now follows by transitivity of algebraic extension. In particular $t_{k+1}$ is algebraic over $F(t_1, \ldots, t_k)$, which proves the lemma.

As in section 8.7, all transcendence bases for $E$ over $F$ have the same cardinality, which may be finite or infinite. This cardinality is called the transcendence degree. (In the proof for the infinite case, $S_c$ is a subset of $B$ on which $c$ is dependent).

In the case $|S| = 1$, a proof of Lemma 4 can be given using the resultant. Given polynomials $p_1(t_1, s)$ and $p_2(t_2, s)$, by considering the coefficients to be in $F[t_1, t_2]$, by lemma 7.12 it follows that $\rho(t_1, t_2) = 0$ whenever $p_1(t_1, s) = 0$ and $p_2(t_2, s) = 0$.

If $S$ is algebraically independent over $F$ and $\bar{F}$ is the algebraic closure of $F$ then $S$ is algebraically independent over $\bar{F}$. Indeed, if not then choose $s \in S$; $s$ is algebraic over $\bar{F}(S - \{s\})$, whence over $F(S - \{s\})$.

Transcendence bases find one use in the construction of automorphisms of the field $C$ of complex numbers. To begin with we note the following.

- An automorphism of $R$ or $\mathcal{C}$ fixes pointwise the prime subfield $\mathbb{Q}$.
- In $R$, $x \geq 0$ iff $x = w^2$ for some $w$. It follows that an automorphism of $R$ preserves $\leq$. It then follows that it is the identity, since the set of rationals below a real determines it.
- An automorphism of $\mathcal{C}$ must act on the roots of any rational polynomial; in particular it must map $i$ to $\pm i$.
- An automorphism of $\mathcal{C}$ over $\mathcal{R}$ is either the identity or complex conjugation.

However, other automorphisms of $\mathcal{C}$ do exist. If $S$ is a transcendence base of $\mathcal{C}$ and $\phi$ is an automorphism of $\mathcal{C}$ then $\phi[S]$ is a transcendence base for $\mathcal{C}$. On the other hand, suppose $S_1$ and $S_2$ are transcendence bases, and $\phi_0 : S_1 \mapsto S_2$ is a bijection; we will show that there is an automorphism of $\mathcal{C}$ extending $\phi_0$. Firstly, there is a unique field isomorphism $\phi_1 : \mathbb{Q}(S_1) \mapsto \mathbb{Q}(S_2)$ extending $\phi_0$. Secondly, if a field isomorphism $\phi_2$ extends $\phi_1$, and $x \notin \text{Dom}(\phi_1)$, then there is a field isomorphism $\phi_3$ extending $\phi_2$ with $x \in \text{Dom}(\phi_3)$. By Zorn’s lemma, there is a field isomorphism extending $\phi_0$ from $\mathcal{C}$ to a subfield of $\mathcal{C}$. It must be surjective, since every element of $\mathcal{C}$ is algebraic over $\phi_1(S_2)$.

Letting $\bar{\mathbb{Q}}$ denote the algebraic closure of $\mathbb{Q}$ in $\mathcal{C}$ (the algebraic numbers), an automorphism of $\mathcal{C}$ maps $\bar{\mathbb{Q}}$ to $\mathbb{Q}$. On the other hand, given an automorphism $\psi$ of $\bar{\mathbb{Q}}$, and a bijection $\phi_0$ of transcendence bases as above, there is an automorphism of $\mathcal{C}$ extending both $\phi_0$ and $\psi$. It suffices to observe that there is a unique field isomorphism $\phi_1 : \bar{\mathbb{Q}}(S_1) \mapsto \bar{\mathbb{Q}}(S_2)$ extending $\phi_0$ and $\psi$.

3. **Local rings.** A commutative ring $R$ is said to be local if it has a unique maximal ideal $M$.

**Theorem 5.** Let $U$ be the multiplicative group of units of a commutative ring $R$. The following are equivalent.

a. $R$ is local.
b. \( R - U \) is an ideal.

c. If \( x \in R - U \) and \( y \in R - U \) then \( x + y \in R - U \).

d. If \( x + y \in U \) then \( x \in U \) or \( y \in U \).

e. If \( x + y = 1 \) then \( x \in U \) or \( y \in U \).

If \( R \) is local then \( R - U \) is the maximal ideal.

**Proof:** Suppose \( R \) is local, with maximal ideal \( M \). If \( r \in R - U \) then \( rR \) is a proper ideal, so \( rR \subseteq M \); this shows that \( R - U \subseteq M \), and since \( M \) is proper \( M \subseteq R - U \). On the other hand, if \( R - U \) is an ideal then since any ideal proper ideal is contained in \( R - U \), \( R - U \) is the unique maximal ideal. Thus, a and b are equivalent. Clearly b implies c; conversely if \( xy \) is a unit then \( y \) is a unit, so c implies b. Clearly c and d are equivalent. Clearly d implies e; conversely, supposing e, if \( x + y = u \) where \( u \in U \) then \( xu^{-1} + yu^{-1} = 1 \), and it follows that \( x \in U \) or \( y \in U \).

The quotient \( R/M \) of a local ring by its maximal ideal is a field, called the residue class field.

If \( \phi : R \to S \) is a homomorphism of commutative rings, an ideal \( J \subseteq S \) is said to lie above an ideal \( I \subseteq R \) if \( I = \phi^{-1}[J] \). The terminology “lie over” is also used.

Following is one important example of a local ring; another will be given in theorem 14.

**Theorem 6.** Suppose \( P \) is a prime ideal in a commutative ring \( R \), and \( S = A - P \); then \( A_S \) is local, with maximal ideal \( P_S \), which lies above \( P \).

**Proof:** That \( S \) is multiplicative was observed in section 6.5. Clearly if \( x \in P \) then \( [x/1] \in P_S \); conversely if \( [x/1] \in P_S \) then \( txs = ty \) for some \( t \in S \) and \( y \in P \), whence \( txs \in P \), whence \( x \in P \). This shows that \( P_S \) lies above \( P \); in particular \( P_S \) is proper. An element of \( R_S - P_S \) is a unit in \( R_S \), since if \( a \notin P \) and \( s \in S \) then \( [s/a] \in R_S \). On the other hand since \( P_S \) is proper any element of \( P_S \) is a nonunit of \( R_S \). Thus, the nonunits of \( R_S \) are precisely \( R_S - P_S \).

As noted in section 6.5, under the circumstances of the lemma, the ring \( R_S \) is also denoted \( R_P \). A ring \( R_P \) for a prime ideal \( P \) is called a localization of \( R \) (some authors refer to any \( R_S \) as a localization). Localization has various uses in algebra. For example, given a prime ideal \( P \) in a commutative ring \( R \), a fact about \( P \) may be proved by localizing, reducing the problem to the case where \( P \) has been “transformed” into a maximal ideal; in particular the quotient ring is a field. For an example, see theorem 23 below.

**4. Absolute values.** An absolute value on a field \( K \) is a map \( x \mapsto |x| \) from \( K \) to \( \mathbb{R}^\geq \) such that

1. \( |x| = 0 \iff x = 0 \);
2. \( |xy| = |x||y| \);
3. \( |x + y| \leq |x| + |y| \) (triangle inequality).

The function \( |0| = 0 \) and \( |x| = 1 \) for \( x \neq 0 \) is an absolute value, called the trivial absolute value; any other absolute value is called nontrivial.

The notation \( |x| \) conflicts with one used in chapter 10, where it denoted \( \sqrt{x^*x} \), which is a map from \( K \) to \( K_\geq^\times \). Unless otherwise specified, from hereon an absolute value will be assumed to be a map to \( \mathbb{R} \). Of course, in the cases \( K = \mathbb{R} \) or \( K = \mathbb{C} \), \( \sqrt{x^*x} \) is the standard absolute value.

**Lemma 7.** Suppose \( |x| \) is a map from \( K \) to \( \mathbb{R}^\geq \) such that conditions 1 and 2 above hold. Then

a. \( |1| = |-1| = 1 \);

b. \( |-x| = |x| \); and

c. \( |x^n| = |x|^n \) for all \( n \in \mathbb{Z} \).

**Proof:** For part a, \( |1| = |1^2| = |1| \), so \( |1| = 1 \) since \( |1| \neq 0 \). Also, \( |-1|^2 = |(-1)^2| = |1| = 1 \), so \( |-1| = 1 \) since 1 is the only square root of 1 in \( \mathbb{R}^\geq \). Part b follows from \( |-x| = |-1||x| \). For part c, \( |x^{-1}||x| = 1 \), so \( |x^{-1}| = |x|^{-1} \), and the rest follows by induction.
Lemma 8. Suppose $|x|$ satisfies the restrictions of lemma 7. Then $|x|$ satisfies the triangle inequality iff $|x| \leq 1$ implies $|x + 1| \leq 2$.

Proof: If the triangle inequality holds and $|x| \leq 1$ then $|x + 1| \leq 2$. For the converse, supposing $|x| \leq |y|$, $|x + y| = |y||x/y + 1| \leq 2|y| = 2 \max(|x|, |y|)$. Inductively, $|x_1 + \cdots + x_n| \leq 2^n \max(|x_j|)$, from which it follows that $|x_1 + \cdots + x_n| \leq 2n \max(|x_j|)$. Thus,

$$|x + y|^n = |(x + y)^n| = \left| \sum_{i=0}^{n} \binom{n}{i} x^{i} y^{n-i} \right| \leq 2(n+2) \max \left\{ \binom{n}{i} |x|^i |y|^{n-i} \right\} \leq 4(n+2) \max \left\{ \binom{n}{i} |x|^i |y|^{n-i} \right\} \leq 4(n+2) \sum_{i=0}^{n} \binom{n}{i} |x|^i |y|^{n-i} = 4(n+1)(|x| + |y|)^n.$$

Taking $n$th roots and letting $n \to \infty$ the triangle inequality follows.

Recall that a metric on a set $X$ is a function $d(x, y)$ from $X \times X$ to $\mathbb{R} \geq$ such that $d(x, y) = 0$ iff $x = y$, $d(x, y) = d(y, x)$, and $d(x, z) \leq d(x, y) + d(y, z)$.

Theorem 9. Suppose $|x|$ is an absolute value on a field $K$. The function $|x - y|$ is a metric function on $K$. The functions $x + y$, $-x$, $x \cdot y$, and $1/x$ (on $K^\neq$) are continuous on the resulting metric space. Also, $|x|$ is continuous.

Proof: All requirements for the first claim follow almost immediately. That $+$ is continuous follows using $|(x + y) - (x_0 + y_0)| \leq |x - x_0| + |y - y_0|$ (given $\epsilon$ suppose $|x - x_0| < \epsilon/2$ and $|y - y_0| < \epsilon/2$). That $\cdot$ is continuous follows using $|(xy) - (x_0y_0)| \leq |x||y - y_0| + |y||x - x_0|$ (given $\epsilon$, if $x_0 = 0$ or $y_0 = 0$ suppose $|x|, |y| < \min(1, \epsilon)$; otherwise suppose $|x - x_0| < \epsilon/(2|y_0|)$, $|x - x_0| < |x|$, whence $|x| \leq |x - x_0| + |x_0| < 2|x_0|$, and $|y - y_0| \leq \epsilon/(4|x_0|)$). That $1/x$ is continuous follows using $|1/x - 1/x_0| = |x - x_0|/(|x||x_0|)$ (given $\epsilon$ suppose $|x - x_0| < |x_0|/2$ whence $|x| \geq |x_0|/2$, and $|x - x_0| < \epsilon|x_0|^2/2$). That $|x|$ is continuous follows using $|x| = ||x||$ (the outer absolute value taken in $\mathbb{R}$).

Two absolute values are called equivalent if they induce the same metric topology on $K$.

Lemma 10. Suppose $|x|_1$ and $|x|_2$ are two absolute values on $K$; then the following are equivalent.

a. The two norms are equivalent.

b. Both norms are trivial, or $|x|_1 < 1$ implies $|x|_2 < 1$.

c. There is a constant $\lambda > 0$ such that $|x|_2 = |x|_1^\lambda$.

Proof: Suppose the norms are equivalent, and $|x|_1 < 1$. Then $|x|_1^n$ converges to 0, so $|x|_2^n$ converges to 0, so $|x|_2 < 1$. Suppose $|x|_1 < 1$ implies $|x|_2 < 1$, and suppose $|x|_1$ is not trivial. Choose any $w$ such that $|w|_1 > 1$ (take $w^{-1}$ if $|w|_1 < 1$), and consider any $x$ with $|x|_1 > 1$. Then $|x|_1 = |w|_1^n$ for some real number $r$; we claim that $|x|_2 = |w|_2^n$. This being so, it follows that if $\lambda = \log(|w|_2)/\log(|w|_1)$ then $|x|_2 = |x|_1^n$ for all $x$. To prove the claim, if $r < m/n$ for $m \in \mathbb{Z}$, $n \in \mathbb{N}$, then $|x|_1 < |w|_1^{m/n}$ iff $|x^n/w^m|_1 < 1$ iff $|x^n/w^m|_2 < 1$ iff $|x|_2 < |w|_2^{m/n}$. For $c$ implies $a$, if $|x|_2 = |x|_1^n$, then $\{x : |x|_1 < \epsilon\} = \{x : |x|_2 < \epsilon^\lambda\}$.

In particular the only absolute value equivalent to the trivial one is the trivial one. The topology on $K$ induced by the trivial absolute value is the discrete topology.

An absolute value which satisfies

- $|x + y| \leq \max(|x|, |y|)$ (ultrametric inequality)

is called non-Archimedean; otherwise it is Archimedean.

Lemma 11. Suppose $|x|$ satisfies the restrictions of lemma 7. Then $|x|$ satisfies the ultrametric inequality iff $|x| \leq 1$ implies $|x + 1| \leq 1$. 242
Lemma 12. An absolute value $X$ is non-Archimedean if $|x| \leq 1$ for all $x$ in the prime field of $K$.

Proof: One direction follows by lemma 11. The proof in the other direction proceeds similarly to that of the converse direction of lemma 10; using the binomial theorem and the hypothesis one obtains $|x + y|^n \leq (n + 1)(\max(|x|, |y|))^n$. Taking $n$th roots and letting $n \to \infty$ yields the claim.

Further observations include the following.
- Equivalent absolute values are either both Archimedean or both non-Archimedean.
- The only absolute value on a finite field is the trivial one (if not the values $|x|^n$ are distinct).
- The trivial absolute value on a field $K$ is non-Archimedean, immediately from the definition.
- An absolute value is non-Archimedean if it is non-Archimedean on a subfield (by lemma 12).
- Any absolute value on a field of nonzero characteristic is non-Archimedean.

In addition to the usual absolute value in the rational numbers $\mathbb{Q}$, there are the $p$-adic valuations for each prime $p$. For a nonzero rational $q$ let $o_p(q)$ be the order at $p$, as defined in chapter 6. Let $r$ be any real number with $0 < r < 1$. Then $|q|_p = p^{-o_p(q)}$ is readily verified to be a non-Archimedean absolute value on $\mathbb{Q}$. For definiteness $r$ is often taken to be $1/p$. The usual absolute value is Archimedean. These are, up to equivalence, all the absolute values on $\mathbb{Q}$. We omit a proof; see [Jacobson] for example.

5. Valuations. A non-Archimedean absolute value is also called a real valuation. There is a more general notion of a valuation; its usefulness is illustrated by theorem 15 below.

Consider an ordered commutative group $G$, written multiplicatively; the “positive” elements are those greater than 1. We may adjoin an element 0, satisfying $0 < x$ and $0x = 0$ for all $x \in G$. We call such a structure a value group; $\mathbb{R}_{\geq 0}$ with multiplication is an example. By a subgroup of a value group we mean a substructure containing 0 and 1 and closed under times and inverse of nonzero elements; a subgroup of a value group is a value group.

A valuation on a field $K$ is a map $|x|$ from $K$ to a value group $G$ satisfying $|x| = 0$ iff $x = 0$, $|xy| = |x||y|$, and the ultrametric inequality. The range of a valuation may be a subgroup of $G$; we call this subgroup the value group of the valuation $|x|$. A valuation satisfies lemmas 7 and 11; the proofs are unchanged. Another important property of valuations is the following.

Lemma 13. If $|x|$ is a valuation on $K$ then
a. if $|x| < |y|$ then $|x + y| = |y|$, and
b. if $n \geq 2$ and for some $i$, $|x_i| > |x_j|$ for all $j \neq i$, then $x_1 + \cdots + x_n \neq 0$.

Proof: For part a, $|y| = |x + y - x| \leq \max(|x + y|, |x|)$; thus if $|x| < |y|$ then $|y| \leq |x + y|$. But if $|x| < |y|$ then $|x + y| \leq |y|$, and so $|x + y| = |y|$. Part b follows, since $|\sum_{j \neq i} x_j| < |x_i|$.

A subring $Q \subseteq K$ of a field $K$ is called a valuation ring if for every nonzero $x \in K$, either $x \in Q$ or $x^{-1} \in Q$. Note that if $Q$ is a valuation ring in a field $K$ then $K$ is the field of fractions of $Q$.

Theorem 14. If $Q \subseteq K$ is a valuation ring then $Q$ is a local ring.

Proof: Suppose $x, y$ are nonzero nonunits and without loss of generality $x/y \in Q$; then $(x + y)/y = 1 + x/y$ is in $Q$. If $x + y$ is a unit then $1/y \in Q$, a contradiction. Thus, the nonunits are closed under + and $Q$ is local by lemma 3.1.

Theorem 15. Suppose $K$ is a field. If $|x|$ is a valuation on $K$ then $Q = \{x \in K : |x| \leq 1\}$ is a valuation ring. Conversely if $Q \subseteq K$ is a valuation ring then there is a valuation $|x|$ on $K$ with $Q = \{x \in K : |x| \leq 1\}$.
Proof: If \(|x|, |y| \leq 1\) then \(|x + y| \leq 1, |−x| \leq 1, |xy| \leq 1, |0| \leq 1, |1| \leq 1; thus, \(Q\) is a ring. If \(x \in \mathbb{K}\) and \(|x| > 1\) then \(|x^{-1}| < 1\), so \(Q\) is a valuation ring. For the converse, let \(U\) be the group of units of \(Q\): the nonzero elements of the value group will be the elements of the group \(\mathbb{K}^\times / U\). Let \(P\) consist of the cosets \(xU\) of \(\mathbb{K}^\times / U\) where \(x \notin Q\). If \(xU = yU\) then \(x \in Q\) if \(y \in Q\); using this and the fact that \(Q\) is a valuation ring it follows that exactly one of \(xU \in P\), \(xU = U\), or \(x^{-1}U \in P\) holds. Again since \(Q\) is a valuation group, if \(x, y \notin Q\) then \(xy \notin Q\), so \(xU, yU \in P\) implies \(xyU \in P\). The map \(x \mapsto xU\) maps nonzero elements of \(\mathbb{K}\) to nonzero elements of the value group; certainly \(|xy| = |x||y|\) for nonzero elements and hence for all elements. Finally if \(x \in Q\) then \(1 + x \in Q\), which proves the ultrametric inequality by lemma 11.

A valuation is called trivial as in the real case. A valuation is called discrete if the value group is cyclic (hence infinite cyclic for nontrivial valuations). The cyclic group may be (order preserving) embedded in the reals; discrete valuations are thus absolute values.

6. Integral extensions. As for fields, an extension \(B \supseteq A\) of rings may be considered as a pair. In this section, rings will be assumed to be commutative. If \(B \supseteq A\) is an extension of commutative rings, an element \(b \in B\) is called integral over \(A\) if there is a monic polynomial \(p(x) \in A[x]\) such that \(p(a) = 0\). The notion is of most interest for integral domains, but various facts hold in general. For factorial domains we may require that \(p(x)\) be irreducible over the field of fractions, by Gauss’ lemma.

Theorem 16. Suppose \(B \supseteq A\) is an extension of commutative rings, and \(b \in B\). The following are equivalent.

\begin{enumerate}
  \item \(b\) is integral over \(A\).
  \item \(A[b]\) is a finitely generated \(A\)-module.
  \item \(A[b] \subseteq M\) where \(M\) is a faithful \(A[b]\)-module which is finitely generated as an \(A\)-module.
\end{enumerate}

If \(A\) is an integral domain it suffices in part c that \(M\) be nonzero.

Proof: For \(a \Rightarrow b\), if \(b^n + a_{n-1}b^{n-1} + \cdots + a_0, a_i \in A\), then any element of \(A[b]\) can be written as a linear combination of \(1, \ldots, b^{n-1}\) with coefficients in \(A\), since \(b^n\), and so all powers of \(b\), can. For \(b \Rightarrow c\), let \(M = A[b]\). For \(c \Rightarrow a\), let \(w_1, \ldots, w_n\) generate \(M\) over \(A\). Then \(bw = cw\) where \(C\) is an \(n \times n\) matrix with entries in \(A\). Left multiplying \(C - bI\) by its adjugate yields \(\det(C - bI)w = 0\). By either hypothesis, \(\det(C - bI) = 0\). Thus, \(b\) is a root of the polynomial \(\det(C - xI)\), which has coefficients in \(A\) and leading coefficient \(\pm 1\).

By the proof, if \(b\) is the root of a monic polynomial of degree \(n\) then \(1, b, \ldots, b^{n-1}\) generate \(A[b]\) as an \(A\)-module. Say that the extension \(B \supseteq A\) is integral if every element of \(B\) is integral over \(A\).

Corollary 17. Suppose \(A, B, C, D\) are commutative rings.

\begin{enumerate}
  \item If \(B\) is a finitely generated \(A\)-module then \(B \supseteq A\) is integral.
  \item If \(B \supseteq A\) is integral and \(B\) is finitely generated over \(A\) as a ring then \(B\) is a finitely generated \(A\)-module.
  \item If \(C \supseteq B \supseteq A\) then \(C \supseteq A\) is integral if \(C \supseteq B\) and \(B \supseteq A\) are.
  \item Suppose \(D \supseteq B \supseteq A\), \(D \supseteq C \supseteq A\), and \(B \supseteq A\) is integral; and \(E\) is the join of the subrings \(B, C\) of \(D\). Then \(E \supseteq C\) is integral.
\end{enumerate}

Proof: For part a, let \(M = B\) in the theorem. For part b, first note that if \(B\) is a finitely generated \(A\)-module, say by \(u_1, \ldots, u_m\), and \(C\) is a finitely generated \(B\)-module, say by \(v_1, \ldots, v_n\), then \(C\) is a finitely generated \(A\)-module, by \(u_1v_1, \ldots, u_nv_n\). Now write

\[ B = A[a_1, \ldots, a_k] \supseteq A[a_1, \ldots, a_{k-1}] \supseteq \cdots \supseteq A; \]

each extension is a finite extension of modules, so \(B \supseteq A\) is. For part c, if \(C \supseteq A\) is integral then certainly \(C \supseteq B\) and \(B \supseteq A\) are. For the converse, if \(a \in C\) then \(a^n + c_{n-1}a^{n-1} + \cdots + c_0 = 0\) where \(c_i \in B\); the ring \(B' = A[c_{n-1}, \ldots, c_0]\) is a finitely generated \(A\)-module, so \(B'[a]\) is a finitely generated \(A\)-module; it is
also closed under multiplication by \( a \), so \( a \) is integral over \( A \). For part d, if \( e \in E \) then \( e = \sum_{i=1}^{n} b_i c_i \) where \( b_i \in B \), \( c_i \in C \). \( C[b_1, \ldots, b_n] \) is a finitely generated \( C \)-module since the \( b_i \) are integral over \( A \) and hence over \( C \); it is closed under multiplication by \( e \) and contains \( e \), so \( e \) is integral over \( C \).

Note that in part d, if \( C \supseteq A \) is also integral then \( E \supseteq A \) is integral by part c. Corollary 17 applies a fortiori if \( A, B, C, D \) are fields, in which case modules become vector spaces and integral elements or extensions algebraic; some of these facts have been proved earlier. The join of two subfields is also called their compositum.

A class of field (or ring) extensions with properties c and d of the corollary is called distinguished; thus, algebraic extensions are distinguished. Finite extensions are distinguished; property c was noted in section 9.1, and property d follows because a basis for \( B \) over \( A \) generates \( E \) over \( C \). Separable extensions are distinguished; property c was noted in section 9.4, and for property d, every element of \( B \) is separable over \( C \), and every element of \( E \) lies in \( C[b_1, \ldots, b_t] \) for some \( b_1, \ldots, b_t \in B \), which is separable over \( C \) by property c.

**Corollary 18.** If \( C \supseteq A \) and \( B \) is those elements of \( C \) integral over \( A \) then \( B \) is a subring of \( C \).

**Proof:** If \( a, b \) are integral over \( A \) then \( R[a, b] \) is an integral extension and contains \( a \pm b \) and \( ab \); clearly 0 and 1 are integral.

The ring \( B \) of corollary 18 is called the integral closure of \( A \) in \( C \). A commutative ring \( A \) which equals its integral closure in \( C \) is said to be integrally closed in \( C \). An integral domain \( A \) which equals its integral closure in its field of fractions \( A_{A^\#} \) is said to be integrally closed (or normal by some authors).

**Theorem 19.**

a. If \( A \) is a factorial domain then \( A \) is integrally closed.

b. If \( A \) is a valuation ring in its field of fractions then \( A \) is integrally closed.

**Proof:** Suppose \( A \) is factorial. Suppose \( a/b \in A_{A^\#} \) is integral over \( R \), say \( c_0 + \cdots + (a/b)^n = 0 \). Suppose that \( p | b \), where \( p \) is a prime of \( A \); then \( p | c_i b^{n-i} \) for \( i < n \), so \( p | a^n \), so \( p | a \). It follows that \( a/b \in A \). Suppose \( A \) is a valuation ring in its field of fractions, and \( c_0 + \cdots + x^n = 0 \) for \( x \in A_{A^\#} \). Suppose \( x \not\in A \); multiplying the polynomial by \( x^{-n+1} \) shows that \( x \in A \), a contradiction.

For some examples, \( Z \) is integrally closed. The ring \( Z[\sqrt{-3}] \) is not integrally closed. Its field of fractions is \( Q[\sqrt{-3}] \), and this contains \( \zeta = (-1 + \sqrt{-3})/2 \), which is a root of \( x^2 + x + 1 \). The ring \( Z[\zeta] \) is integrally closed, by exercise 1 and the theorem.

Suppose \( A \) is an integral domain. If \( a \) is integral over \( A \) then clearly it is algebraic over the field of fractions \( A_{A^\#} \). On the other hand, if \( a \) is algebraic over \( A_{A^\#} \), then \( ca \) is integral over \( A \) for some \( c \in A \). Indeed, if \( c_0 + \cdots + c_n a^n = 0 \) where \( c_i \in A \) then \( c_0 c_n^{-1} + \cdots + (c_n a)^n = 0 \), so \( c_n a \) is integral. Also, letting \( A_c \) denote the ring of fractions with denominator a power of \( c \), if \( ca \) is integral over \( A \) then \( a \) is integral over \( A_c \).

One verifies by usual arguments that, if \( B \supseteq A \) is an extension of integral domains, \( S \) is a multiplicative subset of \( A \), \( T \) is a multiplicative subset of \( B \) with \( S \subseteq T \), by ignoring some canonical embeddings one may suppose that \( A \subseteq A_S \), \( B \subseteq B_T \), and \( A_S \subseteq B_T \). In particular the field of fractions \( A^\#_{A^\#} \) may be viewed as a subfield of the field of fractions \( B^\#_{B^\#} \).

If \( B \supseteq A \) is integral then \( B_{B^\#} \supseteq A_{A^\#} \) is algebraic; it suffices to observe that if \( b \in B \) then \( b^{-1} \in (A_{A^\#})[b] \).

**Theorem 20.** Suppose \( B \supseteq A \) is an integral extension of integral domains.

a. If \( A \) is integrally closed then \( B \cap A_{A^\#} = A \).

b. If \( B \subseteq C \) then the integral closure of \( B \) in \( C \) equals the integral closure of \( A \) in \( C \).

c. \( B \) is integrally closed iff \( B \) is the integral closure of \( A \) in \( B_{B^\#} \), and in this case \( B_{B^\#} = A_{A^\#} \).
d. If $F$ is an algebraic extension of $A_{S^2}$ and $C$ is the integral closure of $A$ in $F$ then $F = C_{S^2} = C_{A_{S^2}}$.

**Proof:** For part a, certainly $A \subseteq B \cap (A_{S^2})$, and if $a \in B \cap (A_{S^2})$ then $a \in A$ since $a$ is integral over $A$ and $A$ is integrally closed. For part b, the integral closure of $A$ is trivially included in the integral closure of $B$; if $B$ is integral over $A$ the opposite inclusion also holds. For part c, the first claim is immediate from part b. For the second claim, if $w \in B_{A_{S^2}}$, then $w$ is algebraic over $A_{S^2}$, so $aw$ is integral over $A$ for some $a \in A$, so $aw \in B$, so $w \in B_{A_{S^2}}$. For part d, if $w \in F$ then by hypothesis $w$ is algebraic over $A_{S^2}$, and $w \in C_{A_{S^2}}$ as in the proof of part c.

**Theorem 21.** Suppose $R' \supseteq R$ is an integral extension of commutative rings.

a. if $\phi : R' \rightarrow R''$ is a ring homomorphism then $\phi[R']$ is integral over $\phi[R]$.

b. If $I' \subseteq R'$ is an ideal lying over $I \subseteq R$ then $R'/I'$ is integral over $R/I$.

c. If $S$ is a multiplicative subset of $R$ then $R'_S$ is integral over $R_S$.

d. If $R'$ is a field then $R$ is.

e. If $R$ is a field and $R'$ is an integral domain then $R'$ is a field.

**Proof:** For part a, if $p$ is a monic polynomial over $R$ satisfied by $r \in R'$ if $\phi(p)$ is a monic polynomial over $\phi[R]$ satisfied by $\phi[r]$. Part b follows by part a, with $\phi(r) = r + I'$; the map $r + I \mapsto r + I'$ from $R/I$ to $\phi[R]$ is clearly surjective, and (as noted in the next section) is injective since $I'$ lies above $I$. For part c, if $p$ is a monic polynomial satisfied by $r$, dividing the $i$th coefficient by $s^i$ yields a monic polynomial satisfied by $r/s$. For part d, if $x \in R_{A_{S^2}}$ then $x^{-1} \in R'$, so $x^{-1}$ is integral over $R$, so $\sum a_ix^{-i} = 0$ for some $a_i \in R$ with $a_0 = 1$, from which $x^{-1} \in R$ follows. For part e, if $x \in R_{A_{S^2}}$ then $\sum a_ix^{-i} = 0$ for some $a_i \in R$ with $a_0 = 1$. If $R'$ is an integral domain it can be assumed that $a_0 \neq 0$, and it follows that $x^{-1} \in R'$.

7. **Ring homomorphisms and ideals.** Recall the category of pairs $\langle R, M \rangle$ mentioned in chapter 13, where $R$ is a ring and $M$ an $R$-module; a morphism is a pair $\langle \phi, \psi \rangle$ where $\phi : R \rightarrow S$ is a ring homomorphism, $\psi : M \rightarrow N$ is a group homomorphism, and $\psi(rm) = \phi(r)\psi(m)$ for $r \in R$, $m \in M$. $N$ may be viewed as an $R$-module, with the action $\langle r, n \rangle \mapsto \phi(r)n$; and $\psi$ becomes an $R$-module homomorphism. Immediately, $\psi[M']$ may be considered an $R$-module (of course, unlike $\psi[M]$ it is not necessarily an $S$-module).

If $N'$ is a submodule of $N$ then $\psi^{-1}[N']$ is a submodule of $M$, and not merely a subgroup; this follows because if $\psi(m) \in N'$ for $m \in M$ then $\psi(rm) \in N'$. If $M'$ is a submodule of $M$, let $[\psi[M']]$ denote the submodule of $N$ generated by $\psi[M']$. The maps $M' \mapsto [\psi[M']]$ and $N' \mapsto \psi^{-1}[N']$ between the algebraic closure systems of submodules form a Galois adjunction. Indeed, both are clearly order preserving, and $[\psi[M']] \subseteq N'$ iff $\psi[M'] \subseteq N'$ iff $M' \subseteq \psi^{-1}[N']$. Thus, the left adjoint $M' \mapsto [\psi[M']]$ preserves joins, and the right adjoint $N' \mapsto \psi^{-1}[N']$ preserves meets. These maps are sometimes given names, for example extension and contraction.

Various facts of interest concern additional properties of these maps in various circumstances. In the case $M = R$, $N = S$, and $\psi = \phi$, $I \mapsto [\phi[I]]$ maps ideals (left for noncommutative rings) of $R$ to ideals of $S$, and $J \mapsto \phi^{-1}[J]$ maps ideals of $S$ to ideals of $R$. The systems of ideals have additional properties, whose preservation is of interest. In particular, $I \mapsto [\phi[I]]$ preserves ideal product (in unabbreviated notation, $[\phi[I_1]I_2] = [[\phi[I_1]][\phi[I_2]]]$), as the reader may verify. Also, if $P$ is a prime ideal then $\phi^{-1}[P]$ is prime, as again the reader may verify.

Recall from section 3 that an ideal $J \subseteq S$ is said to lie above an ideal $I \subseteq R$ if $I = \phi^{-1}[J]$. There is such a $J$ iff $[\phi[I]]$ lies above $I$.

Other general observations include the following.

- If $\phi$ is surjective then $[\psi[M']] = \psi[M']$.
- If $\phi$ and $\psi$ are both surjective then $M' \mapsto [\psi[M']]$ is a bijection from the submodules $M' \subseteq M$ which contain $\ker(\psi)$, to the submodules $N' \subseteq N$.

246
- If $M' \rightarrow [\psi[M']]$ is surjective, if $M$ satisfies either the ascending or descending chain condition so does $N$.
- Given $N$ and $\phi$ an action of $R$ on $N$ is determined, namely $\langle r, n \rangle \mapsto \phi(r)n$ (this operation is sometimes called restriction of scalars).
- In particular, given $\phi$, ideals of $S$ become $R$-modules.
- If $\phi$ is inclusion then $\phi^{-1}[J] = J \cap R$.
- If $\phi[I] \subseteq J$ then the map $r + I \mapsto \phi(r) + J$ is a well defined ring homomorphism from $R/I$ to $S/J$.
- If $J$ lies above $I$ the homomorphism $r + I \mapsto r + J$ is injective.

If $\phi : R \rightarrow R/I$ is a canonical epimorphism then $J \mapsto \phi^{-1}[J]$ is an order preserving bijection from the ideals of $R/I$ to the ideals of $R$ which contain $I$. Clearly $J$ is maximal iff $\phi^{-1}[J]$ is. One also verifies that $J$ is prime iff $\phi^{-1}[J]$ is.

By facts already noted, if $\phi : R \rightarrow S$ is a flat homomorphism then changing rings is an exact functor from $\text{Mod}_R$ to $\text{Mod}_S$. In the case where the codomain is $R_S$ where $S$ is a multiplicative subset of $R$ much more can be said. Recall that $R_S \otimes M$ is isomorphic to $M_S$, and that $\phi(r) = [r/1]$ and $\psi(m) = [m/1]$. Also, for a submodule $M' \subseteq M$, $M'_S$ may be identified with $[\psi[M']]$. If $R$ is an integral domain $\phi$ is an embedding, and may be taken as inclusion.

We leave it to the reader to verify the following facts, some of which have already been proved.

- $(\psi^{-1}[N'])_S = N'$; in particular $N' \mapsto \psi^{-1}[N']$ is injective, and $M' \mapsto M'_S$ is surjective.
- In addition to inclusion and join, $M' \mapsto M'_S$ respects pairwise meet and 0 (for meet, observe that if $tn_1s_2 = tn_2s_1$ then both sides are in $M'_S \cap M'_S$).
- $(m + N)/s \mapsto (m/s) + N_S$ yields an $R_S$-module isomorphism from $(M/N)_S$ to $M_S/N_S$.
- If $\text{Ann}(M) \cap S \neq 0$ then $M_S = 0$; the converse holds if $M$ is finitely generated.

In the case of ideals $I \subseteq R$ and $J \subseteq R_S$, we take $I_S$ to denote $\{r/s : r \in I, s \in S\}$, the equivalence relation being on $R \times S$; this equals $\langle \phi[I]\rangle$ rather than merely being isomorphic. The following additional facts hold.

- $I_S$ is proper iff $I \cap S = \emptyset$. Let $\mathcal{I}_D$ denote $\{I \subseteq R : I \cap S = \emptyset\}$, and $\mathcal{J}_P$ the proper ideals of $R_S$; then $I \mapsto I_S$ maps $\mathcal{I}_D$ surjectively to $\mathcal{J}_P$, and $J \mapsto \phi^{-1}[J]$ maps $\mathcal{J}_P$ injectively to $\mathcal{I}_D$.
- $I_S = R_S$ iff $R_S/I_S = 0$ iff $(R/I)_S = 0$.
- $r \in \phi^{-1}[I_S]$ iff $(rs - j)t = 0$ for some $s, t \in S$ and $j \in I$, iff $rs \in I$ for some $s \in S$. Thus, $I = \phi^{-1}[I_S]$ iff $r \in I$ whenever $rs \in I$ for some $s \in S$. In particular if $P \subseteq R$ is a prime ideal and $P \cap S = 0$ then $P = \phi^{-1}[P_S]$.
- If $P$ is prime and $P \cap S = 0$ then $P_S$ is prime. Indeed, if $[r_1/s_1][r_2/s_2] \in P_S$ then $s_1r_2 \in P$ for some $s \in S$, so $r_1 \in P$ or $r_2 \in P$, so $[r_1/s_1] \in P_S$ or $[r_2/s_2] \in P_S$.
- $J \subseteq R_S$ is prime iff $\phi^{-1}[J] \cap S = \emptyset$ and $\phi^{-1}[J]$ is prime.
- Let $\mathcal{I}_D^p$ be the prime ideals of $\mathcal{I}_D$ and $\mathcal{J}_P^p$ the prime ideals of $\mathcal{J}_P$ (i.e., of $R_S$). Then $P \mapsto P_S$ and $Q \mapsto \phi^{-1}[Q]$ are inverse maps between these two sets.

In the case of localization at a prime ideal $P$, The notation $I_P$ is often used for $I_S$ where $S = R - P$, and more generally $M_P$ for an $R$-module $P$. From the foregoing, $(R/P)_P = R_p/P_p$, the residue class field. This field is also the field of fractions of $R/P$; the map $(r + P)/s \mapsto (r + P)/(s + P)$ from $(R/P)_P$ to the field of fractions is readily verified verified to be an isomorphism.

The case where $\phi$ is an integral extension is also of interest.

**Lemma 22.** Suppose $R' \supseteq R$ is an integral extension of commutative rings, where $P \subseteq R$ and $P' \subseteq R'$ are prime ideals with $P'$ lying over $P$. Then $P'$ is maximal iff $P$ is maximal. In particular if $R$ is local with maximal ideal $P$ then the prime ideals of $R'$ lying over $P$ are the maximal ideals of $R'$. 247
THEOREM 23. Suppose \( R' \supseteq R \) is an integral extension of commutative rings, and \( P \subseteq R \) is a prime ideal. Let \( S = R - P \), and let \( \phi \) be the map from \( R' \) to \( R_S \). Then \( Q' \mapsto \phi^{-1}[Q] \) maps the maximal ideals of \( R'_S \) bijectively to the prime ideals \( P' \subseteq R' \) lying over \( P \). In particular, there are such \( P' \), and no inclusion relation holds between any two such.

Proof: By the usual identification with \( \phi[R] \) we can assume that \( R_S \subseteq R'_S \). In particular \( R'_S \) is nontrivial, so has maximal ideals. Suppose \( Q' \) is such, and let \( P' = \phi^{-1}[Q] \). By lemma 22, \( Q' \cap R_S = P_S \), so \( P = \phi^{-1}[P_S] \subseteq \phi^{-1}[P'_S] = P' \). By facts noted above \( Q' \mapsto \phi^{-1}[Q'] \) maps the maximal ideals of \( R'_S \) bijectively to the prime ideals \( P' \subseteq R' \) with \( P'_S = Q' \) and \( P' \cap S = 0 \), or \( P' \cap R \subseteq P \); thus it maps injectively to the \( P' \) lying over \( P \). Given \( P' \) lying over \( P \), \( P'_S \) is a prime ideal of \( R'_S \). Further, \( P'_S \cap R_S = P_S \).

8. The radical of an ideal. In this section \( R \) is a commutative ring. The radical of an ideal \( I \) is defined to be \( \{ x \in R : \exists n \in N (x^n \in I) \} \). We will use the notation \( \text{Rad}(I) \) for this; no confusion will arise with the notation \( \text{Rad}(R) \) for the radical of a ring. If \( x^n \in I \) and \( y^m \in I \) then \( (x+y)^{m+n} \in I \), since every term in the expansion is in \( I \). It follows that \( \text{Rad}(I) \) is an ideal. \( \text{Rad}(0) \), the nilpotent elements, is called the nilradical.

An ideal is called radical if it equals its radical. A commutative ring is said to be reduced if it contains no nonzero nilpotent elements, that is, \( \text{Rad}(0) = 0 \). The following are left to the reader.

- For an ideal \( I \subseteq R \), \( R/I \) is reduced iff \( I \) is radical.
- \( R/\text{Rad}(0) \) is a reduced ring.
- A prime ideal is radical.
- If \( \phi : R \mapsto S \) is a ring homomorphism and \( J \subseteq S \) is a radical ideal then \( \phi^{-1}[J] \) is a radical ideal.
- If \( \phi : R \mapsto R/I \) is a canonical epimorphism and \( J \subseteq S \) is an ideal then \( \phi^{-1}[J] \) is radical iff \( J \) is.

Lemma 24. If \( S \) is a multiplicative subset and \( I \) is maximal among the ideals disjoint from \( S \) then \( S \) is prime.

Proof: If \( x \notin I \) then \( rx - s \in I \) for some \( r \in R \), \( s \in S \). If also \( x' \notin I \) then \( r'x' - s' \in I \) for some \( r', s' \). Then \( xx' \notin I \), else \( ss' \in I \), but \( ss' \in S \).

Theorem 25. Suppose \( I \subseteq R \) an ideal. Then \( \text{Rad}(I) \) equals the intersection of the prime ideals containing \( I \).

Proof: By considering \( R/I \) it suffices to show that \( \text{Rad}(0) \) is the intersection \( S \) of all prime ideals. If \( x^n = 0 \) then \( x \) is contained in any prime ideal since \( x^n \) is; thus, \( \text{Rad}(0) \subseteq S \). If \( x \notin \text{Rad}(0) \) then \( \{ x^i : i \in N \} \) is a multiplicative subset. By lemma 24 there is a prime ideal not containing \( x \).

Theorem 26. \( \text{Rad}(\cap_{1 \leq j \leq r} I_j) = \cap_{1 \leq j \leq r} \text{Rad}(I_j) \).

Proof: \( x \in \text{Rad}(\cap_{1 \leq j \leq r} I_j) \) iff \( \exists n (x^n \in \text{Rad}(\cap_{1 \leq j \leq r} I_j)) \) iff \( \exists n \forall j (x^n \in I_j) \). But this is so iff \( \forall j \exists n (x^n \in I_j) \), iff \( \forall j (x \in \text{Rad}(I_j)) \), iff \( x \in \cap_{1 \leq j \leq r} \text{Rad}(I_j) \).

The nilradical, being the intersection of the prime ideals, is a subset of the radical of the ring. A ring is said to have dimension 0 if every prime ideal is maximal. In this case the nilradical equals the radical. Since 0 is a prime ideal in an integral domain, an integral domain has dimension 0 iff it is a field. In section 9 it will be seen that Artinian rings have dimension 0. An integral domain is said to have dimension 1 if every nonzero prime ideal is maximal. In section 25.4 the general notion of the dimension of a commutative ring will be considered.
Other properties of the radical operation include the following.

- If \( \phi : R \rightarrow S \) is a homomorphism of commutative rings and \( J \subseteq S \) is an ideal then \( \phi^{-1}[\text{Rad}(J)] = \text{Rad}(\phi^{-1}[J]) \). Indeed, \( \phi(x)^n \in J \) for some \( n \) if and only if \( \phi(x^n) \in J \) for some \( n \).
- Also, if \( I \subseteq R \) is an ideal then \( [\phi(\text{Rad}(I))] \subseteq \text{Rad}(\phi[I]) \). Indeed, if \( x^n \in I \) then \( (\phi(x))^n \in \phi[I] \).
- If \( S \subseteq R \) is a multiplicative subset then \( \text{Rad}(I)_S = \text{Rad}(I_S) \). Indeed, both sides equal \( \{ [r/s] : r^n s^2 \in I \text{ for some } s_2 \in S, n \} \) (exercise 3).

9. Primary decomposition. In this section \( R \) is a commutative ring. An ideal \( Q \subseteq R \) is said to be primary if \( Q \) is proper, and in \( R/Q \) a zero divisor is nilpotent; equivalently if \( ax \in Q \) and \( x \notin Q \) then \( a^n \in Q \) for some \( n \). A primary decomposition of an ideal \( I \) in \( R \) is an expression of \( I \) as the intersection of finitely many primary ideals.

The notion of primary decomposition can be defined for submodules of an \( R \)-module \( M \), and the main facts proved with little additional effort from the case of ideals. Say that an element \( a \in R \) is a zero divisor on \( M \) if \( ax = 0 \) for some nonzero \( x \in M \); and is nilpotent on \( M \) if there is a nonnegative integer \( n \) such that \( a^n M = 0 \). A submodule \( N \subseteq M \) is said to be primary if \( N \) is proper, and a zero divisor on \( M/N \) is nilpotent on \( M/N \). Since \( a \) is nilpotent on \( R/I \) as an \( R \)-module if and only if \( a + I \) is a nilpotent element of \( R/I \), the definitions agree for an ideal. A primary decomposition of a submodule \( N \subseteq M \) is an expression of \( N \) as the intersection of finitely many primary submodules.

**Lemma 27.** If \( N \) is a primary submodule of \( M \) then \( \text{Ann}_R(M/N) \) is a primary ideal.

**Proof:** By hypothesis, if \( a \in R, x \in M, ax \in N, \) and \( x \notin N \) then \( a^n M \subseteq N \) for some \( n \). Suppose \( a, y \in R, ayM \subseteq N, \) and \( yM \not\subseteq N \); then there is a \( z \in M \) such that \( yz \notin N \), and since \( a(yz) \in N, a^n M \subseteq N \).

**Lemma 28.** If \( Q \subseteq R \) is a primary ideal then \( \text{Rad}(Q) \) is a prime ideal, and hence the smallest prime ideal containing \( Q \).

**Proof:** If \( xy \in \text{Rad}(Q) \) then \( (xy)^n \in Q \) for some \( n \), whence either \( x^n \in Q \) or for some \( m \) \( y^mn \in Q \); but then either \( x \in \text{Rad}(Q) \) or \( y \in \text{Rad}(Q) \).

If \( Q \) is a primary ideal and \( P = \text{Rad}(Q) \) then \( Q \) is said to be \( P \)-primary. A primary submodule \( N \) is said to be \( P \)-primary when \( \text{Ann}_R(M/N) \) is \( P \)-primary.

Some general facts will be required below, whose proofs are left to the reader. If \( M \) is an \( R \)-module and \( N_0 \) is a submodule the map \( I \mapsto [IN_0] \) from ideals to submodules is the left adjoint of a Galois adjunction. The right adjoint maps \( N \) to \( \{ r : rN_0 \subseteq N \} \); this ideal is commonly denoted \( (N : N_0) \). The annihilator \( \text{Ann}(M) \) is the special case \( (0 : M) \); more generally \( \text{Ann}(M/N) = (N : M) \). The fact that \( N \mapsto (N : N_0) \) preserves ideals will be used below. Finally, if \( \cap_1 \leq j \leq r \) \( I_j \subseteq P \) where the \( I_j \) are ideals and \( P \) is a prime ideal then for some \( j, I_j \subseteq P \); if the intersection equals \( P \) then some \( I_j \) does.

**Lemma 29.** Suppose \( N_i \) is a \( P \)-primary submodule for \( 1 \leq i \leq r \) and \( N = \cap_1 \leq i \leq r N_i \); then \( N \) is \( P \)-primary.

**Proof:** By hypothesis \( \text{Rad}(\text{Ann}(M/N_i)) = P \) for all \( i \), so by theorem 26 and the preceding paragraph \( \text{Rad}(\text{Ann}(M/N)) = P \). Suppose \( ax \in N \) and \( x \notin N \); then for some \( i \) \( ax \in N_i \) and \( x \notin N_i \). Hence for some \( n \) \( a^n \in \text{Ann}(M/N_i) \), so \( a \in P \), so for all \( i \) there is an \( n \) with \( a^n \in \text{Ann}(M/N_i) \), so there is an \( n \) with \( a^n \in \text{Ann}(M/N) \) for all \( i \). Thus, there is an \( n \) with \( a^n \in \text{Ann}(M/N) \).

Say that a primary decomposition \( N = \cap_1 \leq i \leq r N_i \) of a submodule \( N \) is reduced if the ideals \( P_i = \text{Rad}(\text{Ann}(M/N_i)) \) are distinct, and for \( S \subseteq \{ 1, \ldots, r \} \) and \( i \notin S, \cap_{j \in S} N_j \subseteq N_i \). Given any primary decomposition, lemma 29 can be used to ensure the first requirement; the second requirement can then be ensured by successively deleting \( N_i \) which violate it. Thus, if an ideal has a primary decomposition it has a reduced one. Theorem 31 shows that the ideals \( P_i \) do not depend on the particular reduced decomposition.
LEMMA 30. Suppose $N$ is a $P$-primary submodule, and $x \in M$. Then $\text{Rad}((N : x)) = R$ if $x \in N$, else $P$.

PROOF: If $x \in N$ then clearly $(N : x) = R$; suppose $x \notin N$. If $a \in P$ then $a^n \in \text{Ann}(M/N)$ for some $n$, so $a^n \in (N : x)$, so $a \in \text{Rad}((N : x))$. If $a \notin P$ then for any $n$ $a^n \notin \text{Ann}(M/N)$, so $a^n x \notin N$ for any $N$; thus, $a \in \text{Rad}((N : x))$.

THEOREM 31. Suppose $N = N_1 \cap \cdots \cap N_r$ is a reduced primary decomposition, where $N_i$ is $P_i$-primary. Then $P \in \{P_i\}$ iff $P = \text{Rad}((N : x))$ for some $x \in N$.

PROOF: $\text{Rad}((N : x)) = \text{Rad}(\cap_i (N_i : x)) = \cap_i \text{Rad}((N_i : x)) = \cap_i x \notin P_i P_i$. If $\text{Rad}((N : x))$ is prime then it must be $P_i$ for some $i$ with $x \notin P_i$, by remarks above. Given $i$, choose $x \in \cap_{j \neq i} P_j - P_i$; then $\text{Rad}((N : x)) = P_i$.

THEOREM 32. If $M$ is a Noetherian module over a commutative ring, any proper submodule $N$ has a primary decomposition.

PROOF: Say that a submodule $N$ is irreducible if, whenever $N = N_1 \cap N_2$, either $N = N_1$ or $N = N_2$. It follows by a standard argument that every submodule of a Noetherian module $M$ can be written as in intersection of irreducible submodules. Indeed, if $S$ is the set of submodules which can be so written, if $S^e$ is nonempty it contains a maximal element $N_0$. $N_0 = N_1 \cap N_2$ where $N \subset N_1$ and $N \subset N_2$, a contradiction since $N_1 \subset S$ and $N_2 \subset S$. We claim that an irreducible submodule is primary, proving the theorem. Suppose $ax \in N$ for $a \in R, x \in M - N$. Let $S_n = \{m \in M : a^m m \in N\}$; these form an ascending chain of submodules, so for some $n$ $S_m = S_n$ for $m > n$. Let $T = a^n M + N$, and suppose $g \in S_n \cap T$. Then $y = a^n y_1 + y_2$ where $y_2 \in N$, and $a^n y \in N$. Thus, $a^n y_1 \in N$, and by maximality of $S_n$ $a^n w_1 \in N$, so $w \in N$. Thus, $S^n \cap T \subseteq N$. If $S_n \notin M$ then $N \subset T$; since $N \subset S_n$ also, $N$ is not irreducible, a contradiction. Thus $S_n = M$, showing that $N$ is primary.

The theory of primary decomposition is most satisfactory for finitely generated modules over a Noetherian ring, but it is considered more generally. In particular, an element $a \in R$ is said to be locally nilpotent on $M$ if for any $x \in M$ there is a nonnegative integer $n$ such that $a^n x = 0$. For finitely generated modules this is equivalent to nilpotence. In the general case, especially when $R$ is Noetherian, some authors (for example [Lang] or [Matsumura]) define primary submodules using local nilpotence.

10. The Nullstellensatz. Let $F$ be a field, and let $F[x]$ denote the ring of polynomials in the variables $x = (x_1, \ldots, x_n)$. As observed following theorem 7.9, $F[x]$ is a Noetherian ring. It is an $F$-algebra, and the canonical map $F \rightarrow F[x]$ is injective, so that $F$ may be regarded as a subring (the “constant polynomials”). An ideal $I \subseteq F[x]$ is proper if it contains no nonzero constant. If $I$ is a proper ideal then the composition $F \rightarrow F[x] \rightarrow F[x]/I$ of canonical maps is injective, and again $F[x]/I$ is an $F$-algebra containing $F$.

LEMMA 33. In $F[x]$, for $a \in F^n$ let $I_a$ denote the ideal $\{p : p(a) = 0\}$, and let $J_a$ denote the ideal generated by $\{x_1 - a_1, \ldots, x_n - a_n\}$.

a. $\ J_a = I_a$.

b. For an ideal $I \subseteq F[x]$, $F[x]/I$ is isomorphic to $F$ as an $F$-algebra iff $I$ equals $I_a$ for some $a$.

PROOF: For $p \in F[x]$ and $a \in F^n$ let $p(x + a) = q(x)$; then $p(x) = q(x - a)$. Thus $p$ has been written as $b + p'$ for some $b \in F^n$ and $p' \in J_a$. If $p \in J_a$ then clearly $p(a) = 0$. Conversely suppose $p(a) = 0$; then in the expression $p = b + p'$ as above, $b$ must equal 0. The evaluation map $p \mapsto p(a)$ is a homomorphism of $F$-algebras, with kernel $I_a$, so $F[x]/I_a$ is isomorphic to $F$. Suppose $j : F \rightarrow F/I$ is an isomorphism of $F$-algebras. Then $j(a) = a + I$, and it follows that for any $p \in F[x]$ there is an $a \in F$ such that $p - a \in I$. Let $a_i$ be such that $x_i - a_i \in I$; then $J_a \subseteq I$. But $J_a$ is maximal, since adding a polynomial would add a nonzero constant polynomial. Thus, $J_a = I$. 

250
In general, even $F[x]$ can have other maximal ideals; however if $F$ is algebraically closed this is all of them. This result generalizes to $F[x]$, but some proof is required. The following theorem is attributed to Zariski.

**Theorem 34.** Suppose $F$ is a field and $E \supseteq F$ is an extension field, and suppose $E = F[a_1, \ldots, a_n]$ for some $a_1, \ldots, a_n \in E$. Then $E$ is a finite extension of $F$.

**Proof:** The proof is by induction on $n$. If $n = 1$, $a_1^{-1} = p(a_1)$ for some $p \in F[x]$, whence $a_1$ is algebraic over $F$ and $E$ is finite over $F$. If $n > 1$ let $F_1 = F[a_1]$, let $R = F[a_1]$, and for $c \in R$ let $R_c$ be the fractions with denominator a power of $c$. Then $E = F_1[a_1, \ldots, a_n]$, so by induction $E$ is a finite extension of $F_1$. For each $i$ with $2 \leq i \leq n$ there is a $p_i \in F_1[[x]]$ such that $p(a_i) = 0$. Let $c_i$ be the leading coefficient of $p_i$ after clearing denominators; then $a_i$ is integral over $R_{c_i}$. Letting $d = c_2 \cdots c_n$, each $a_i$ is integral over $R_d$. It follows that every element of $E$ is integral over $R_d$, whence that $a_1$ satisfies a polynomial with coefficients in $R$, whence that $a_1$ is algebraic over $F$, and the theorem follows.

**Corollary 35.** If $F$ is algebraically closed and an ideal $I$ in $F[x]$ is maximal iff it is the ideal generated by $\{x_1 - a_1, \ldots, x_n - a_n\}$ for some $a_1, \ldots, a_n \in F$.

**Proof:** One direction is proved above. Let $I$ be maximal, so that $F[x]/I$ is a field. By the theorem, this field is an algebraic extension of $F$, and since $F$ is algebraically closed it equals $F$. Lemma 33 then shows that it has the specified form.

We introduce some notation discussed at greater length in Chapter 25. If $I$ is an ideal in $F[x]$ let $V(I)$ denote the “zero set” of $I$, namely, $\{a \in F^k : p(a) = 0 \text{ for all } p \in I\}$.

**Corollary 36 (Weak Nullstellensatz).** If $F$ is algebraically closed and $I$ is a proper ideal in $F[x]$ then $V(I)$ is nonempty.

**Proof:** Choose a maximal ideal $J$ containing $I$, and suppose it is generated by $\{x_1 - a_1, \ldots, x_n - a_n\}$. Then $a \in V(I)$ where $a = (a_1, \ldots, a_n)$.

**Theorem 37 (Nullstellensatz).** Suppose $F$ is algebraically closed, and $I$ is an ideal in $F[x]$. Suppose $p \in F[x]$ vanishes on $V(I)$. Then $p^m \in I$ for some integer $m$, that is, $p \in \text{Rad}(I)$.

**Proof:** Suppose $p_1, \ldots, p_t$ generate $I$, and let $J$ be the ideal in $F[x, y]$ generated by $p_1, \ldots, p_t, 1 - yp$ (this is called Rabinowitsch’s trick). $J$ is readily verified to have an empty zero set, whence it is the entirety of $F[x, y]$. In particular $1 = q_1p_1 + \cdots + q_tp_t + r(1 - yp)$ for some $q_1, \ldots, q_t, r \in F[x, y]$. Now substitute $1/p$ for $y$ and clear denominators; it follows that for some $m$, $p^m = Q_1p_1 + \cdots + Q_tp_t$ for some $Q_1, \ldots, Q_t \in F[x]$, which completes the proof.

**11. Dedekind domains.** Dedekind domains occur in algebraic number theory, and can be studied abstractly. A Dedekind domain is an integral domain which is integrally closed, Noetherian, and dimension 1 (that is, every nonzero prime ideal is maximal).

Let $A$ be an integral domain, and $K$ the field of fractions. A fractional ideal is defined to be an $A$-module $I \subseteq K$ such that for some $c \in A$, $cI \subseteq A$. The (A-module) meet, and join or sum, of two fractional ideals is a fractional ideal. The product $[IJ]$ of two fractional ideals may be defined as the submodule of $K$ generated by the subset $IJ$; this is also a fractional ideal, as is readily verified.

**Lemma 38.** In a Noetherian ring $A$, every ideal contains a product of prime ideals.

**Proof:** The proof is similar to that of the first fact in theorem 32. Let $S$ be the set of submodules which contain a product of prime ideals. If $S^c$ is nonempty it contains a maximal element $I$. $I$ cannot be prime, so there are $a, b \in A$ with $a, b \notin I$ but $ab \in I$. If $J = aA + I$, $K = bA + I$ then $J$ and $K$ each contain a product of prime ideals. Further $[JK] \subseteq I$, a contradiction.
Lemma 39. If \( A \) is a Dedekind domain with field of fractions \( K \), and \( I \subseteq A \) is a nonzero fractional ideal, then \( I' = \{ c \in K : cI \subseteq A \} \) is the unique fractional ideal such that \([I'] = A\).

Proof: If \( J \) is any fractional ideal such that \([J] = A\), call \( J \) an inverse for \( I \). Clearly \( J \subseteq I' \). Conversely, if \( x \in I' \) then \( xI \subseteq A \), so \( x[I] \subseteq J \), so \( xA \subseteq J \), so \( x \in J \). Thus, an inverse equals \( I' \) if it exists. If \( I \) is an ideal then \( I \subseteq [IJ] \subseteq A \), \( I' \) is a fractional ideal (it is clearly an \( A \)-module, and \( cI' \subseteq A \) for any \( c \in I \)), and \( A \subseteq I' \). If \( I \) is maximal then \([IJ] = I \) or \([IJ'] = A \). If \([IJ'] = I \) then any element of \( I' \) leaves a finitely generated \( A \)-module invariant and hence is integral over \( A \), and since \( I \) is integrally closed \( I' \subseteq A \); thus, to show \([IJ'] = A \) it suffices to show \( A \subseteq I' \). Indeed we claim that for any \( a \in I \) there is a \( b \in A \) such that \( b/a \in I' - A \), or equivalently \( bI' \subseteq aA \) but \( b \notin aA \). To show this it suffices to show that there is an ideal \( H \) such that \( IH \subseteq aA \) but \( H \nsubseteq aA \); for then let \( b \) be any element of \( H - aA \). To find \( H \), let \([P_1 \cdots P_r] \) be a product of prime ideals contained in \( aA \), with \( r \) least. We may assume without loss of generality that \( P_1 = I \), since some \( P_i \) must be contained in \( I \) (because their product is and \( I \) is prime) and each \( P_i \) is maximal. Then letting \( H = [P_2 \cdots P_r] \), \( H \nsubseteq aA \) because \( r \) was least. We have shown that any maximal ideal has an inverse. If some ideal lacks an inverse, choose such an ideal \( I \) such that any larger \( I \) has an inverse; \( I \) is not maximal, so \( I \subseteq M \) for some maximal ideal \( M \), and \([MM'] = A \). We cannot have \([IM'] = I \), else \( M' \subseteq A \), so, letting \( J = [IM'] \), \( I \subseteq J \subseteq A \), and \([JJ'] = A \). We claim \([M'J'] = A \) is an inverse for \( I \), a contradiction; indeed, \([IM'J'] = [JJ'] = A \). We have shown that every ideal has an inverse. If \( I \) is a fractional ideal, and \( cI \subseteq A \), then \( cI \) has an inverse, and \( cJ \) is an inverse for \( I \).

An ideal \( I \) in a commutative ring \( A \) is said to be invertible if there is a fractional ideal such that \([I'] = A \). Thus, in a Dedekind domain any ideal is invertible. In a commutative ring an ideal \( I \) is said to divide an ideal \( J \) iff there is an ideal \( K \) such that \([IK] = J \). If this is so then \( I \supseteq J \). If \( I \) is invertible, in particular in a Dedekind domain, the converse holds; indeed, if \([I'] = A \) then \([II'] = J \), and since \( I \supseteq J \), \( A = [II'] \supseteq [IJ] \).

It also follows that the fractional ideals of a Dedekind domain form a group under multiplication of fractional ideals. Further the homomorphism \( a \mapsto aA \) of the multiplicative monoids extends to a homomorphism of multiplicative groups on the nonzero elements of \( A_{\neq} \).

Theorem 40. Any ideal in a Dedekind domain \( A \) equals a product of prime ideals in an essentially unique way.

Proof: Let \( S \) be the set of nonzero ideals \( I \subseteq A \) which have a factorization as a product of prime ideals. If some nonzero ideal is not in \( S \), let \( I \) be a maximal such. Thus, \( I \subseteq P \) for some prime ideal \( P \). Let \( P' \) be the inverse fractional ideal to \( P \). Then \( I \subseteq [IP'] \subseteq A \), so \([IP'] \) has a factorization into prime ideals, so \([[IP']P] = I \) does. Thus, every nonzero ideal has some factorization as a product of prime ideals. If \( P, Q, \) and \( R \) are prime ideals and \([QR] \subseteq P Q \subseteq P \) or \( R \subseteq P \); since \( A \) is dimension 1 equality must hold. The theorem now follows, as in the proof of theorem 6.5.c.

If \( I \) is a fractional ideal, it may be written in an essentially unique way as

\[
\frac{[P_1^{r_1} \cdots P_r^{r_r}]}{[Q_1^{t_1} \cdots Q_s^{t_s}]}.
\]

where the \( P_i, Q_j \) are distinct prime ideals. This may be seen by choosing \( c \) so that \( cI \subseteq A \), factoring \( cI \) and \( c \), and canceling common factors; the result does not depend on \( c \). The fractional ideal \( I \) is said to be of order \( e_i \) at \( P_i \), or of order \( -f_j \) at \( Q_j \). The order is a homomorphism from the multiplicative group of fractional ideals to the additive group of the integers. A homomorphism is induced on \( A_{\neq} \), by the above map \( x \mapsto xA \). For another consequence of unique factorization of ideals, for a fractional ideal \( I \), \( I^n \subset I \).
The property of the theorem in fact characterizes Dedekind domains. Also, an integral domain is Dedekind iff every ideal is invertible. We will not prove these facts; for proofs, and many other facts about Dedekind domains, see [Jacobson].

By Theorem 6.5, a principal ideal domain is Noetherian and factorial. By theorem 19 it is integrally closed. By theorem 6.5, a nonzero prime ideal in a principal ideal domain is one generated by a prime element; since for two non-associated primes, neither divides the other, a principal ideal domain is dimension 1. Thus, a principal ideal domain is Dedekind.

**Theorem 41.** If a Dedekind domain is factorial it is a principal ideal domain.

**Proof:** Suppose $A$ is a Dedekind domain, and suppose there is a nonprincipal ideal; since the product of principal ideals is principal there must be a nonprincipal prime ideal, say $P$. If $P'$ is the inverse to $P$ in the fractional ideal group, and $cP'$ is an ideal $I$, then $[PI] = cA$; thus there is an ideal $K$ which is maximal among the ideals $I$ such that $[PI]$ is principal. Let $p \in A$ be such that $[PK] = pA$. Then $p$ is irreducible; for if $p = ab$ then since $P$ is prime either $bA = [PI]$ for some $I$ or $cA = [PI]$ for some $I$; in either case, by maximality of $K$, $I = K$. Since $P$ and $K$ are both proper neither $[PK] = P$ nor $[PK] = K$ holds, whence there are elements $a \in P - pA$ and $b \in K - pA$. Then $p|ab$ but $p \not| a$ and $p \not| b$.

We give an example of the use of invertible ideals, referring to the literature for proofs. If $R$ is an integral domain then an ideal is invertible iff it is projective ([Jacobson], Proposition 10.3). The ideal generated by $\{2, 1 + \sqrt{5}\}$ in $\mathbb{Z}[\sqrt{5}]$ is invertible but not principal ([Eisenbud], section 11.3). It is not free, since if so it would have to have to be principal, by considering dimensions.

**12. Complete fields.** If $K$ is a field equipped with an absolute value, $K$ is called complete if the metric space is complete. In section 17.8, we proved that $\mathcal{R}$ with the usual absolute value is complete. If $F$ is any field with an absolute value, the standard construction of the completion $\bar{K}$ of the metric space of $F$ can be expanded, to yield $\bar{K}$ as an extension field of $F$, equipped with an extension of the absolute value.

To accomplish this, define addition and scalar multiplication componentwise on the Cauchy sequences, yielding a commutative ring $R$; the embedding of $F$ via the constant sequences is a ring homomorphism. The Cauchy sequences which converge to 0 are an ideal $N$ (called, among other things, the null ideal) in $R$. The completion has the structure of the quotient ring $R/N$ defined on it. The quotient is in fact a field ($N$ is a maximal ideal). Indeed, suppose $\langle a_i \rangle$ is a Cauchy sequence which is not in $N$. Then there is an $r > 0$ and an $n_0$ such that $|a_n| \geq r$ for $n \geq n_0$. Let $b_n = 1$ if $n < n_0$, and $b_n = a_n^{-1}$ if $n \geq n_0$. Clearly $\langle b_i \rangle$ is a Cauchy sequence which represents the inverse of $\langle a_i \rangle$ in $R/N$.

The embedding of $F$ remains a ring homomorphism, so $R/N$ is an extension field. The norm of a Cauchy sequence $\langle a_i \rangle$ is defined as the limit, as the limit of the Cauchy sequence $\langle |a_i| \rangle$. On $R$ this is in fact a pseudo-norm (or semi-norm), meaning that $|x| = 0 \Rightarrow x = 0$ may fail to hold; however the equivalence relation it determines is identical to that determined by $N$, so the pseudo-norm induced on $R/N$ is a norm.

The field structure is uniquely determined. By the proof of theorem 9, $+$ is in fact uniformly continuous, and hence determined on $K$ since $F$ is dense in $K$. Similarly $\cdot$ is determined on the elements whose norm is bounded by any fixed bound, hence everywhere.

Ostrowski's theorem states that the only fields complete with respect to an Archimedean absolute value are $\mathcal{R}$ and $\mathcal{C}$, with their usual absolute values. A proof may be found in [Jacobson].

The completion of $Q$ with respect to the $p$-adic valuation defined in section 4 may be defined for each prime $p$. These fields were first studied by Hensel starting in the late nineteenth century, and have since become an important tool in various branches of mathematics. We will give a brief introduction; among many texts containing further discussion are [Ebbinghaus], [Robert], and [Schikhof].

The metric $d(x, y) = |x - y|$ arising from a non-Archimedean absolute value satisfies the ultrametric inequality $d(x, z) \leq \max(d(x, y), d(y, z))$. This has as a consequence the "isosceles triangle property", that
at least two sides of a triangle have the same length, and the length of the third is no greater. Indeed, if \( d(x, z) \) is greatest then by the ultrametric inequality \( d(x, z) = \max(d(x, y), d(y, z)) \). If \( x_n \) is a sequence converging to \( x \) in an ultrametric space, and \( y \neq x \), let \( N \) be such that \( n \geq N \) then \( d(x_n, x) < d(y, x) \); then for \( n \geq N \), \( d(x_n, y) = d(x_n, x) \).

If \( |x| \) is a non-Archimedean absolute value on a field \( F \), recall that \( Q = \{ x \in F : |x| \leq 1 \} \) is a ring, called the valuation ring. It is a local ring, with maximal ideal \( M = \{ x \in F : |x| < 1 \} \) (the units are clearly \( \{ x \in F : |x| = 1 \} \)). The field \( Q/M \) is called as usual the residue class field.

Let \( \hat{F} \) denote the completion of \( F \) with respect to \( |x| \). Let \( |x| \) denote the norm on \( \hat{F} \) as well; as in remarks preceding theorem 17.34, it is non-Archimedean. Let \( \hat{Q} \) denote the closure of \( Q \) in \( \hat{K} \), and similarly for \( \hat{M} \). The following facts may be verified.

- \( \hat{Q} = \{ x \in \hat{K} : |x| \leq 1 \} \). Clearly if \( x \in \hat{Q} \) then \( |x| \leq 1 \). If \( |x| < 1 \) any Cauchy sequence \( w_n \) for \( x \) eventually has \( |w_n| < 1 \). If \( |x| = 1 \), replacing \( w_n \) by \( w_n^{-1} \) where necessary yields a Cauchy sequence with \( |w_n| < 1 \).
- \( \hat{M} = \{ x \in \hat{K} : |x| < 1 \} \).
- The canonical map \( Q \mapsto \hat{Q}/\hat{M} \) is surjective. Indeed, given \( x \in \hat{Q} \), let \( q \) be such that \( |x - q| < 1 \).
- Identifying \( K \) with the constant Cauchy sequences as usual, \( |x| \) on \( K \) is the restriction of \( |x| \) on \( \hat{K} \). In particular, \( Q = \hat{Q} \cap K \) and \( M = \hat{M} \cap K \).
- There is an induced map \( Q/M \mapsto \hat{Q}/\hat{M} \), and it is an isomorphism. Thus, the residue class fields are isomorphic.
- The value groups are also isomorphic, because by the isosceles triangle property, for any \( x \in \hat{K} \) if \( w \in K \) and \( |w - x| < |x| \) then \( |x| = |w| \).

If \( |x| \) is a discrete valuation, with \( r < 1 \) generating the value group, let \( \pi \in F \) be such that \( |\pi| = r \). It is readily verified that every nonzero element of \( F \) can be written uniquely as \( u\pi^n \) for a unit \( u \) and an integer \( n \). It follows that \( |u\pi^n| = r^n \), \( M = \pi Q \), and \( F \) equals \( Q_S \) where \( S = \{ \pi^n : n > 0 \} \). A nonzero ideal in \( Q \) contains \( \pi_n \) for some \( n \geq 0 \), and equals \( \pi M \) for the least such \( n \). Thus, \( Q \) is a principal ideal domain. A local principal ideal domain is called a discrete valuation ring; note that it is factorial.

**Theorem 42.** Every discrete valuation ring \( Q \) is a valuation ring in \( Q_{Q^x} \), where the valuation is discrete.

**Proof:** Let \( M \) be the maximal ideal, with \( M = \pi Q \). Since \( M \) is the only prime ideal, up to a unit \( \pi \) is the only prime, and since \( Q \) is factorial, \( Q \) equals \( \{ u\pi^n \} \) where \( u \) is a unit and \( n \in \mathbb{Z}^\geq \). The fractions are readily seen to be \( \{ u\pi^n \} \) where \( n \in \mathbb{Z} \). The valuation may be taken as \( r^n \) where \( r \) is a real with \( 0 < r < 1 \).

There are various other characterizations of discrete valuation rings; see [AtiMac], Proposition 9.2. For example a local Noetherian domain of dimension 1 is a discrete valuation ring iff its maximal ideal is principal.

In the case of a discrete valuation \( |x| \), one verifies by arguments as above that

- \( \hat{\pi^n}Q = \pi^n \hat{Q} = \{ x : |x| \leq r^n \} \),
- \( \pi^n Q = \pi^n \hat{Q} \cap K \),
- \( Q \mapsto \hat{Q}/\pi^n \hat{Q} \) is surjective, and
- \( Q/\pi^n Q \mapsto \hat{Q}/\pi^n \hat{Q} \) is an isomorphism.

Since \( \pi^n Q \subseteq \pi^{n+1} Q \) there is a homomorphism \( Q/\pi^{n+1} Q \mapsto Q/\pi^n Q \). The system of these forms an inverse system. The inverse limit is an instance of a general construction, the completion of a ring by a filtration; see [AtiMac] for example. The inverse limit may be taken in either the category of rings, or \( Q \)-modules, with the same result, namely the consistent sequences of theorem 13.5.

There is a map \( \hat{Q} \mapsto Q/\pi^n Q \), mapping \( x \) to \( q + \pi^n Q \) where \( |x - q| \leq r^n \). We claim that the induced map from \( \hat{Q} \) to the inverse limit is an isomorphism. Given a consistent sequence \( \langle q_n + \pi^n Q \rangle \), the sequence \( \langle q_n \rangle \) satisfies \( |q_n - q_{n+1}| \leq r^n \). It follows by the ultrametric property that it is a Cauchy sequence. Further,

254
it maps to the given sequence; thus, the map is surjective. It is also injective; if \( x_1 \) and \( x_2 \) map to the same consistent sequence then their difference satisfies \( |x_1 - x_2| < r^n \) for all \( n \).

Let \( S \) be a system of coset representatives for \( Q/\pi Q \) in \( Q \). Every \( x \in \hat{Q} \) can be written uniquely as \( a_0 + \pi x_1 \) where \( a_0 \in S \); continuing, there is a unique expression \( x = \sum a_i \pi^i \) with \( a_i \in S \). The series is finite iff \( x \in \hat{Q} \). For any \( x \in K \) there is an \( m \geq 0 \) with \( \pi^m x \in \hat{Q} \), and elements of \( K \) may be written as series \( x = \sum_{n \geq m} a_i \pi^i \). The \( a_i \) may be considered as elements of the field \( Q/\pi Q \).

In the case of the \( p \)-adic valuation, the residue class field equals the finite field \( F_p \) (see section 9.6). The coefficients \( a_i \) may be taken as \( 0, \ldots, p-1 \), and \( \pi \) as \( p \). The series is finite iff \( x \) is an ordinary integer, and the series is its \( p \)-adic notation. We let \( Z_{\text{pad}} \) denote the ring of series starting at 0, and \( Q_{\text{pad}} \) the field of series starting at some integer. \( Q_{\text{pad}} \), the field of \( p \)-adic numbers, is the completion of \( Q \) with respect to the \( p \)-adic valuation. \( Z_{\text{pad}} \), the ring of \( p \)-adic integers, is its valuation ring. Rules for adding and multiplying the series may be given, involving “carrying”, but we omit this.

For convenience we have used the notation \( Z_n \) for the integers mod \( n \) in this book. A common alternative is to use \( Z_p \) for the \( p \)-adic integers, \( Q_p \) for the \( p \)-adic numbers, and \( Z/nZ \) for the integers mod \( n \). We will use the notation \( Q_{\text{pad}} \) and \( Z_{\text{pad}} \) only infrequently.

13. Real algebras. As mentioned in chapter 6, the associative law for multiplication may be dropped from the requirements for a ring; the requirement for a multiplicative identity may also be dropped. Thus, multiplication need only be left and right distributive. We will call such structures generalized rings. They are also called nonassociative rings, or simply rings; the former terminology is somewhat strained in that nonassociative rings might be associative. Other terminology (for example “rng”) is sometimes adopted.

A generalized algebra over a field \( F \) is a vector space together with a bilinear form; this is a special type of generalized ring. An introduction to this topic may be found in [Jacobson]. An important example is Lie algebras, which will be considered in chapter 27.

In this section some topics concerning generalized algebras over \( R \) will be covered, in particular the hypercomplex, which are associative. The complex numbers are a two-dimensional real algebra. Interest in “hypercomplex” numbers, which are higher dimensional algebras over \( R \) with properties in common with the complex numbers, goes back to the 19th century; see [Ebbinghaus] for a discussion.

An important tool in the study of real algebras is the notion of a quadratic algebra. Such may be defined over any field \( F \), to be a generalized algebra with unit 1, where every \( x \) satisfies an equation \( x^2 + bx + c1 = 0 \) where \( b, c \in F \). The following lemma is known as Frobenius’ lemma.

**Lemma 43.** Suppose \( F \) has characteristic other than 2, and \( A \) is a quadratic algebra over \( F \). Let \( I \) be 0, together with those \( x \notin F1 \) such that \( x^2 \in F1 \). Then \( I \) is a subspace, and \( A = F1 \oplus I \).

**Proof:** Suppose \( x, y \in I \). If \( y = ax \) for some \( a \in F \) then \( x + y = (1 + a)x \) is in \( I \), so suppose \( x \) and \( y \) are linearly independent. We first note that no nontrivial linear combination if \( x \) and \( y \) is in \( F1 \); for if \( y = a1 + bx \) squaring yields that \( 2abx \in F \), so \( 2ab = 0 \), so \( a = 0 \), and a contradiction is obtained from either \( a = 0 \) or \( b = 0 \). From the equations \( (x + y)^2 + b1(x + y) + c11 = 0 \) and \( (x - y)^2 + b2(x - y) + c21 = 0 \) we conclude that \( (b1 + b2)x + (b1 - b2)y \in F \). Hence \( b1 + b2 = 0 \) and \( b1 - b2 = 0 \), and \( b1 = 0 \), so \( (x + y)^2 \in F1 \); and we have already seen that \( x + y \notin F1 \). Suppose \( x \) is an element of \( A - F1 \). Then \( x^2 + bx + c1 = 0 \) for some \( b, c \in F \), and so \( (x - (b/2))1^2 = (b^2 - c)1 \). Since \( x - (b/2)1 \notin F1 \), \( x - (b/2)1 \in I \). Thus, \( x \in F1 + I \). This shows that \( A = F1 + I \), and since \( F1 \cap I = \{0\} \), \( A = F1 \oplus I \).

Henceforth we will assume that for quadratic algebras, the characteristic of \( F \) is not 2. The elements of \( F1 \) are analogous to the real elements, and those of \( I \) to the purely imaginary elements, of the complex numbers. The projection onto \( F1 \) induces a linear functional \( \rho : A \mapsto F \); \( \rho(x)1 \) is analogous to the “real part”. The map \( \iota(x) = x - \rho(x)1 \) is an \( F \)-linear map on \( A \), and maps an element to its “imaginary part”. An
involution in an algebra $A$ over a field $F$ is an $F$-linear map $x \mapsto \bar{x}$ such that $\bar{x} = x$. A quadratic algebra may be equipped with the involution $\bar{x} = \rho(x)1 + i(x)$, analogous to conjugation; we will call this the standard involution.

Squaring $x = \rho(x)1 + i(x)$ and writing $i(x)^2 = \theta(x)1$ yields $x^2 = 2\rho(x)(\rho(x)1 + i(x)) - \rho(x)^21 = 2\rho(x)x + (\theta(x) - \rho(x)^2)1$. Applying $\rho$, $\rho(x^2) = 2\rho(x)^2 + \theta(x) - \rho(x)^2$, whence $\theta(x) = \rho(x^2) - \rho(x)^2$. Thus, $x^2 = 2\rho(x)x + (2\rho(x)^2 - \rho(x^2))1 = 0$, giving explicitly a quadratic equation satisfied by $x$. Also, $\bar{x} = x\bar{x} = \rho(x)^21 - i(x)^2 = (2\rho(x)^2 - \rho(x^2))1$.

Suppose the characteristic of $F$ is not $2$. Recall from section 10.7 that a quadratic form on a vector space $V$ may be defined to be a function $q : V \mapsto F$ such that for some symmetric bilinear form $b(x,y)$ on $V$, $q(x) = b(x,x)$. Exercise 4 shows that $q$ is a quadratic form iff $q(sx) = s^2q(x)$ for $s \in F$, and $f(x + y) - f(x) - f(y)$ is bilinear. One verifies that if $\rho(x)$ is a linear form on an $F$-algebra, then $\rho(x)^2$ and $\rho(x^2)$ are quadratic forms, and also $2\rho(x)^2 - \rho(x^2)$. In a quadratic algebra, the quadratic form $2\rho(x)^2 - \rho(x^2)$, analogous to the “norm squared”, will be called the standard quadratic form.

The “Cayley-Dickson” or “doubling” procedure generalizes the way in which the complex numbers are obtained from the real numbers. Suppose $A$ is an algebra over a field $F$, equipped with an involution $x \mapsto \bar{x}$. Define a multiplication on the vector space $A \times A$ by the formula

$$(x_1, x_2)(y_1, y_2) = (x_1y_1 - \bar{y}_2x_2, x_2\bar{y}_1 + y_2x_1)$$

(there are variations and generalizations of this law).

**Theorem 44.** For an $F$-algebra $A$ equipped with an involution $x \mapsto \bar{x}$, let $B$ be the doubled algebra, as above. Then $B$ is an algebra. The map $(x,y) \mapsto (\bar{x},-y)$ is an involution on $B$. The elements $(x,0)$ comprise a subalgebra isomorphic to $A$. If $A$ has unit 1 then $(1,0)$ is the unit of $B$. The involution on the subalgebra is isomorphic to the given one in $A$. If $A$ is a quadratic algebra with the standard involution then $B$ is a quadratic algebra, and the above involution is the standard one.

**Proof:** The bilinearity of the product is a routine exercise left to the reader; and also the proof that the map is an involution. The statements concerning the embedding of $A$ are also straightforward. Noting that $x + \bar{x} = 2\rho(x)1$, the square in $B$ of $(x_1, x_2)$ equals $(2\rho(x_1)x_1 + a1, 2\rho(x_1)x_2)$ where $a \in F$, proving that $B$ is a quadratic algebra. Further the square is in $F(1,0)$ iff $\rho(x_1) = 0$, and the last claim follows.

The double of the real numbers with the identity involution is the complex numbers with the usual conjugation; indeed the latter are frequently defined in this manner.

**Lemma 45.** Suppose $A$ is a generalized real algebra such that

1. $A$ has a unit and no zero divisors, and
2. $A$ is two dimensional.

Then $A$ is isomorphic to $\mathcal{C}$.

**Proof:** We may identify $r \in \mathcal{R}$ with $r1$ where 1 is the unit of $A$. Suppose $u \in A - \mathcal{R}$; then $u^2 + bu + c = 0$ for some $b, c \in \mathcal{R}$. If $v = u + (b/2)$ then $v^2 = s$ where $s \in \mathcal{R}$. If $s$ is nonnegative then, since $A$ has no zero divisors, $v = \pm \sqrt{s}$, a contradiction; thus $s$ is negative. Let $i = v/\sqrt{-s}$. Then 1, $i$ is a basis, and the structure constants are those of $\mathcal{C}$.

For the next theorem we note a consequence of the fundamental theorem of algebra, proved in section 7.1. Suppose $p$ is a polynomial with real coefficients. Letting $\bar{x}$ denote the complex conjugate, if $p(a) = 0$ for a complex number $a$ then $p(\bar{a}) = \bar{p}(a) = 0$. If $a$ is not real the polynomial $x^2 - (a + \bar{a})x + (a\bar{a})$ has real coefficients and divides $p(x)$. By the Euclidean property of $\mathcal{R}[x]$, the quotient is a real polynomial. Thus, any real polynomial $p$ may be written as a product of factors, each of which is a linear, or quadratic polynomial with real coefficients and complex conjugate roots. Polynomials of the latter type are irreducible over the reals, since a factor would have to be $x - a$ for a root $a$. 

256
Theorem 46. Suppose $A$ is a real algebra such that
1. $A$ is an integral domain, and
2. every element of $A$ is algebraic over $R$.
If $A$ is a proper extension of $R$ then $A$ is isomorphic to $C$.

Proof: Suppose $u \in A - R$; then $p(u) = 0$ for some $p \in R[x]$. Since $u \notin R$ and $A$ has no zero divisors, by remarks above $q(u) = 0$ for some quadratic $q \in R[x]$. Let $B = R + Ru$; by the existence of $q B$ is a two dimensional subalgebra of $A$. By lemma 45 there is an element $i \in B$ making $B$ a copy of $C$ with basis $1, i$. Suppose $v \in A$; again $p(u) = 0$ for some $p \in R[x]$. By the fundamental theorem of algebra $p$ splits into factors $x - b_i$ where $b_i \in B$. Further, $u$ must satisfy one of these factors. Thus, $A = B$.

The double of the complex numbers is a four dimensional real algebra, known as the quaternions, and denoted $H$. Any finite dimensional real algebra satisfies property 2 of the theorem (the powers of any element must be linearly dependent), so $H$ must fail to be an integral domain.

The map
$$\langle x_1, x_2 \rangle \mapsto \begin{bmatrix} x_1 & x_2 \\ -\bar{x}_2 & \bar{x}_1 \end{bmatrix}$$
from $H$ to the $2 \times 2$ matrices over $C$ is readily verified to be injective and preserve multiplication; let $H_M$ denote the image of $H$ under this map. An element of $H_M$ is singular iff it is 0, by considering the determinant. Thus, $H$ is an associative algebra without zero divisors.

A generalized algebra is said to be a division algebra if every equation $z = xy$ and $z = yx$ with $y \neq 0$ has a solution $x$. Clearly such an algebra has no zero divisors; if it is finite dimensional the converse is true. Indeed, the map $x \mapsto xy$ is injective, so bijective. In particular $H$ is a division algebra, indeed a skew field ($H$ is not commutative; this will follow directly from the next characterization).

Let
$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad j = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad k = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$These matrices clearly form a basis for $H_M$. It is readily verified that $i^2 = j^2 = k^2 = -1$, $ij = k$, $jk = i$, $ki = j$, $ji = -k$, $kj = i$, and $ik = -j$. This gives $H$ via structure constants for the basis $1, i, j, k$. The standard involution on $H$ maps $a_0 + a_1i + a_2j + a_3k$ to $a_0 - a_1i - a_2j - a_3k$. The standard quadratic form is $a_0^2 + a_1^2 + a_2^2 + a_3^2$. The square root of this is the Euclidean norm on $R^4$ in this basis.

One verifies readily that the norm squared of a quaternion equals the determinant of its matrix in $R_M$.

Recalling from section 10.12 that the adjoint $M^\dagger$ of a square matrix $M$ equals $\bar{M}^t$, one also verifies that the matrix of the conjugate quaternion is the adjoint to the matrix of the quaternion; and for $M \in H_M$, $MM^\dagger = M^\dagger M = \det(M)1$ where 1 is the identity matrix.

Let $U(2)$ denote the $2 \times 2$ complex matrices $M$ such that $MM^\dagger = M^\dagger M = 1$. These form a group (hence a real algebra), called the unitary group, and for $M \in U(n) \det(M) = \pm 1$. Let $SU(2)$ denote those $M \in U(2)$ with det$(M) = 1$. $SU(2)$ is a subgroup of $U(2)$, called the special unitary group. By the preceding paragraph, if $M \in H_M$ and det$(M) = 1$ then $M \in SU(2)$.

Conversely, for $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in $SU(2),$ $M^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ and $M^\dagger = \begin{bmatrix} \bar{a} & \bar{\bar{c}} \\ \bar{b} & \bar{d} \end{bmatrix},$

and $M \in H_M$ follows. That is, $SU(2)$ is isomorphic to the quaternions of norm 1.

Let $O(n)$ denote the $n \times n$ real matrices $M$ such that $MM^t = M^tM = 1$. These form a group, called the orthogonal group, and for $M \in O(n) \det(M) = \pm 1$. Exercise 5 shows that $M \in O(n)$ iff $M$ preserves
the Euclidean norm, if it preserves the inner product. Let $SO(n)$ denote those $M \in O(n)$ with $\det(M) = 1$. $SO(n)$ is a subgroup of $O(n)$, called the rotation group. Further facts concerning $SO(3)$ are given in the exercises, including Euler’s theorem that a rotation in $\mathbb{R}^3$ leaves an “axis of rotation” fixed pointwise; and the fact that $SU(2)$ is a “double cover” of $SO(3)$, so that each rotation in $\mathbb{R}^3$ has two representations as a norm 1 quaternion.

It is a theorem of Frobenius that the only associative quadratic real algebras without zero divisors are $\mathbb{R}, \mathbb{C}$, and $\mathbb{H}$; see [Ebbinghaus] for a proof. The double of the quaternions is an 8 dimensional algebra $\mathcal{O}$, called the octonions. It is has no zero divisors, but is not associative; we omit a proof. See [Baez] for an overview of relevance of the octonions to modern algebra.

Higher doublings are of interest; see [BremHentsz] for some recent research. However the only real division algebras are $\mathbb{R}, \mathbb{C}, \mathbb{H}$, and $\mathcal{O}$. The proof makes use of the “Bott periodicity theorem”. See [Ebbinghaus] for an overview, and [Atiyah] for a proof of the Bott periodicity theorem.

**14. Lagrange’s four square theorem.** Write a quaternion $x$ as $x_0 + x_1 i + x_2 j + x_3 k$. The conjugate $\bar{x}$ is $x_0 - x_1 i - x_2 j - x_3 k$; write $Q(x)$ for the quadratic form $x_0^2 + x_1^2 + x_2^2 + x_3^2$. Using observations in the preceding section, $Q(x) = \bar{x} x$, $\sum y, Q(xy) = Q(x)Q(y)$, and $x^{-1} = \bar{x}/Q(x)$.

The identity $Q(xy) = Q(x)Q(y)$ shows that the product of two sums of four squares is a sum of four squares itself; indeed of squares of bilinear forms in the $x_i$ and $y_i$, with integer coefficients. We leave it to the reader to work out the bilinear forms.

Lagrange’s four square theorem states that a nonnegative integer is a sum of four integers. By the preceding it suffices to prove this for a prime $p$. The quaternions may be used to give a proof of this, which can be found in [HardWir] for example. Let $\mathcal{H}_H$ denote $\{x : x_i$ is an integer for all $i$, or a half-integer for all $i\}$, where a half-integer is a rational number $n/2$ for $n$ an odd integer. One readily verifies that $\mathcal{H}_H$ is a subring of $\mathcal{H}$ (consider the above mentioned bilinear forms mod 4). One also verifies that for $x \in \mathcal{H}_H$ $Q(x)$ is an integer, and $x$ is a unit of $\mathcal{H}_H$ iff $Q(x) = 1$.

**Lemma 47.** Suppose $x, d \in \mathcal{H}_H$ and $d \neq 0$. Then there are $q, r \in \mathcal{H}$ with $x = qd + r$ and $Q(r) < Q(d)$.

**Proof:** First we show the theorem when $d$ is an integer $m$. Let $\zeta = (1 + i + j + k)/2$. Every $x \in \mathcal{H}_H$ can be written uniquely as $x_0^\prime \zeta + x_1^\prime i + x_2^\prime j + x_3^\prime k$, indeed $x_0^\prime = 2x_0$ and for $1 \leq i \leq 3, x_i^\prime = x_i - x_0$; the $x_i^\prime$ are integers. If $r = x - mq$ then $Q(r) = (r_0^\prime/2)^2 + \sum_{i=1}^3 (r_i^\prime/2 + r_i)^2$. Choose $q_0$ so that $|r_0^\prime| \leq m/2$; and then for $i > 0$ choose $q_i$ so that $|r_i^\prime/2 + r_i| \leq m/2$. For the resulting $r$, $Q(r) < m^2$. In the general case, let $m = Q(d)$; $q$ can be chosen so that $Q(xd - qm) < m^2$, whence $Q(x - qd)Q(d) < Q(d)^2$, whence $Q(x - qd)m < m^2$, whence $Q(x - qd) < m$.

Thus, $\mathcal{H}_H$ is a noncommutative version of a Euclidean ring. It follows by essentially the same argument as the proof of theorem 6.5.a that every left ideal is principal. Define $d$ to be a right divisor of $x$ if $x = qd$ for some $q$. Define $d$ to be a greatest common right divisor of $x$ and $y$ if $d$ is a right divisor if $x$ and $y$, and any other right divisor of $x$ and $y$ is a right divisor of $d$. It follows by essentially the same argument as the proof of theorem 6.5.g.1 that two elements $x, y \in \mathcal{H}_H$ have a greatest common right divisor, of the form $px + qy$; and two greatest common right divisors differ by a unit on the left.

**Lemma 48.** If $p$ an odd prime then for some $r, s$ with $0 \leq r, s < p$, $1 + r^2 + s^2 \equiv 0 \mod p$.

**Proof:** The numbers $x^2$ with $0 \leq x \leq (p + 1)/2$ are incongruent mod $p$, and likewise the numbers $1 - y^2$ with $0 \leq y \leq (p + 1)/2$. Therefore $x^2 \equiv 1 - y^2 \mod p$ for some such $x, y$.

**Lemma 49.** A prime integer $p$ is not a prime in $\mathcal{H}_H$.

**Proof:** Since $2 = (1 + i)(1 - i)$ we may assume $p$ is odd. Choose $r, s$ as in lemma 48, and let $x = 1 + ri + sj$. We claim that the greatest common right divisor $d$ of $x$ and $p$ is not a unit. Otherwise, $1 = ux + vp$ for some
\[
Q(u)Q(x) = Q(ux) = Q(1 - vp) = (1 - pv)(1 - p\bar{v}) = 1 - pv - p\bar{v} + p^2Q(v).
\]

Since \(p|Q(x)\) this is a contradiction. Now, \(x = ud\) and \(p = vd\) for some \(u, v \in \mathcal{H}_p\). If \(v\) is a unit then \(x = ud = uv^{-1}p\), which is clearly impossible.

**Theorem 50.** Any prime integer \(p\) is the sum of four integer squares.

**Proof:** Suppose \(p = xy\) where \(x, y \in \mathcal{H}_h\) and neither \(x, y\) a unit; then \(Q(x) = Q(y) = p\). If \(x\) has integer components the theorem is proved. Otherwise, write \(x\) as \(x' + u\) where \(x'\) is an even integer and \(u = \pm (1/2)\). Since \(Q(u) = 1\) \(u\) is a unit. Also \(x\bar{u} = x'v + 1\) has integral components. Finally, \(p = (x\bar{u})(v^{-1}y)\).

There are shorter proofs of Lagrange’s four square theorem, which use Fermat’s “method of descent”; see [IreRos]. The fact that a product of two sums of eight squares is a sum of eight squares may be proved using octonion multiplication. A theorem of Hurwitz states that a product of two sums of \(n\) squares is a sum of \(n\) squares, of bilinear forms in the original numbers, only holds when \(n = 1, 2, 4, 8\); see [Ebbinghaus] for a proof.

\(\mathcal{H}_h\) is called the Hurwitz quaternion ring. It has a variety of characteristics of interest. It is a maximal “order” in the \(Q\)-algebra of rational quaternions, where an order in a finite dimensional \(Q\)-algebra is a subring which is a full rank point lattice (point lattices are defined in chapter 23). The lattice of \(\mathcal{H}_h\) is in fact the \(D_4\) lattice (see chapter 23, or [ConSi]). The group of units has order 24; its elements comprise the shortest vectors of \(D_4\), and also the vertices of the 4-dimensional regular polytope called the 24-cell. See [Weil] for basic facts about orders in algebraic number fields.

**Exercises.**

1. Let \(\zeta = (−1 + \sqrt{-3})/2\). For an element \(\alpha = m + n\zeta\) in \(\mathbb{Z}[\zeta]\) let \(\nu(\alpha) = m^2 - mn + n^2\). Show that \(\nu\) is a Euclidean norm on \(\mathbb{Z}[\zeta]\). Hint: Using \(\zeta + \zeta^* = -1\) and \(\zeta\zeta^* = 1\) it follows that \(\nu(\alpha) = \alpha\alpha^*, \) and \(\nu\) is multiplicative. Given \(\alpha, \beta \in \mathbb{Z}[\zeta]\) with \(\beta \neq 0\), write \(\alpha = r + s\zeta\) where \(r, s \in \mathbb{Q}\), and choose \(m, n \in \mathbb{Z}\) with \(|r - m| \leq 1/2\) and \(|s - n| \leq 1/2\); let \(\gamma = m + n\zeta\). Then either \(\alpha\gamma = \beta, \) or \(\nu(\alpha/\beta - \gamma) < 1\) and \(\nu(\alpha - \gamma\beta) < \nu(\beta)\).

2. Suppose \(G\) is a finite group, \(F\) is a field of characteristic 0, and \(F\) splits \(G\), and recall the notation of chapter 15.7. Give an alternate proof that \(n_1\) divides \(|G|\), which uses theorem 16 and does not assume that \(F = \mathbb{C}\). Hint: From the proof of theorem 15.26, we have \(e_i = \sum\alpha a_{i\alpha}g\) where \(a_{i\alpha} = n_i\chi_i(g)^*/|G|\). Let \(t\) be an exponent for \(G\), \(\xi\) a primitive \(t\)-th root of unity, and \(M\) the \(\mathbb{Z}\)-module in \(F[G]\) generated by the elements \(\xi^n g e_i, 0 \leq u < t, g \in G, 1 \leq i \leq s\). Multiplication by \(|G|/n_1\) (which we know exists) maps \(M\) into itself, by the above expression for \(e_i\).

3. Show that if \(R\) is a commutative ring, \(I \subseteq R\) is an ideal, and \(S \subseteq R\) is a multiplicative subset then \(\text{Rad}(I)S = \text{Rad}(IS) = \{[r/s] : r^n s_2 \in I \text{ for some } s_2 \in S, n\}\). Hint: \([r/s] \in \text{Rad}(I)S\) iff \(r s_1 t = r_1 st\) where \(r^n_1 \in I\) for some \(n\) and \(s_1, t \in S\). For such \([r/s]\) let \(s_2 = s^n t^n\). If \(r^n s_2 \in I\) then \([r/s] = [rs_2/ss_2]\) where \((rs_2)^n \in I\). \([r/s] \in \text{Rad}(I)S\) iff \(r^n s_1 t = r_1 s^n t\) for some \(n, r_1 \in I, s_1, t \in S\). For such \([r/s]\) let \(s_2 = s_1 t\). If \(r^n s_2 \in I\) \([r^n/ss^n] = [r^n s_2/ss_2]\).

4. Suppose the characteristic of \(F\) is not 2. Show that \(q\) is a quadratic form on a vector space \(V\) (i.e., \(b(x, x)\) for a symmetric bilinear form \(b\)) if \(q(sx) = s^2 q(x)\) for \(s \in F\), and \(q(x + y) - q(x) - a(y)\) is bilinear.

5. Show that in \(\mathbb{R}^n\), the matrix \(M\) preserves the Euclidean norm iff \(M^t M = MM^t = I\). Hint: \(|Mx| = |x|\) iff \(x^t M^t M x = x^t x\), and \(M\) preserves the quadratic form iff it preserves the bilinear from by exercise 4.

6. Show that a rotation in \(SO(2)\) is a matrix of the form

\[
R_\theta = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}
\]
for a unique $\theta$ with $0 \leq \theta < 2\pi$. Further $R_{\theta+\phi} = R_\theta R_\phi$.

7. Show that for $M \in SO(n)$ for $n$ odd, $M$ it fixes a line through the origin pointwise. Hint:

$$\det(M - I) = \det((M - I)^t) = \det(1 - M^t) = \det(1 - M) = -\det(M - I).$$

Thus, $\det(M - I) = 0$ and there is a nonzero vector $x$ such that $Mx = x$.

8. Suppose $M \in SO(3)$, and $n$ is a unit vector along an axis fixed by $M$. Show that there is a unique $\theta$ such that for all $x$,

$$Mx = (n \cdot x)n + [x - (n \cdot x)n] \cos \theta + (n \times x) \sin \theta.$$

Conversely, an expression as on the right is a rotation. Hint: For the first claim, choose an orthogonal triple $u, v, n$, and let $T$ be the matrix with these columns. Then $T^tMT$ is of the form

$$\begin{bmatrix} R_\theta & 0 \\ 0 & 1 \end{bmatrix},$$

where $\theta$ is independent of the choice of $u, v$.

9. Given an expression as in exercise 8, let

$$e_0 = \cos(\theta/2), \quad e_I = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = n \sin(\theta/2).$$

Show that

$$Mx = (e_0^2 - e_1^2 - e_2^2 - e_3^2)x + 2(e_1 \cdot x)e_I + 2e_0(e_I \times x).$$

Show that conversely, given $e_i$ with $\sum_i e_i = 1$, a rotation is determined via these expressions.

10. Suppose

$$M = \begin{bmatrix} e_0^2 + e_1^2 - e_2^2 - e_3^2 & 2(e_1e_2 - e_0e_3) & 2(e_1e_3 + e_0e_2) \\ 2(e_1e_2 + e_0e_3) & e_0^2 - e_1^2 + e_2^2 - e_3^2 & 2(e_2e_3 + e_0e_1) \\ 2(e_1e_3 - e_0e_2) & 2(e_2e_3 - e_0e_1) & e_0^2 - e_1^2 - e_2^2 + e_3^2 \end{bmatrix},$$

where $\sum_i e_i = 1$. Show that $M$ is a rotation (the $e_i$ are called real Euler parameters for the rotation). Let

$$Q = \begin{bmatrix} e_0 + e_1i \\ -e_2 + e_3i \\ e_2 + e_3i \end{bmatrix}, \quad \text{and} \quad P_x = \begin{bmatrix} x_1i \\ -x_2 + x_3i \\ -x_2 + x_3i \end{bmatrix},$$

for a vector $x$. Show that $y = Mx$ iff $P_y = QP_xQ^\dagger$. Show that the map $Q \mapsto M$ is a group homomorphism. Show that its image is $SO(3)$ and its kernel is $\mathbb{Z}_2$. Hint: Use exercise 9 to show that $M$ is a rotation. Flipping the signs of all the $e_i$ corresponds to flipping the signs of $n$ and $\theta$ in the expression in exercise 8; the identity is a special case.

11. For this exercise modify the correspondence of exercise 10 by letting

$$Q = \begin{bmatrix} e_0 + e_3i & e_2 + e_1i \\ -e_2 + e_1i & e_0 - e_3i \end{bmatrix}. $$

Let

$$Q_\phi = \begin{bmatrix} e^{-i\phi/2} & 0 \\ 0 & e^{i\phi/2} \end{bmatrix}, \quad Q_\theta = \begin{bmatrix} \cos \theta & -i \sin \theta \\ -i \sin \theta & \cos \theta \end{bmatrix}, \quad Q_\psi = \begin{bmatrix} e^{-i\psi/2} & 0 \\ 0 & e^{i\psi/2} \end{bmatrix}. $$

Let

$$M_\phi = \begin{bmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad M_\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix}, \quad M_\psi = \begin{bmatrix} \cos \psi & \sin \psi & 0 \\ -\sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{bmatrix}. $$

260
Show that $Q_{\phi}$ maps to $M_{\psi}$, etc., under the modified correspondence. Show that

$$Q = Q_\psi Q_\theta Q_\phi = \begin{bmatrix} e^{-i(\psi+\phi)/2} \cos \theta/2 & -i e^{-i(\psi-\phi)/2} \sin \theta/2 \\ -i e^{i(\psi-\phi)/2} \sin \theta/2 & e^{i(\psi+\phi)/2} \cos \theta/2 \end{bmatrix}$$

maps to

$$M = M_\psi M_\theta M_\phi = \begin{bmatrix} \cos \phi \cos \psi - \sin \phi \cos \theta \sin \psi & \sin \phi \cos \psi + \cos \phi \cos \theta \sin \psi & \sin \psi \sin \theta \\ -\cos \phi \sin \psi - \sin \phi \cos \theta \cos \psi & -\sin \phi \sin \psi + \cos \phi \cos \theta \cos \psi & \cos \psi \sin \theta \\ \sin \phi \sin \theta & -\cos \phi \sin \theta & \cos \theta \end{bmatrix}.$$ 

Let

$$e_0 = \cos \frac{\psi+\phi}{2} \cos \frac{\theta}{2}, \quad e_1 = -\cos \frac{\psi-\phi}{2} \sin \frac{\theta}{2}, \quad e_2 = \sin \frac{\psi-\phi}{2} \sin \frac{\theta}{2}, \quad e_3 = -\sin \frac{\psi+\phi}{2} \cos \frac{\theta}{2}.$$ 

Show that these are real Euler parameters for $M$. The angles $\phi, \theta, \psi$ are called Euler angles for the rotation, and have a geometric interpretation which can be seen from the expression of $M$ as the product. It is conventional to first rotate about the $z$ axis, necessitating the change in the map from $SU(2)$ to $SO(3)$. The entries in $Q$ are called Cayley-Klein parameters.
21. Lattice orders.

1. Varieties of lattices. Lattices are defined in chapter 3, and considered further in chapters 11 and 16, exemplifying their utility in algebra. Lattice theory is also useful in other areas, and is a subject of study in its own right. [Gratzer], [Johnstone] and [Gierz et al] are among numerous texts in the subject. The title of this chapter is “lattice orders” because the term “lattice” has another important use, as will be seen in chapter 23.

Let PoSet denote the category of partially ordered sets. In chapter 3 the subcategories of join sublattices, meet sublattices, and lattices have been defined; these will be denoted SLat⊔, SLat⊓, and Lat. These categories are of the form MdlT for T a set of equations; such categories are called “equational classes” or “varieties” of “algebras” (i.e., structures). As observed earlier in the text, many basic facts about the category (such as the construction of limits and colimits) follow.

In an equational class the quotient of a structure by a congruence relation, as defined in chapter 2, is in the class; indeed this can be verified when the axioms are Horn formulas where the only predicate is equality. This is not true for Horn formulas in general, as the case of PoSet shows. Any equivalence relation is a congruence relation, and examples where antisymmetry is not preserved are easy to construct.

In the former case every congruence relation is the kernel of an epimorphism (and coinage-image-factorizations exist). In the latter the canonical epimorphism of structures may no longer be in the category, and there may be congruence relations which do not arise from epimorphisms. In PoSet strong homomorphisms are isomorphisms, and the discrete relation is the only strong congruence relation. Congruence relations where the quotient is a partial order are an intermediate type.

SLat⊔ and SLat⊓ have the same axioms. The category SLat can be defined as MdlT for these axioms. However it cannot be considered a subcategory of PoSet since which of the two ways of choosing the order is not prescribed.

In SLat⊔ a morphism must preserve ⊔, and SLat⊔ is not a full subcategory of PoSet. Neither is Lat a full subcategory of SLat⊔. Any of these categories C may be restricted by requiring the existence of a least element (C₀), a greatest element (C₁), or both (CB). Again these are not full subcategories, since morphisms must preserve the constant(s). They are all equational classes; the axiom for 0 is either x⊔0 = x or x∩0 = 0, etc. LatB is often called the bounded lattices. As mentioned earlier some authors require lattices to be bounded. The category SLat⊔₀ is of such importance in some areas that semilattices are required to be of this type; etc.

The varieties of modular and distributive lattices have also been defined already; these will be denoted MLat and DLat. These are full subcategories of Lat. DLat is clearly an equational class. MLat is also; the modular law may be replaced by

\[(x \cap y) \cup (x \cap z) = x \cap (y \cup (x \cap z)).\]

Indeed, if this is true and z ≤ x then x ∩ z = z, so (x ∩ y) ∪ z = x ∩ (y ∪ z). On the other hand, since x ∩ z ≤ z this equation is implied by the modular law.

It was observed in section 16.2 that the submodules of a module comprise a modular lattice. The subspaces of the vector space F² over a field F are not a distributive lattice, though; indeed for three distinct one dimensional subspaces a, b, c, (a ⊔ b) ∩ c = c but (a ∩ c) ⊔ (b ∩ c) = {0}. The “fundamental theorem of projective geometry” states that for a finite dimensional vector space of dimension at least 3 over a division ring, the dimension and the division ring are uniquely determined by the lattices of subspaces; see [Jacobson] for a proof.

In a lattice L ∈ LatB an element y ∈ L is a complement for x ∈ L if y ∩ x = 0 and y ∪ x = 1. Recall (lemma 3.6) that for lattices in DLatB, a complement is unique if it exists. Note also that lattice morphisms preserve any complements which exist.
A lattice in $\text{DLat}_B$ in which every element has a complement is called a Boolean algebra. This category is denoted $\text{Bool}$. The Boolean algebras are an equational class (axioms are given in chapter 3, involving a unary complementation operator $x^c$). $\text{Bool}$ is a full subcategory of $\text{Lat}_B$, since lattice morphisms preserve complements.

Consider the following axioms for a unary function in a bounded lattice:

1. $x \cap x^c = 0$,
2. $x \cup x^c = 1$,
3. $x^{cc} = x$, and
4. $x \leq y \Rightarrow y^c \leq x^c$.

A bounded lattice satisfying 1 and 2 is called a lattice with complementation. A bounded lattice in which every element has a unique complement is called uniquely complemented; such a lattice satisfies axioms 1, 2, and 3. Note that a distributive lattice with complementation is the same thing as a Boolean algebra.

Suppose 3 and 4 are satisfied. DeMorgan’s laws $(x \cup y)^c = x^c \cap y^c$, $(x \cap y)^c = x^c \cup y^c$ are satisfied (exercise 11). Also, $x^c \leq 1$, so $1^c \leq x^{cc}$, so $1^c = 0$, and similarly $0^c = 1$. Finally either of 1 or 2 imply the other. A lattice satisfying all 4 axioms is said to be an ortholattice. The ortholattices are an equational class; by an argument as above $x \leq y \Rightarrow y^c \leq x^c$ may be replaced by $(x \cup y)^c \leq x^c$.

In an ortholattice the following are equivalent:

1. $x \leq y \Rightarrow y = x \uplus (y \cap x^c)$,
2. $x \uplus ((x \uplus y) \cap x^c) = x \uplus y$,
3. if $x = (x \cap y) \uplus (x \cap y^c)$ then $y = (y \cap x) \uplus (y \cap x^c)$, and
4. if $x \leq y$ and $y \cap x^c = 0$ then $x = y$

(exercise 12). An ortholattice satisfying them is called an orthomodular lattice. By 2 the orthomodular lattices are an equational class. A modular ortholattice is an orthomodular lattice, since 1 is a special case of the modular law. In chapter 24 the closed subspaces of an infinite dimensional Hilbert space will be shown to comprise an orthomodular lattice which is not modular.

In a lattice, an element $z$ is said to be a pseudo-complement of $x$ relative to $y$ if $z$ is largest among the elements $z'$ such that $x \cap z' \leq y$. The notation $x \rightarrow y$ is used to denote a relative pseudo-complement. The following facts are readily verified.

- The relative pseudo-complement is unique if it exists.
- If $y_1 \leq y_2$, and $x \rightarrow y_1$ and $x \rightarrow y_2$ both exist, then $x \rightarrow y_1 \leq x \rightarrow y_2$.
- $z \leq x \rightarrow y$ iff $x \cap z \leq y$.
- $x \rightarrow x$ exists for some $x$ iff 1 exists, in which case $x \rightarrow x = 1$ for all $x$.
- In a Boolean algebra $x \rightarrow y = x^c \cap y$.

The definition of relative pseudo-complement in fact can be given in a $\cap$-semilattice.

A Heyting algebra is defined to be a lattice in which $x \rightarrow y$ exists for all $x$ and $y$. This is readily seen to be so iff the map $y \mapsto x \cap y$ has a right adjoint for all $x$; and the right adjoint is $x \rightarrow y$. Since $y \mapsto x \cap y$ is a left adjoint, $\cap$ distributes over any joins that exist; in particular a Heyting algebra is a distributive lattice (a direct proof can readily be given by the usual adaptation).

It is interesting to note that a Heyting algebra is a poset which, as a small category, has all finite products and coproducts, and is Cartesian closed.

**Theorem 1.** In a Heyting algebra the following are true:

a. $x \rightarrow x = 1$;
b. $x \cap (x \rightarrow y) = x \cap y$;
c. $y \cap (x \rightarrow y) = y$;
d. $x \rightarrow (y \cap z) = (x \rightarrow y) \cap (x \rightarrow z)$.
Conversely if a lattice satisfies these axioms it is a Heyting algebra, with \( x \to y \) the relative pseudo-complement.

**Proof:**

Part a was observed above (\( z \leq x \to x \) since \( x \sqcap x \leq x \)). For part b, let \( L \) denote \( x \sqcap (x \to y) \). Clearly \( L \leq x \), and \( L \leq y \) by the adjunction. On the other hand \( x \sqcap y \leq x \), and \( x \sqcap y \leq x \to y \) by the adjunction. For part c, \( y \sqcap (x \to y) \leq y \), and by the adjunction \( y \leq x \to y \). Part d follows since \( y \to x \to y \) is a right adjoint. For the converse, suppose \( z \leq x \to y \); then \( x \sqcap z \leq x \sqcap x \to y \), so by b \( x \sqcap z \leq x \sqcap y \leq y \). Now, by c \( z = z \sqcap (x \to z) \), so by a \( z \leq (x \to x) \sqcap (x \to z) \), so by d \( z \leq (x \sqcap z) \). Also, by d \( y \to x \to y \) is order preserving. Thus, if \( x \sqcap z \leq y \) then \( z \leq x \to y \).

In particular the Heyting algebras are an equational class. \( \text{HAlg} \) will be used to denote the category; morphisms of course must preserve \( \to \). \( \text{HAlg}_0 \) will denote the Heyting algebras with 0. Unlike \( \text{Bool} \), \( \text{HAlg} \) is not a subcategory of \( \text{Lat} \); an example will be given below. It is true that if \( f \) is a lattice homomorphism then \( f(x \to y) \leq f(x) \to f(y) \); indeed,

\[
f(x \sqcap f(y) = f(x \sqcap x \to y) = f(x \sqcap y) = f(x) \sqcap f(y) \leq f(y).\]

In a Heyting algebra with 0, \( x \to 0 \) is the largest \( w \) such that \( w \sqcap x = 0 \). This may be defined for \( L \in \text{SLat}_{0} \); it is unique if it exists, and is called the pseudo-complement. Suppose \( L \in \text{SLat}_{0} \) is such that the pseudo-complement exists for all \( x \in L \); let \( x^\text{pp} \) denote it. It is readily verified that \( x \leq y \) iff \( x^p \geq y^p \), and that \( z \leq y^p \) iff \( z^p \geq y^p \); thus, the map \( x \mapsto x^p \) from \( L^\text{pp} \) to \( L \) is a right adjoint, with left adjoint the same map in the opposite direction. Also, \( 0^p = 1 \) and \( L \) is bounded.

In particular \( x \leq x^\text{pp} \) and \( x^\text{pp} = x^p \). It follows that \( L^p = \{ x^p : x \in L \} \) equals \( \{ x \in L : x^p = x \} \). For \( x, y \in L^p \), \( x = x^{pp} \geq (x \sqcap y)^{pp} \), and similarly \( y \geq (x \sqcap y)^{pp} \), so \( x \sqcap y \geq (x \sqcap y)^{pp} \), and \( x \sqcap y = (x \sqcap y)^{pp} \). For \( x, y \in L^p \), \( x = x^{pp} \geq (x \sqcap y)^{pp} \) since \( x^p \geq x^p \sqcap y^p \), and similarly \( y \leq (x \sqcap y)^{pp} \). Further if \( x, y \leq w \) where \( w \in L^p \) then \( x^p \sqcap y^p \geq w^p \), so \( (x^p \sqcup y^p)^p \leq w^{pp} = w \). Thus, \( L^p \) is a sub-lattice-\( B \)-algebra. L, and is a lattice with join \( (x \sqcup y)^{pp} \).

It follows by the foregoing that if \( L \) is a Heyting algebra with 0 then \( L^p \) is a Boolean algebra, where \( x^p \) the complement; in fact this is true for \( L \in \text{SLat}_{B} \) with a pseudo-complement (see [Gratzer]). Also \( (x \sqcup y)^{pp} = x^p \sqcap y^p \) by the adjunction, in fact for \( L \in \text{Lat}_0 \) with a pseudo-complement. It also follows that \( (x \sqcap y)^{pp} = x^{pp} \sqcup y^{pp} \) (exercise 1).

Heyting algebras where \( L^p \) is a sublattice are of interest; we leave it as exercise 2 to show that in a Heyting algebra \( L \) with 0 (indeed, pseudo-complemented distributive bounded lattice) the following are equivalent.

\[ a. \ L^p \text{ is a sublattice (i.e., for } x, y \in L^p, x \sqcup y \in L^p). \]
\[ b. \text{ For all } x \in L \ x^p \sqcup x^{pp} = 1. \]
\[ c. \text{ For all } x, y \in L \ (x \sqcap y)^{pp} = x^p \sqcap y^p. \]

It is easily seen that a complement is a pseudo-complement, so b holds iff the elements of \( L \) possessing a complement are those of \( L^p \).

In \( \text{SLat}_{0} \), if \( h \) is a homomorphism (or \( \equiv \) a congruence relation) then \( h^{-1}(0) \) (or \( [0] \)) is readily verified to be an ideal. Dually in \( \text{SLat}_{1} \), \( h^{-1}(1) \) (or \( [1] \)) is a filter. In general the map from congruence relations to filters is neither surjective nor injective (examples will be given below). In a Heyting algebra it is bijective; let \( x \leftrightarrow y \) denote \( (x \to y) \sqcap (y \to x) \).

**Theorem 2.** If \( F \) is a filter in a Heyting algebra define \( x \equiv y \) iff \( x \leftrightarrow y \in F \); then \( \equiv \) is a congruence relation. The map \( F \mapsto \equiv \), and the map \( \equiv \mapsto [1] \), are inverse.

264
Proof: Reflexivity follows from \( x \leftrightarrow x = 1 \), and symmetry from \( x \leftrightarrow y = y \leftrightarrow x \). Now,
\[
x \sqcap ((x \to y) \sqcap (y \to z)) = (x \sqcap y) \sqcap (y \to z) \leq y \sqcap (y \to z) = y \sqcap z \leq z.
\]
Thus \( (x \to y) \sqcap (y \to z) \leq x \to z \), and transitivity follows. The second map is left inverse to the first because \( x = x \to 1 \). It is right inverse because \( x \to y \equiv 1 \) iff \( x \equiv y \). The reverse implication is clear; for the forward implication,
\[
x \equiv x \sqcap (x \to y) = x \sqcap y = y \sqcap (x \to y) \equiv y.
\]

The correspondence between filters and congruence relations is in fact an isomorphism of algebraic closure systems, because it respects inclusion. The first claim for Boolean algebras was proved in chapter 11. In a Boolean algebra \( F \) is a filter iff \( \text{co-}F \) is an ideal, and the correspondence is bijective. In a Heyting algebra if \( F \) is a filter then \( F^p \) is an ideal, where \( F^p \) is the pseudo-complements, as is readily seen.

Suppose \( L \in \text{SLat}_{\infty}, \) and \( F \) is a proper filter in \( L \). For \( x \in L \), the filter generated by \( F \) and \( x \) is proper if \( w \sqcap x \neq 0 \) for any \( w \in F \). It follows that \( F \) is maximal iff any such \( x \) is already in \( F \); or iff, for any \( x \notin F \), there is a \( w \in F \) with \( w \sqcap x = 0 \). If \( L \in \text{HAlg}_{\infty} \) the latter is true iff \( x^p \in F \); that is, \( F \) is maximal iff \( x \notin F \) implies \( x^p \in F \).

In a Boolean algebra the dual statement holds also, that is, a proper ideal \( I \) is maximal iff \( x \notin I \) implies \( x^c \in I \). In a Heyting algebra with \( 0 \), if \( I \) is maximal and \( x \notin I \) then \( x^p \in I \). This follows because there is some \( w \in I \) with \( w \sqcap x = 1 \), and for such a \( w \) \( x^p \leq w \); indeed,
\[
x^p = x^p \sqcap 1 = x^p \sqcap (w \sqcup x) = (x^p \sqcap w) \sqcup (x^p \sqcap x) = x^p \sqcup w.
\]

However, there can be ideals with this property, which are not maximal. The unit interval in \( R \) with the usual order is a Heyting algebra, and the reader may verify that there are non-maximal ideals having this property.

Theorem 3. Suppose \( D \) is a complete category and \( G \) is a functor from \( D \) to a category \( C \). Then \( G \) has a left adjoint iff \( G \) preserves limits, and for all \( c \in C \) there is a set \( S \) of arrows \( \{ f_i : c \to G(d_i) \} \), such that for any arrow \( f : c \to G(d) \) there is an \( f_i \in S \) and an arrow \( h : d_i \to d \) such that \( f = G(h) f_i \).

Proof: If \( G \) has a left adjoint then by theorem 13.13 it preserves limits; and by the remarks following lemma 13.3 \( S \) can be taken as the universal arrow from \( c \) to \( GF(c) \). The converse will not be proved here; the reader is referred to Theorem V.6.2 of [MacLane].

Corollary 4. Suppose \( D \) is a complete lattice and \( G \) is an order preserving map from \( D \) to a poset \( C \). Then \( G \) has a left adjoint iff it preserves meets.

Proof: It has already been shown that a right adjoint preserves meets. For the converse, let \( F(c) = \sqcap \{ d \in D : c \leq G(d) \} \). Since \( G \) preserves meets \( GF(c) = \sqcap \{ G(d) : c \leq G(d) \} \), whence \( GF(c) \geq c \). Since \( FG(d) = \sqcap \{ d' : G(d) \leq G(d') \} \), \( FG(d) \leq d \). By remarks at the beginning of section 13.7, \( F \) is a left adjoint to \( G \).

The theorem is called Freyd’s adjoint functor theorem. The dual of the corollary states that an order preserving map from a complete lattice to a poset has a right adjoint iff it preserves joins. From this, a finite distributive lattice is a Heyting algebra.

In giving examples of finite partial orders (and for other purposes) the notion of the Hasse diagram of a poset is useful. This is a directed graph, where there is an arrow from node \( x \) to node \( y \) if \( x < y \) and for
no \( w \) does \( x < w < y \) hold (\( y \) is said to cover \( x \)). Figure 1 gives some examples; the arrowheads are omitted, with the understanding that the edges are directed up in the drawing.

It is easy to see that these posets are all lattices. Call them I to III from left to right; I is not modular, II is modular but not distributive, and III is distributive. The last fact is easy to verify from the theorem that a lattice is modular iff it does not contain I as a sublattice; and distributive iff it does not contain either I or II. A proof of this will not be given here; see [Gratzer].

The ideal \( \{0, a\} \) in lattice I is not \([0]\) for any congruence relation. Indeed, if \( a \equiv 0 \) then taking the join with \( c \), \( 1 \equiv c \). Now taking the meet with \( b \), \( b \equiv 0 \). On the other hand, in the three element chain, the ideal \( \{0\} \) is \([0]\) for both the congruence relations \( \{0\} \), \( \{1, 2\} \) and \( \{0\} \), \( \{1\} \), \( \{2\} \).

Lattice III has 5 filters; by theorem 2 each determines a unique Heyting algebra congruence relation. It may be verified that there are 8 lattice congruence relations.

2. Complete lattices. Let Clat be the category of complete lattices, with morphisms preserving all joins (in particular 1) and all meets (in particular 0). By lemma 3.4, in the category of complete join semilattices, with morphisms preserving all joins, the objects are the same as those of Clat; but the morphisms are different. Let Clat\(_\uparrow\) denote this category. Requiring in addition that finite meets, or directed meets, be respected results in further categories, denoted Clat\(_{\uparrow\vee}\) and Clat\(_{\uparrow\wedge}\) (requiring both yields Clat). Clat\(_{\downarrow\vee}\), Clat\(_{\downarrow\wedge}\), and Clat\(_{\downarrow\wedge}\) are defined dually.

Although in practice some of the foregoing categories are of greater interest, general facts can be considered. For example, limits exist; the proof will make use of theorem 13.12, although the reader may verify that in Set the resulting limits are the same as those of theorem 13.5.

For some preliminary facts, let \( P \) be a poset and let \( Q \) be a suborder (subset with the restriction order). Suppose \( S \subseteq Q, m \in Q, \) and \( m = \sqcap S \) in \( P \); then \( m = \sqcap S \) in \( Q \), as is readily verified. Let \( h \) be a map between posets which preserves either binary or directed joins; then \( h \) is order preserving. The dual facts hold.

**Theorem 5.** Let \( C \) be any of the categories of the first paragraph.

- a. The product in \( C \) is the Cartesian product.
- b. The equalizer in \( C \) is the same as that in Set.
- c. The quotient by the equivalence relation of a morphism yields a coimage-image factorization.

**Proof:** For part a, let component \( i \) of the join of a set of sequences be the join of the \( i \)th component of the sequences. Given \( \beta_i : P \to L_i \), let \( h : P \to \times_i L_i \) be the induced map. If all \( \beta_i \) preserve a join of the \( i \)th components then \( h \) does. Dual statements hold for meets. For part b, \( \{ x : f(x) = g(x) \} \) is closed under either joins or meets, so is a complete lattice. For part c, let \( h : L \to M \) be a morphism (say preserving all joins, with the remaining cases being dual); in particular it is order preserving. Let \( \equiv \) be the equivalence relation \( h(x) \equiv h(y) \). Since \( h \) preserves joins, they may be defined on \( L/\equiv \), whence it is an object in the
category; further the map \( e : L \rightarrow L/\equiv \) mapping \( x \) to \([x]\) preserves joins. Now, \([x] \leq [y]\) iff \([x] \cup [y] = [y]\) iff \([x] \cup [y] = [y]\) iff \( h(x \cup y) = h(y)\) iff \( h(x) \cup h(y) = h(y)\) iff \( h(x) \leq h(y)\); that is, the corestriction of the map \( i : L/\equiv \rightarrow M \) mapping \([x]\) to \(h(x)\) to \(h[L]\) is an order isomorphism. It follows that \( e \) preserves meets which \( h \) does; and \( i \) also.

Colimits will be given a more cursory discussion. Firstly, the free algebra (needed for example for theorem 13.11) might not exist; see [Johnstone] for a proof that in CLat it does not. Exercise 3 shows that the free algebra exists in CLat, and that colimits exist. There are general theorems giving circumstances under which the existence of the free algebra implies the existence of colimits.

Other categories of interest are obtained by restricting the objects of CLat, for example to be modular lattices (CMLat), distributive lattices (CDLat), or Boolean algebras (CBool). Theorem 5 is readily verified to hold in these categories.

By corollary 4 if all joins exist and meet distributes over them the lattice is a complete Heyting algebra. The category Frm (the category of frames) has these as objects; morphisms must preserve finite meets and all joins (in particular 0 and 1). Let CHAlg denote the category where morphisms must preserve \( \rightarrow \), all meets, and all joins. Again, theorem 5 is readily verified to hold in these categories. Also, colimits exist in Frm (exercise 4).

If \( L \) is a complete Heyting algebra then the Boolean algebra \( L^p \) defined in section 1 is complete. Indeed, the meet in \( L^p \) is the meet in \( L \), by a slight generalization of the proof for finite meets.

Another category of interest, which we denote PoSet\(_{D\cup} \), is the posets which are closed under directed joins; morphisms must preserve directed joins. The dual category PoSet\(_{D\cap} \) may also be defined (but is of lesser interest). Theorem 5 may be verified to hold for these categories. Also, colimits exist in PoSet\(_{D\cup} \) (exercise 5).

If \( P \in \text{PoSet}_{D\cup} \) say that \( U \subseteq P \) is Scott open if \( U \) is \( \geq \)-closed; and whenever \( \sqcup D \in P \) for some directed \( D \subseteq P, D \cap U \neq \emptyset \). The Scott open sets form a topology on \( P \) (exercise 6). The closed sets are those which are \( \leq \)-closed and closed under directed joins. Various topologies are defined on posets; for example the interval topology for some collection of intervals is that where the intervals form a subbase for the closed sets. In the upper interval topology, the intervals are the sets \( \{x^\leq\} \); for \( P \) in PoSet\(_{D\cup} \), this topology is contained in the Scott topology (by exercise 7).

The following facts needed for the proof of the next theorem are left to the reader. Let \( P, Q \) be partial orders.

- Suppose \( f : P \rightarrow Q \) is order preserving. If \( V \subseteq Q \) is \( \geq \)-closed then \( f^{-1}[Q] \) is \( \geq \)-closed.
- If \( S \subseteq P \) then \( \{x \in P : x \notin S\} \) is open in the upper interval topology.

**Theorem 6.** Suppose \( P, Q \in \text{PoSet}_{D\cup} \), and \( f : P \rightarrow Q \). Then \( f \) preserves directed joins iff \( f \) is a continuous function when \( P \) and \( Q \) are equipped with their Scott topologies (is “Scott continuous”).

**Proof:** Suppose \( f \) preserves directed joins, and \( V \subseteq Q \) is Scott open. As observed above, \( f^{-1}[V] \) is \( \geq \)-closed. If \( D \) is directed and \( \sqcup D \in f^{-1}[D] \) then \( \sqcup f[D] = f(\sqcup D) \in V \), whence there is some \( p \in D \) with \( f(p) \in V \), or \( p \in D \cap f^{-1}[V] \). Suppose \( f \) is Scott continuous. Suppose for \( x, y \in P \) that \( x \leq y \). Since \( V = \{q \in Q : q \notin f(y)\} \) is Scott open \( f^{-1}[V] \) is Scott open. If \( f(x) \notin f(y) \) then \( x \in f^{-1}[V] \), whence \( y \in f^{-1}[V] \), a contradiction. That is, \( f \) is order preserving. Suppose \( D \subseteq P \) is directed. Then, \( V = \{q \in Q : q \notin f(D)\} \) is Scott open, and so \( f^{-1}[V] \) is Scott open. Since \( f^{-1}[V] \cap D = \emptyset, \sqcup D \notin f^{-1}[V] \), and it follows that \( f(\sqcup D) \leq \sqcup f[D] \). Since \( f \) has already been shown to be order preserving, \( f(\sqcup D) = \sqcup f[D] \).

If \( Y \) is a poset the set \( Y^X \) of functions from \( X \) to \( Y \) may be made a poset, by defining \( f \leq g \) iff \( f(x) \leq g(x) \) for all \( x \); this is just the product order. If \( Y \) is closed under finite, directed, or arbitrary joins so is \( Y^X \), with the join taken “pointwise”, e.g., \((f \sqcup g)(x) = f(x) \sqcup g(x)\). The analogous fact holds for meets.
In a category of posets, $\text{Hom}(X,Y)$ has the order induced as a subset of $Y^X$. One verifies that in the categories of theorem 5, $\text{Hom}(X,Y)$ is closed under the joins or meets preserved by the morphisms. This also holds for $\text{PoSet}_{\mathcal{D},\cup}$.

**Theorem 7.** In $\text{ Frm Hom}(X,Y)$ with the product order is closed under directed joins.

**Proof:** Suppose $\{f_i\}$ is a set of frame morphisms, directed in the product order on $\text{Hom}(X,Y)$, and let $f = \sqcup f_i$. Then

$$f(\sqcup_j x_j) = \sqcup_i f_i(\sqcup_j x_j) = \sqcup_i \sqcup_j f_i(x_j) = \sqcup_j \sqcup_i f_i(x_j) = \sqcup_j f(x_j),$$

showing that $f$ preserves arbitrary joins. Also, $f(a) \sqcap f(b) = (\sqcup_i f(a)) \sqcap (\sqcup_i f(b))$; by distributivity this equals $\sqcup_i f(a) \sqcap f(b)$. Since $\{f_i\}$ is directed this equals $\sqcup_k f_k(a) \sqcap f_k(b)$. By the usual argument this equals $f(a \sqcap b)$.

**Theorem 8.** $\text{PoSet}_{\mathcal{D},\cup}$ is Cartesian closed.

**Proof:** Let $C$ denote $\text{PoSet}_{\mathcal{D},\cup}$. Suppose $f : A' \to A$ and $g : B \to B'$; then $h \mapsto ghf$ is the arrow in $\text{Set}$ from $\text{Hom}(A,B)$ to $\text{Hom}(A',B')$. In fact this arrow is in $C$, that is, preserves directed joins. This may be seen by applying $g(\sqcup_i h_i)f$ and $\sqcup_i (gh_i)f$ to an element of $A'$, where $\{h_i\}$ is a directed set, noting that $g$ preserves directed joins. In particular $\text{Hom}(\_,\_)$ may be regarded as a functor from $C^{op} \times C$ to $C$. Recall the notation $\bar{g}(x)(y) = g(x,y)$ introduced prior to theorem 13.4, and suppose $g$ is an arrow in $C$. In particular it is continuous in $y$, whence $g_x$ is in $C$ for all $x \in X$. (A function in several variables is said to be continuous in one of them if it is continuous in the variable, for every value of the other variables). To see that $\bar{g}$ preserves directed sups, apply $\sqcup_i \bar{g}(x_i)$ and $\bar{g}(\sqcup_i x_i)$ to $y$ and use continuity of $g$ in $x$. Now suppose $\bar{g}$ is continuous.

If $\bar{g}$ is in $C$ then $g_x$ is in $C$ for any $x$, which shows that $g$ is continuous in $y$. Since $\bar{g}$ preserves directed joins, applying $\bar{g}(\sqcup_i x_i)$ and $\sqcup_i \bar{g}(x_i)$ to $y$ for $\{x_i\}$ directed shows that $g$ is continuous in $x$. To see that $g$ is continuous, suppose $\{\langle x_i, y_i \rangle\}$ is a directed set in $X \times Y$; then $\sqcup_i \langle x_i, y_i \rangle = \langle \sqcup_i x_i, \sqcup_j y_j \rangle$. By continuity in each argument, $f(\sqcup_i x_i, \sqcup_j y_j) = \sqcup_{i,j} f(x_i, y_j)$. Since $\{\langle x_i, y_i \rangle\}$ is directed, and $f$ being monotone in each argument) is monotone, $\sqcup_{i,j} f(x_i, y_j) \leq \sqcup_i f(x_i, y_i)$. It has been shown that $f(\sqcup_i \langle x_i, y_i \rangle) \leq \sqcup_i f(x_i, y_i)$, and the opposite inequality follows by monotonicity. Having shown that the correspondence between $g$ and $\bar{g}$ restricts to $C$, the theorem is proved.

The argument that continuity in each argument implies joint continuity does not require that the Scott topology in the product order be the product of the Scott topologies. Indeed this need not hold; see [Gierz et al].

If a poset $P \in \text{PoSet}_{\mathcal{D},\cup}$ is closed under finite joins it is a complete lattice. The full subcategory of $\text{PoSet}_{\mathcal{D},\cup}$ of the complete lattices, which may be denoted $\text{Clat}_{\mathcal{D},\cup}$, is also of interest, especially some subcategories which will be considered below. One can verify that theorem 5 holds, colimits exist, and the category is Cartesian closed.

Finally, some authors consider the posets, with maps preserving directed joins which exist.

3. **Algebraic lattices.** Let $P$ be a poset in $\text{PoSet}_{\mathcal{D},\cup}$. An element $a \in P$ is said to be compact if whenever $a \leq \sqcup S$ for some directed $S \subseteq P$, there is an $f \in S$ with $a \leq f$. If $P$ has all joins, a is compact iff whenever $a \leq \sqcup S$ for some $S \subseteq P$, there is a finite $F \subseteq S$ with $a \leq \sqcup F$, as is easily seen. The similarity to the definition of compactness in topological spaces is more than superficial, as will be seen.

For $x \in P$ let $K_x$ denote the set of compact $a$ with $a \leq x$. $P \in \text{PoSet}_{\mathcal{D},\cup}$ is said to be an algebraic poset if for every $x \in P$, $K_x$ is directed and $x = \sqcup K_x$. We use AlgPos to denote the algebraic posets, considered as a subcategory of $\text{PoSet}_{\mathcal{D},\cup}$.

For the following it is convenient to generalize the notation $x^\leq$; for a poset $P$ and $S \subseteq P$ let $S^\leq$ denote $\sqcup_{x \in S} x^\leq$, the “downward closure” of $S$. No confusion will arise with the notation $R^\leq$ for ordered rings.

268
Theorem 9. Suppose $P \in \text{PoSet}$, and let $I$ be the ideals, ordered by inclusion. Let $G$ denote the forgetful functor from $\text{AlgPos}$ to $\text{PoSet}$.

a. $I \subseteq \text{AlgPos}$; the compact elements are the principal ideals.

b. The map $x \mapsto x^\leq$ from $P$ to $I$ is a universal arrow from $P$ to $G$.

c. The map $P \mapsto I$ is the object map of a left adjoint $\text{Idl}$ to $G$; for $f : P \mapsto Q \text{Idl}(f)$ maps the ideal $J$ to $f[J]^\leq$, which is the ideal generated by $f[J]$.

d. If $A \in \text{AlgPos}$ let $K_A$ be the compact elements with the inherited order, and $I_A$ the ideals of $K_A$; then the map $x \mapsto K_x$ is an order isomorphism from $A$ to $I_A$.

Proof: Suppose $S \subseteq I$ is directed. If $x \subseteq \bigcup S$ and $y \leq x$ then $x \in J$ for some $J \in S$, whence $y \in J$, whence $y \subseteq \bigcup S$. If $x_1, x_2 \in \bigcup S$ then for $i = 1, 2$, $x_i \in J_i$ for some $J_i \in S$, whence for some $J \in S$ both $x_1, x_2 \in J$, whence $x_1 \cup x_2 \in J$, whence $x_1 \cup x_2 \subseteq \bigcup S$. Thus, $\bigcup S$ is an ideal, so $I$ is closed under directed sups. If $x^\leq \subseteq \bigcup S$ where $S \subseteq I$ then $x \in J$ for some $J \in I$, so $x^\leq \subseteq J$. Suppose $K \in I$ is compact; since $K \subseteq \bigcup_{x \in K} x^\leq$, $K \subseteq x_1^\leq \cup \cdots \cup x_k^\leq$ for some $x_1, \ldots, x_k \in K$; it follows that $K = (x_1 \cup \cdots \cup x_k)^\leq$. Thus, the compact elements are the principal ideals. If $J \in I$, $K_J$ is the principal ideals contained in $J$; this is directed because $J$ is, and clearly $J = \bigcup K_J$. This completes the proof of part a. Suppose $f : P \mapsto G(A)$ where $A \in \text{AlgPos}$. As is readily verified, if $S$ is directed and $f$ is order preserving then $f[S]$ is directed; thus, for $J \in I$, $f[J]$ is directed, and so $g : I \mapsto A$ may be defined by $g(J) = \bigcup f[J]$. Since for directed $S \subseteq I$,
\[ g(\bigcup S) = \bigcup_{x \in \bigcup S} f(x) = \bigcup_{J \in S} \bigcup_{x \in J} f(x) = \bigcup_{J \in S} g(J), \]
g is a morphism. Since $f(w) \leq f(x)$ if $w \leq x$, in $A \cup_{w \leq x} f(w) = f(x)$. This shows that $f = G(g)e$ where $e(x) = x^\leq$. On the other hand if $g$ is a morphism with $f = G(g)e$, then $g$ must be as given. This completes the proof of part b. Part c follows by theorem 13.2, noting that the image of $J$ under the induced map is $\bigcup_{J \in J} f(j)^\leq$. Clearly, if $x \leq y$ then $K_x \subseteq K_y$; and since $x = \bigcup K_x$, if $K_x \subseteq K_y$ then $x \leq y$. In particular $x \mapsto K_x$ is injective. Suppose $J \in I_A$. Let $x = \bigcup J$; $J \subseteq K_x$ because $J \subseteq K_A \subseteq x^\leq$. If $w \in K_x$ then $w \leq \bigcup J$, and since $w$ is compact and $J$ is directed $w \in J$. That is, $I = K_x$; thus $x \mapsto K_x$ is surjective. This completes the proof of part d.

Suppose $F : C \mapsto D$ is the left adjoint to $G : D \mapsto C$, $C'$ is a full subcategory of $C$, $D'$ is a full subcategory of $D$, $F[C'] \subseteq D'$, and $G[D'] \subseteq C'$. The induced functor $F' : C' \mapsto D'$ is readily verified to be the left adjoint to $G' : D' \mapsto C'$, with the Hom set bijections being the same. Further, the unit is the same, i.e., for $c' \in C'$ $(G'F')(c') \in C'$ and the unit map is the same; and similarly for the counit.

The algebraic posets which are complete lattices are called algebraic lattices. The full subcategory of $\text{AlgPos}$ consisting of these will be denoted $\text{AlgLat}$. Recall that the closure systems on a set are in bijective correspondence with the closure operators; and a closure system is algebraic iff the closure operator preserves directed unions. Also recall that for $L \in \text{SLat}_{\leq,0}$, the ideals of $L$ form an algebraic closure system.

Theorem 10.

a. If $A \subseteq \text{Pow}(S)$ is an algebraic closure system then $A$ is an algebraic lattice, whose compact elements are the closures of the finite sets.

b. If $K$ is the compact elements of an algebraic lattice with the inherited order then $K \in \text{SLat}_{\leq,0}$.

c. Theorem 9 holds, with $\text{PoSet}$ replaced by $\text{SLat}_{\leq,0}$, and $\text{AlgPos}$ by $\text{AlgLat}$.

Proof: Suppose $F$ is finite and $K = F^\leq$. Suppose $D \subseteq A$ is directed and $K \subseteq \bigcup D$. Then $F \subseteq \bigcup D$, so $F \subseteq G$ for some $G \subseteq D$, so $K = F^\leq \subseteq G$. Hence, $K$ is a compact element of $A$. Given $X \in A$, $X = \bigcup \{F \subseteq X : F \text{ is finite}\}$, whence $X = \bigcup \{F^\leq : F \subseteq X, F \text{ finite}\}$. Hence, $A$ is algebraic. In fact, $X = \bigcup \{F^\leq : F \subseteq X, F \text{ finite}\}$, and the latter set is directed since $F^\leq \cup G^\leq \subseteq (F \cup G)^\leq$. If $X$ is compact, it follows that $X = F^\leq$ for some finite $F$. This proves part a. Clearly $0 \in K$. If $x_1, x_2 \in K$ and $x_1 \cup x_2 \leq \bigcup S$
for some $S \subseteq L$, then for $i = 1, 2$, $x_i \leq \sqcup F_i$ for some finite $F_i \subseteq S_i$; and so $x_1 \sqcup x_2 \leq \sqcup (F_1 \sqcup F_2)$. This shows that $x_1 \sqcup x_2 \in K$; and so $K$ is closed under join. This proves part b. That $I \in \text{AlgLat}$ follows by part a. The proof of the remainder of theorem 9 is virtually unchanged; indeed some claims follow by remarks above.

The algebraic lattices may also be considered as the objects of a subcategory of $\text{CLat}^D$, which we denote $\text{AlgLat}^D$. We state the following without proof.

- The Cartesian product of algebraic lattices is algebraic ([Gierz et al], I-14.4).
- The Cartesian product is a product in either $\text{AlgLat}$ or $\text{AlgLat}^D$.
- In $\text{CLat}^D$, the algebraic lattices are closed under subobject.
- Theorem 5 holds in $\text{AlgLat}^D$.
- $\text{AlgLat}$ is Cartesian closed.
- $\text{AlgPos}$ is not Cartesian closed; for example $\text{Hom}(\mathcal{Z}^<, \mathcal{Z}^<)$ has no compact elements.

4. Stone Duality. An equivalence of categories is an adjunction in which the unit and counit are natural equivalences. The functors are not inverse, but are “up to isomorphism”. In particular, if the images of two objects are isomorphic then the objects are.

**Lemma 11.** Suppose $F$ is a functor from $A$ to $B$, $G$ is a functor from $B$ to $A$, and $FG$ and $GF$ are naturally equivalent to the identity functor. Then $F$ and $G$ are faithful, full, and both left and right adjoints to each other.

**Proof:** Let $\alpha$ be the natural equivalence in $A$ from $GF$ to the identity. Suppose $f : a_1 \mapsto a_2$ is an arrow in $A$; the equation $f \alpha_{a_1} = \alpha_{a_2}GF(f)$ shows that $f$ is determined by $F(f)$, i.e., $F$ is faithful. Symmetrically $G$ is faithful. Given an arrow $g : F(a_1) \mapsto F(a_2)$ in $B$, let $f = \alpha_{a_2}G(g)\alpha^{-1}_{a_1}$. The equation $f \alpha_{a_1} = \alpha_{a_2}h$ holds with $h$ either $G(g)$ or $GF(f)$; by uniqueness $G(g) = GF(f)$, and since $G$ is faithful $F(f) = g$. The map $g \mapsto \alpha_{a_2}G(g)$ is a bijection from $\text{Hom}(b, F(a))$ to $\text{Hom}(G(b), a)$, which is readily seen to be natural. The remaining claims follow by symmetry.

The open sets of a topological space, under the subset order, constitute a frame, since the required distributive law is inherited from the power set. If $f$ is a continuous function between topological spaces, the function $U \mapsto f^{-1}[U]$ is a morphism between the frames, in the opposite direction. We let $\Omega$ denote this functor from Top to $\text{Frm}^\text{op}$.

Stone duality is concerned with an equivalence of categories between $\text{DLat}_B$ and a subcategory of Top. In more recent developments, it was noticed that this equivalence can be “factored through” a subcategory of $\text{Frm}^\text{op}$. The phrase “pointless topology” has been used to describe investigating $\text{Frm}^\text{op}$ as an alternative to Top. $\text{Frm}^\text{op}$ is given a name, the category of locales; but we will use the former. This “factorization” serves to organize the argument, and leads to facts of interest in themselves.

For $L \in \text{Lat}_B$, an ideal is prime if and only if it equals $h^{-1}(0)$ for some homomorphism $h : L \mapsto C_2$ where $C_2$ is the two element chain. It was observed above that $h^{-1}(0)$ is an ideal for any $h$; if the codomain is $C_2$ and $h(x) \sqcap h(y) = 0$ then $h(x) = 0$ or $h(y) = 0$, and so $h^{-1}(0)$ is prime. On the other hand if $I$ is a prime ideal then the map where $h(x) = 0$ for $x \in I$, else 1, is a homomorphism to $C_2$. The dual statements hold for filters and $h^{-1}(1)$.

As usual $\text{Hom}_{\text{Lat}_B}(-, C_2)$ is a functor from $\text{Lat}^\text{op}_B$ to $\text{Set}$. It may be “enriched” to a functor to Top. For $A \in \text{Lat}_B$ and $a \in A$ let $\beta^A_a$ denote $\{f \in \text{Hom}(A, C_2) : f(a) = 1\}$; $A$ may be omitted if clear from context. In terms of prime filters (ideals), $\beta_a$ is the prime filters (ideals) containing (not containing) $a$.

**Lemma 12.** Suppose $A, B \in \text{Lat}_B$ and $f : A \mapsto B$ is a morphism. Let $\beta^A_a$ be as above.

a. For $a, b \in A$, $\beta_a \cap \beta_b = \beta_{a \sqcap b}$.

b. $\{\beta_a\}$ is the base for a topology on $\text{Hom}(A, C_2)$. 

270
c. \((f^R)^{-1}(\beta_A^1) = \beta_B^{f(a)}\).

d. \((f^R)^{-1}\) is continuous with the topologies of part b.

**Proof:** For part a, \(f \in \beta_{a \cap b}\) iff \(f(a \cap b) = 1\) iff \(f(a) \cap f(b) = 1\) iff \(f(a) = 1\) and \(f(b) = 1\) iff \(f \in \beta_a \cap \beta_b\).

Part b is immediate. For part c, \(g \in (f^R)^{-1}(\beta_A^1)\) iff \((g \circ f)(a) = 1\) iff \(g \in \beta_{f(a)}^B\). Part d is immediate.

We denote the functor from \(\text{Lat}^{op}_B\) to \(\text{Top}\) of lemma 12 as \(\text{Spec}\) (the “prime spectrum” of the lattice).

Similarly to part a of the lemma, \(\beta_{a \cup b} = \beta_a \cup \beta_b\). Noting that \(\beta_0 = \emptyset\) and \(\beta_1 = \text{Hom}(A, C_2)\), the map \(a \mapsto \beta_a\) is a \(\text{Lat}_B\) homomorphism from \(A\) to \(\Omega(\text{Spec}(A))\). By lemma 11.16, if \(A\) is distributive the map is injective; for if \(a \neq b\) then \(\beta_a\) cannot equal \(\beta_b\), since there is a prime ideal containing \(b\) but not \(a\).

The Spec functor is of greater significance for \(\text{DLat}_B\) then for \(\text{Lat}_B\); indeed representation theorems for \(\text{Lat}_B\) involve categories further enriched from \(\text{Top}\). The following lemma gives further properties which hold on \(\text{DLat}_B\). Note that it concerns compact open sets, which are not common in ordinary topology (e.g., they don’t exist in Hausdorff spaces).

**Lemma 13.** If \(A \in \text{DLat}_B\), an open set in \(\text{Spec}(A)\) is compact iff it is \(\beta_a\) for some \(a \in A\).

**Proof:** Suppose \(\beta_a \subseteq \bigcup_{b \in S} \beta_b\), i.e., if \(f(a) = 1\) then for some \(b \in S\) \(f(b) = 1\). Let \(I\) be the ideal generated by \(S\) and let \(F\) be the filter \(a^\perp\). By lemma 11.16, if \(I \cap F = \emptyset\) then there is an \(f \in \text{Hom}(A, C_2)\) such that \(f(b) = 0\) for all \(b \in I\) and \(f(a) = 1\). Thus, \(I \cap F \neq \emptyset\), which is to say \(a \leq b_1 \cup \cdots \cup b_k\) for some \(b_1, \ldots, b_k \in S\). It follows that \(\beta_a \subseteq \bigcup_i \beta_{b_i}\). This shows that \(\beta_a\) is compact. If \(K \in \text{Hom}(A, C_2)\) is compact open, say \(K = \bigcup_{b \in S} \beta_b\) where \(S \subseteq A\), then \(K \subseteq \bigcup_i \beta_{b_i}\) for some \(b_1, \ldots, b_k \in S\); and \(K = \beta_{b_1 \cup \cdots \cup b_k}\) follows.

As mentioned above, it is convenient to present some facts concerning Frm, before proceeding with Stone duality itself. Firstly, \(\text{Hom}_{\text{frm}}(-, C_2)\) may also be enriched to \(\text{Top}\). If \(h : A \rightarrow C_2\) then \(h^{-1}(0)\) is closed under arbitrary joins, and so is principal; and conversely a principal prime ideal yields a homomorphism to \(C_2\). Also (taking \(\beta_a\) as the set of frame homomorphisms), for \(S \subseteq A\), \(\beta_{\bigcup S} = \bigcup_{a \in S} \beta_a\), and it follows that \(\{\beta_a\}\) is a topology rather than merely a base. The resulting functor from \(\text{Frm}^{\text{op}}\) to \(\text{Top}\) is denoted \(\text{Pt}\). The map \(a \mapsto \beta_a\) from \(A\) to \(\Omega(\text{Pt}(A))\) is a surjective frame homomorphism, but not necessarily injective; \(A\) is said to be spatial if it is (i.e., if \(A\) “has enough points”). Note finally that by the same argument as lemma 13, the open sets \(\beta_a\) are compact.

A topological space (or as usual subspace) is said to be irreducible if it is nonempty and is not the union of two proper closed subsets. It is readily seen that a closed set \(K\) in a topological space \(X\) is irreducible iff, whenever \(K \subseteq K_1 \cup K_2\) for closed \(K_1, K_2\), either \(K \subseteq K_1\) or \(K \subseteq K_2\). This in turn is the case iff \(K^c\) is a prime element in the lattice \(\Omega(X)\).

It is easily seen that for a topological space \(X\), and \(x \in X\), \(\{x\}^c\) is an irreducible closed set. \(X\) is said to be sober if the map \(x \mapsto \{x\}^c\) is a bijection from \(X\) to the irreducible closed sets. It is also easily seen that \(X\) is \(T_0\) iff the map \(x \mapsto \{x\}^c\) is injective. If \(X\) is \(T_2\), then any two points in a closed set can be separated, and it follows that the only irreducible closed sets are the singletons. Thus, \(T_2\) spaces are sober, and sober spaces are \(T_0\).

Spaces with nontrivial irreducible closed sets are another example of spaces with properties not common in ordinary topology. One example is any infinite set \(X\), with \(X\) and the finite subsets as the closed sets. \(X\) is irreducible; the space is \(T_1\) and so \(\{x\}^c = \{x\}\), and so the space is not sober. An example of a space which is sober but not \(T_1\) is \(\{0, 1\}, \{\{1, 2\}\}\).

One final piece of notation is useful; given a topological space \(X\) and \(x \in X\) let \(e_x\) denote the map from a one element space to \(X\), whose range is \(\{x\}\). Then \(\Omega(e_x)\) is an element of \(\text{Pt}(\Omega(X))\), which takes value 1 on an open set \(U\) iff \(x \in U\).

**Theorem 14.**

a. \(\text{Pt}\) is a right adjoint to \(\Omega\), with the counit component \(Y \mapsto \Omega(\text{Pt}(Y))\) mapping \(y \mapsto \beta_y\).
b. The counit component for \( Y \in \text{Frm} \) is an isomorphism iff \( Y \) is spatial.

c. \( \Omega(X) \) is spatial.

d. The unit component \( X \mapsto \text{Pt}(\Omega(X)) \) maps \( x \mapsto \Omega(\epsilon_x) \).

e. The unit component for \( X \in \text{Top} \) is an isomorphism iff \( X \) is sober.

f. \( \text{Pt}(Y) \) is sober.

g. The adjunction induced on the spatial frames and sober spaces is an equivalence of categories.

PROOF: By remarks above \( y \mapsto \beta_y \) is a frame morphism. Given \( f : Y \mapsto \Omega(X) \) let \( g : X \mapsto \text{Pt}(Y) \) be defined by \( g(x)(y) = 1 \) iff \( x \in f(y) \); then \( g \) is the unique function such that \( g^{-1}[\beta_y] = f(y) \), i.e., \( \Omega(g)(\beta_y) = f(y) \). Since \( g(x) = \Omega(\epsilon_x) \circ f \), \( g(x) \) is a morphism in \( \text{Frm} \). Further since \( g^{-1}[\beta_y] = f(y) \), \( g \) is continuous. This establishes the required universality property for part a. Part b follows by the definition given above of spatial frames. For part c, suppose \( V, W \) are distinct members of \( \Omega(X) \); choose some \( x \in X \) in one but not the other, and let \( p \) be such that \( p(U) = 1 \) iff \( x \in U \). This shows that \( b_V \neq b_W \). Given \( f : X \mapsto \text{Pt}(Y) \) let \( g : Y \mapsto \Omega(X) \) be defined by \( x \in g(y) \) iff \( f(y) = 1 \); then \( g \) is the unique function such that \( \Omega(\epsilon_x) \circ g = f(x) \). Part d follows since \( \text{Pt} \) is the adjoint. For part e, since \( \{x\}^\text{cl} = \{U : \Omega(\epsilon_x)\{U\} = 1\} \), \( x \mapsto \{x\}^\text{cl} \) is bijective iff \( x \mapsto \Omega(\epsilon_x) \) is. For part f, it must be shown that the map \( p \mapsto \Omega(\epsilon_p) \) is a bijection from \( \text{Pt}(Y) \) to \( \text{Pt}(\Omega(\text{Pt}(Y))) \); note that \( \Omega(\epsilon_p)(\beta_y) = 1 \) iff \( p(y) = 1 \). The map is an injection by the triangular identities. Suppose \( p \in \text{Pt}(\Omega(\text{Pt}(Y))) \); let \( p \in \text{Pt}(Y) \) be defined by \( p(y) = P(\beta_y) \). Then \( P(\beta_y) = p(y) = \Omega(\epsilon_p)(\beta_y) \) for all \( y \in Y \), whence \( P = \Omega(\epsilon_p) \); this shows that the map is surjective. Finally, part g is a consequence of the preceding parts.

Define a frame to be coherent if it is algebraic, and the compact elements contain 1 and are closed under \( \sqcup \). (An algebraic lattice is called arithmetic if the compact elements comprise a sub-lattice; thus, a frame is coherent iff it is arithmetic and 1 is compact). Let \( \text{CohFrm} \) be the category whose objects are the coherent frames, and whose morphisms are the frame morphisms which preserve compact elements.

THEOREM 15.

a. For \( L \in \text{DLat}_B \) \( \text{Idl}(L) \in \text{Frm} \).

b. If \( f \) is a morphism in \( \text{DLat}_B \) then \( \text{Idl}(f) \) is a morphism in \( \text{Frm} \).

c. The functor \( \text{Idl} \) from \( \text{DLat}_B \) to \( \text{Frm} \) is the left adjoint to the forgetful functor from \( \text{Frm} \) to \( \text{DLat}_B \); the unit component is \( x \mapsto x^\leq \).

d. In fact, for \( L \in \text{DLat}_B \) \( \text{Idl}(L) \in \text{CohFrm} \), and if \( f \) is a morphism in \( \text{DLat}_B \) then \( \text{Idl}(f) \) is a morphism in \( \text{CohFrm} \).

e. For \( A \in \text{CohFrm} \) let \( \text{Comp}(A) \) denote the compact elements. The map \( A \mapsto \text{Comp}(A) \) is the object map of a functor from \( \text{CohFrm} \) to \( \text{DLat}_B \). For \( f \) a morphism of \( \text{CohFrm} \) \( \text{Comp}(f) \) is just the restriction.

f. \( \text{Idl} \) (considered as a functor from \( \text{CohFrm} \)) is the left adjoint to \( \text{Comp} \). The unit component is \( x \mapsto x^\leq \), and is an isomorphism.

g. With \( K_x \) as in section 3, the map \( x \mapsto K_x \) from \( A \in \text{CohFrm} \) to \( \text{Idl}(\text{Comp}(A)) \) is an isomorphism in \( \text{CohFrm} \).

h. The adjunction of part f is an equivalence of categories.

PROOF: Suppose \( I \in \text{Idl}(L) \) and \( S \subseteq \text{Idl}(L) \); \( \cup J \cap J \subseteq \cap (\cup S) \) in any lattice. Suppose \( x \in J \subseteq I \cap (\cup S) \), whence \( x \in I \) and \( x \leq y_1 \cup \cdots \cup y_k \) where \( y_i \in J_i \) for some \( J_i \in S \). Since
\[
x = x \cap (y_1 \cup \cdots \cup y_k) = (x \cap y_1) \cup \cdots \cup (x \cap y_k),
\]
\( x \in \cup J \cap J \). This proves part a. For part b it must be shown that \( f[J_1] \subseteq \cap f[J_2] \subseteq f[J_1 \cap J_2] \subseteq \cap f[S] \subseteq \cup J \cap J \); these are routine and left to the reader. For part c, it suffices to show that the map \( g \) in the proof of theorem 9b is a morphism in \( \text{Frm} \), that is, \( \cup f[J_1 \cap J_2] = (\cup f[J_1]) \cap (\cup f[J_2]) \), and \( \cup f[S] = \cup f[J] \); these are routine and left to the reader. From theorem 9 \( \text{Idl}(L) \) is an algebraic
lattice, whose compact elements are the principal ideals; and from part a $\text{Idl}(L) \in \text{Frm}$. $L$ is compact since it has a greatest element, and closed under $\sqcup$ since $x \leq y \leq (x \sqcup y) \leq z$; thus, $\text{Idl}(L) \in \text{CohFrm}$. Even on \text{PoSet} $\text{Idl}(f)$ preserves principal ideals, and part d is proved. If $A \in \text{AlgPos}$, $x_1, x_2 \in A$ are compact, and $x_1 \sqcup x_2 \leq \sqcup S$ for some $S \subseteq A$, then for $i = 1, 2$ $x_i \leq \sqcup F_i$ for some finite $F_i \subseteq S$, and $x_1 \sqcup x_2 \leq \sqcup (F_1 \cup F_2)$. Thus, the compact elements are closed under $\sqcup$; clearly 0 is compact, and all other requirements for part e are immediate from the definitions. The map $g$ of theorem 9b has already been shown to be a morphism in $\text{Frm}$; further $g(x^\leq) = f(x)$, so $g$ maps compact elements to compact elements, and is a morphism in $\text{CohFrm}$. It is determined on the principal ideals by the requirement $f = \text{Comp}(g)e$ where $e(x) = x^\leq$; and since it is a frame homomorphism it is determined completely. The map $x \mapsto x^\leq$ is clearly injective, and also clearly surjective to $\text{Comp}(\text{Idl}(L))$ for $L \in \text{DLat}_B$. This proves part f. Since the compact elements are closed under $\sqcap$, $K_x$ is an ideal. The map $J \mapsto \sqcup J$ is a two-sided inverse to $x \mapsto K_x$, proving part g; indeed $\sqcup K_x = x$ by hypothesis. On the other hand clearly $J \subseteq K_{\sqcup J}$; and if $k \in K_{\sqcup J}$ then $k \leq \sqcup J$, and since $J$ is directed $k \leq j$ for some $j \in J$, and $k \in J$. Part h follows by lemma 11, from the fact that the isomorphism of part g is natural; this follows because for a morphism $f : A \to B$ in $\text{CohFrm}$ and $a \in A$, $K_{f(a)} = f[K_a]^\leq$ (where the right side is taken in $\text{Comp}(B)$). To see this, if $x \leq f(k)$ where $k \leq a$ and $x \in \text{Comp}(B)$ then $x \in K_{f(a)}$. Now, $a = \sqcup K_a$, whence $f(a) = \sqcup f[K_a]$, so if $x \in K_{f(a)}$ then $x \leq f(k)$ for some $k \leq a$.

By part c, given a morphism $p : L \to C_2$ in $\text{DLat}_B$ there is a unique morphism $P : \text{Idl}(L) \to C_2$ in $\text{Frm}$ with $p(x^\leq) = P(x^\leq)$ for $x \in L$. This establishes a natural bijection from $\text{Spec}(L)$ to $\text{Pt}(L)$. Since for $I \in \text{Idl}(L)$ $I = \sqcup_{x \in I} x^\leq$, and $P$ is a frame homomorphism, one sees that $P(I) = \sqcup p[I]$. If $I \neq J$ are ideals of $L$, suppose $x \in I - J$ (the case $x \in J - I$ being symmetric); then $x^\leq$ is disjoint from $J$, so there is a $p \in \text{Spec}(L)$ where $p(x) = 1$ and $p(w) = 0$ for $w \in J$. For the corresponding $P$, $P(I) = 1$ and $P(J) = 0$. Thus, $\text{Idl}(L)$ is spatial. It follows by part g that any coherent frame is spatial.

Say that a topological space $X$ is coherent if
- it is sober;
- the compact open sets are a base;
- it is compact; and
- the compact open sets are closed under $\sqcap$.

Let $\text{CohTop}$ be the category whose objects are the coherent spaces, and whose morphisms are the continuous functions where for $K$ compact open $f^{-1}[K]$ is compact. It is straightforward to verify that an open subset of a topological space $X$ is compact iff it is a compact element of the frame $\Omega(X)$. Thus, a space is coherent if it is sober and $\Omega(X)$ is a coherent frame. By definition if $f$ is a morphism in $\text{CohTop}$ then $\Omega(f)$ is a morphism in $\text{CohFrm}^{op}$. Thus, $\Omega$ may be considered as a functor from $\text{CohTop}$ to $\text{CohFrm}^{op}$.

On the other hand if $Y \in \text{CohFrm}$ then $\text{Pt}(Y)$ is sober by theorem 14.f. As already observed $Y$ is spatial, and so by theorem 14b $\Omega(\text{Pt}(Y))$ is isomorphic to $Y$. Thus, $\text{Pt}(Y) \in \text{CohTop}$. For $f$ a morphism in $\text{CohFrm}$ $\text{Pt}(f)$ is a morphism in $\text{CohTop}$ because lemma 12.c holds in $\text{Frm}$.

Recall the notation $e_x$ used in theorem 14.

**Theorem 16.** $\text{Pt}$, considered as a functor from $\text{CohFrm}^{op}$ to $\text{CohTop}$, is the right adjoint to the functor $\Omega$, considered as a functor from $\text{CohTop}$ to $\text{CohFrm}^{op}$. The counit component maps $y \mapsto \beta_y$, and is an isomorphism. The unit component maps $x \mapsto \Omega(e_x)$, and is an isomorphism.

**Proof:** The function $g$ of the proof of theorem 14.a is a $\text{CohFrm}$ morphism, whence the counit is as claimed. The remainder of the theorem follows as in theorem 14.

The above correspondence $p \mapsto P$ from $\text{Spec}(L)$ to $\text{Pt}(L)$ may be verified to be a homeomorphism. Indeed, the image of $\beta_x$ is $\beta_{x^\leq}$, so the map is open; and the inverse image of $\beta_J$ is $\sqcup_{x \in J} \beta_x$, so the map is continuous. In particular, for $L \in \text{DLat}_B$ $\text{Spec}(L) \in \text{CohTop}$. $\text{Spec}$ also maps morphisms of $\text{DLat}_B$ to
morphisms of CohTop, again by lemma 12.c. Comp ◦ Ω from CohTop to DLat_{B} takes a coherent space to its lattice of compact subspaces; we may use Comp to denote this functor as well.

**Theorem 17.** Spec, considered as a functor from DLat_{op} to CohTop, is the right adjoint to the functor Comp, considered as a functor from CohTop to CohFrm_{op}. The counit component maps \( y \mapsto \beta_y \), and is an isomorphism. The unit component maps \( x \mapsto \text{Comp}(e_x) \), and is an isomorphism.

**Proof:** Using \( \alpha_{LY}(h) = \text{Comp}(h)\mu_L \) (see section 13.4) with \( h \) being Comp\((h')\) and \( h' : J \mapsto \beta_J \), one sees that the counit of the composed equivalences of theorems 15 and 16 is \( x \mapsto \beta_x \). Composing this with the above natural isomorphism yields that \( x \mapsto \beta_x \) is the counit. The claim for the unit follows similarly.

Recall that a topological space \( X \) is totally disconnected if, for every \( x \) and \( y \) in \( X \) there are disjoint open sets \( U \) and \( V \) such that \( x \in U \), \( y \in V \), and \( U \cup V = X \). By theorem 17.23 and remarks preceding it, a compact Hausdorff space is totally disconnected if it has a base of clopen sets. Also, in a compact Hausdorff space a subspace is closed iff it is compact.

**Theorem 18.** A coherent space is Hausdorff iff it is totally disconnected. A space is coherent and Hausdorff iff it is compact and totally disconnected.

**Proof:** If \( X \) is coherent and Hausdorff it is compact, and has a basis of compact open sets, whence of clopen sets; so \( X \) is totally disconnected. That totally disconnected spaces are Hausdorff has already been observed. The first claim establishes that a coherent Hausdorff space is compact and totally disconnected. A compact totally disconnected space is Hausdorff, hence sober. The compact open sets (being the clopen sets) form a base; they are also closed under intersection.

The coherent Hausdorff (compact totally disconnected) spaces are known by various other names, in particular Stone spaces; we use CohHTop to denote the full subcategory of CohTop of these spaces. If \( f : X \mapsto Y \) where \( X, Y \in \text{CohHTop} \) is continuous, and \( K \subseteq Y \) is compact open, then \( K \) is clopen, so \( f^{-1}[K] \) is clopen, so \( f^{-1}[K] \) is compact open. Thus, CohHTop is a full subcategory of Top.

**Theorem 19.**

a. If \( L \in \text{Bool} \) then \( \text{Spec}(L) \in \text{CohHTop} \).

b. If \( X \in \text{CohHTop} \) then \( \text{Comp}(X) \in \text{Bool} \).

c. The equivalence of categories of theorem 17 restricts to an equivalence of categories between Bool and CohHTop.

d. The implications of parts a and b are in fact equivalences.

**Proof:** Suppose \( L \in \text{Bool} \) and \( p \) and \( q \) are distinct elements of \( \text{Spec}(L) \). Let \( x \) be such that \( p(x) \neq q(x) \), say \( p(x) = 1 \) (the case \( q(x) = 1 \) being symmetric). Then \( q(x^c) = 1 \), so \( x^c \) and \( (x^c)^c \) separate \( p \) and \( q \). This proves part a. If \( X \in \text{CohHTop} \) then the compact open sets are the clopen sets, and part b follows. Part c is immediate, and part d follows because the unit and counit components are isomorphisms.

Suppose \( U \) is an open set and \( K \) a closed set in a topological space. Then \( U \subseteq U^\text{cl} \), whence \( U \subseteq (U^\text{cl})^\text{int} \); and \( K^\text{int} \subseteq K \), whence \( (K^\text{int})^\text{cl} \subseteq K \). Thus \( \text{cl} \) on the open sets and \( \text{int} \) on the closed sets (both ordered by inclusion) form a Galois adjunction. An open set is said to be regular open if \( (U^\text{cl})^\text{int} = U \). Recall (exercise 17.3.b) that for \( S \) a subset of a topological space, \( (S^\text{int})^c = (S^c)^\text{cl} \).

In the Heyting algebra \( \Omega(X) \) of a topological space \( X \), \( U^p = (U^c)^\text{int} \) since by definition this is the largest open set contained in \( U^c \). If \( U \) is regular then \( V^p = U \) where \( V = (U^\text{cl})^c \), and if \( V^p = U \) for some open \( V \) then \( U \) is regular (these are easily shown using the preceding paragraph). That is, the regular open sets are exactly the sets \( V^p \), which as shown above form a Boolean algebra (sometimes called the regular elements of the Heyting algebra). Since \( \Omega(X) \) is complete the Boolean algebra of regular open sets is.
The members of $\Omega(X)$ which have a complement are the clopen sets. By exercise 2 these are the regular open sets iff the clopen sets are closed under union. If $X$ is coherent Hausdorff, and the clopen sets form a Boolean algebra, then they are closed under union (exercise 8). It follows that if $B$ is a complete Boolean algebra then $B$ is isomorphic to the regular open sets of $\text{Spec}(B)$.

By theorem 7 $\text{Hom}_{\text{Frm}}(X, C_2)$ may be considered an object in $\text{PoSet}_{DL\cup}$. If $\{p_i\}$ is a directed subset, and $f : X \rightarrow Y$ is a frame homomorphism, then $\sqcup_i p_i f$ equals $(\sqcup_i p_i) f$, as may be seen pointwise. Also, $\{p_i f\}$ is readily seen to be directed, and it follows that $f^R$ preserves directed joins, so $\text{Hom}( -, C_2)$ may be considered as a functor from $\text{Frm}^{\text{op}}$ to $\text{PoSet}_{DL\cup}$, which we denote $\Pi$.

If $X \in \text{PoSet}_{DL\cup}$ then $X$ equipped with the Scott topology is an object in $\text{Top}$. Further, by theorem 6 an arrow in $\text{PoSet}_{DL\cup}$ may be considered an arrow in $\text{Top}$. We use $\Sigma$ to denote this functor from $\text{PoSet}_{DL\cup}$ to $\text{Top}$. We use $\Sigma$ to denote $\Omega \circ \text{Set}$.

**Theorem 20.** $\Pi$ is a right adjoint to $\Sigma$. The counit component $Y \mapsto \Sigma(\Pi(Y))$ maps $y \mapsto \beta_y$. The unit component $X \mapsto \Pi(\Sigma(X))$ maps $x \mapsto (e_x)$, where $e_x$ maps the one element poset to $x \in X$.

**Proof:** First, $\beta_y$ is Scott open; indeed $\beta_y = \{ p : p(y) = 0 \}$ is clearly $\leq$-closed and closed under directed (indeed all) joins in $\Pi(Y)$, that is, is Scott closed. That $y \mapsto \beta_y$ is a frame morphism follows as in theorem 14. Given $f : Y \rightarrow \Sigma(X)$ let $g : X \rightarrow \Pi(Y)$ be defined as in theorem 14, by $g(x)(y) = 1$ iff $x \in f(y)$; as before $g$ is the unique function such that $g^{-1}(\beta_y) = f(y)$, i.e., $\Sigma(g)(\beta_y) = f(y)$. Since $g(x) = \Sigma(e_x) \circ f$, $g(x)$ is a morphism in $\text{Frm}$. Suppose $v \leq u$ for $u, v \in X$. If $g(u)(y) = 0$ then $u \neq f(y)$, so $v \neq f(y)$, so $g(v)(y) = 0$; that is, $g(v) \leq g(u)$. This shows that $g$ is monotone. Suppose $D \subseteq X$ is directed. Then $g(D)$ is directed, and $\sqcup g[D] \leq g(\sqcup D)$. If $g(d)(y) = 0$ for all $d \in D$ then $d \neq f(y)$ for all $d \in D$, whence $\sqcup D \neq f(y)$ because $f(y)$ is Scott open, whence $g(\sqcup D)(y) = 0$; that is, $g(\sqcup D) \leq g[D]$. Thus, $g$ preserves directed joins and is Scott continuous. This completes the proof of the required universality property for the counit. The proof that $x \mapsto \Sigma(e_x)$ is the component of the unit is virtually identical to the proof of the corresponding fact in theorem 14.

Many adjunctions involving lattices have been discovered. For some further examples, exercise 4.a shows that the $\leq$-closed sets of a semilattice $L \in \text{SLat}_{\geq 0}$, with the map $x \mapsto x^L$, yield a universal arrow (i.e., a left adjoint) to the forgetful functor from $\text{Frm}$ to $L \in \text{SLat}_{\geq 0}$. Exercise 9 shows that $\text{Hom}( -, C_2)$, considered as a functor from $\text{DLat}^{\text{op}}$ to $\text{PoSet}$, is right adjoint to the functor from $\text{PoSet}$ to $\text{DLat}^{\text{op}}$, which takes a poset to its lattice of $\leq$-closed sets, and an arrow $f : P \rightarrow Q$ in $\text{PoSet}$ to the map $S \mapsto f^{-1}[S]$. Finally, the functor $\text{Pt}$ may be defined on $\text{CLat}^F_{\leq}$ just as on $\text{Frm}$, and is right adjoint to $\Omega$.

Some aspects of the development in this section were taken from [Cheung].

5. **Continuous lattices.** Suppose $P \in \text{PoSet}_{DL\cup}$: for $x, y \in P$ let $x \ll y$ if, whenever $D$ is directed and $y \leq \sqcup D$ there is a $d \in D$ such that $x \leq d$. Thus, $x$ has a certain compactness property with respect to $y$. Various terminology is used for this relation, including “way below” and “compact relative to”. The following are readily verified.

- $x \ll y$ if, whenever $I$ is an ideal and $y \leq \sqcup I$ then $x \in I$.
- If $P$ has all joins then $x \ll y$ iff, whenever $S \subseteq P$ and $y \leq \sqcup S$ then there is a finite subset $F \subseteq S$ such that $x \leq \sqcup F$.
- $x$ is compact iff $x \ll x$.

If $x \ll y$ then $x \leq y$, since $D$ may be taken as $\{x\}$. If $x' \leq x$ and $x \ll y$ then $x' \ll y$, and if $y \leq y'$ and $x \ll y$ then $x \ll y'$; these are immediate from the definition.

We will use $x^{\ll}$ to denote $\{w : w \ll x\}$. $P$ is said to be continuous if for every $x \in P$, $x^{\ll}$ is directed and $x = \sqcup x^{\ll}$. An algebraic poset is continuous. If $x$ is compact and $x \leq y$ then $x \ll y$; on the other hand if $P$ is a continuous and $x \ll y$ implies $x$ is compact, then $P$ is algebraic. A simple example of a continuous poset which is not algebraic is the unit interval $[0,1]$ in $\mathcal{R}$ (exercise 10).
Let ContPos denote the full subcategory of PoSet\_{D⊔} of the continuous posets. The continuous posets which are complete lattices are called continuous lattices. The full subcategory of ContPos of these will be denoted ContLat. The continuous lattices may also be considered as a subcategory of of CLat^D; the resulting category will be denoted ContLat^D.

In a continuous lattice \( x \ll \) is an ideal.

The theory of continuous lattices benefits from a discussion of infinitary distributive laws. If \( S_j \) is a set of elements for each \( j \in J \), \( \bigcap_{j \in J} \cup S_j \) is an arbitrary meet of arbitrary joins. Let \( C \) be the set of functions \( f : J \mapsto \bigcup S_j \) such that \( f(j) \in S_j \) be the set of functions choosing an element of \( S_j \) for each \( j \in J \) (i.e., the product). A complete lattice is said to be completely distributive if \( \bigcap_{j \in J} \bigcup S_j = \bigcup_{f \in C} \bigcap_{j \in J} f(j) \) for all \( J \) and \( \{ S_j \} \). Completely distributive lattices have long been a topic of study; but fairly specialized and we only give the definition in preparation for the next definition. Let DD denote the complete distributivity law, restricted to those \( \{ S_j \} \) where \( S_j \) is directed for each \( j \in J \).

**Theorem 21.** For \( L \in \text{PoSet}_{D⊔} \) let \( J \mapsto \sqcup J \) denote the monotone map from the poset Idl(L) to \( L \). For \( L \in \text{CLat} \) the following are equivalent.

a. \( L \) is continuous.

b. \( J \mapsto \sqcup J \) has a left adjoint.

c. \( J \mapsto \sqcup J \) preserves meets.

d. \( L \) satisfies the infinite distributive law DD defined above.

The left adjoint, if it exists, maps \( x \) to \( x \ll \).

**Proof:** A left adjoint to \( J \mapsto \sqcup J \) would map \( x \) to the smallest ideal \( J \) such that \( x \subseteq \sqcup J; \) clearly this exists iff it equals \( \sqcup x \ll \). The equivalence of b and c is immediate from corollary 4. It is easy to see that DD is equivalent to the version where directed sets are replaced by ideals. This is turn is equivalent to c, since \( \bigcap_{j \in J} f(j) \in \bigcap_{j \in J} S_j \), and for \( x \in \bigcap_{j \in J} S_j \), take \( f \) to be constantly \( x \).

In particular a completely distributive lattice is continuous. We state the following without proof.

- The equivalence of a and b of Theorem 21, and the last claim, hold in \( L \in \text{PoSet}_{D⊔} \).

- The Cartesian product of continuous lattices is continuous. See [Gierz et al], I-2.7; alternatively, this can be seen from Theorem 21. The left side of DD holds in the product iff in each component \( S_j \) is directed and the left side holds, iff the right side holds in each component, iff the right side holds in the product.

- The Cartesian product is a product in either ContLat or ContLat^D.

- In CLat^D, the continuous lattices are closed under subobject.

- Theorem 5 holds in ContLat^D.

- ContLat is Cartesian closed.

- For \( P \in \text{PoSet} \) the map \( x \mapsto x \ll \) from \( P \) to the poset Idl(P) has a left adjoint iff \( P \in \text{PoSet}_{D⊔} \), in which case the left adjoint is \( J \mapsto \sqcup J \). ([Johnstone], Lemma VII.2.1).

**Lemma 22.** In a continuous poset,

a. if \( v \ll x \) then there is a \( w \) with \( v \ll w \ll x \); and

b. if \( x \ll y \) and \( y \leq \sqcup D \) where \( D \) is directed then there is a \( d \in D \) with \( x \ll d \).

**Proof:** Consider the set of chains \( u \ll w \ll x \); the set \( S \) of \( u \) in such chains is nonempty, and is also directed; if \( u_i \ll w_i \ll x \) for \( i = 1, 2 \) then there is a \( w \) with \( u_i \ll w \ll x \) and \( w_i \leq w \), whence a \( u \) with \( u \ll w \ll x \) and \( u_i \leq u \). The join of \( S \) is clearly \( x \), so if \( v \ll x \) then there is a \( u \in S \) and a \( w \) with \( v \leq u \ll w \ll x \), whence \( v \ll w \ll x \). This proves part a. For part b, by by part a \( x \ll w \ll y \) for some \( w \), so for some \( d \in D \) \( w \leq d \), and \( x \ll d \).
The next theorem gives various properties of the Scott topology on a continuous poset. In any poset, \( x^\leq \) is closed in the Scott topology, indeed is the closure of \( \{x\} \). Also \( x^\leq \) is compact, indeed in any topology where the open sets are \( \geq \)-closed. In a continuous poset let \( x^\gg \) denote \( \{y : x \ll y\} \); note that \( x^\gg \subseteq x^\geq \).

**Theorem 23.** Let \( P \) be a continuous poset equipped with the Scott topology.

a. The sets \( x^\gg \) form a base; indeed, if \( V \) is open and \( x \in V \) then there is a \( w \in V \) with \( x \in w^\gg \).

b. If \( V \) is open and \( x \in V \) then there is a compact set \( K \) and an open set \( U \) with \( x \in U \subseteq K \subseteq V \).

c. If \( P \) has 0 (in particular if \( P \) is a lattice) then \( P \) is compact.

d. The irreducible closed sets are the sets \( x^\leq \) for \( x \in P \); in particular \( P \) is a sober space.

**Proof:** Since \( x^\gg \) is \( \geq \)-closed to see that it is open it suffices to show that if \( D \) is directed and \( \sqcup D \subseteq x^\gg \), that is, \( x \ll \sqcup D \), then for some \( d \in D \) \( d \in x^\gg \); that is, \( x \ll d \); this is immediate by lemma 22.b. Suppose \( V \) is open and \( x \in V \); then \( x^\leq \) is directed and \( \sqcup x^\leq = x \in V \), so there is some \( w \in x^\leq \cap V \); then \( x \in w^\gg \) and \( w^\gg \subseteq U \). Part a is now proved. For part b, by part a there is a \( w \) with \( x \in w^\gg \) and \( w \in V \); let \( U = w^\gg \) and \( K = w^\geq \). Part c holds because \( P = 0^\geq \). For part d, since \( x \leq \{x\}^\geq \) it is irreducible. Suppose \( K \) is an irreducible closed set, and let \( K^\leq = \sqcup \{x^\leq : x \in K\} \). Suppose \( x, y \in K^\leq \); then \( x^\leq \nsubseteq K^\leq \) and \( y^\geq \nsubseteq K^\leq \), and because \( K \) is irreducible, \( x^\gg \cap y^\geq \nsubseteq K^\leq \). Choose \( z \in K \) with \( z \ll x \) and \( y \ll z \); since \( z^\leq \) is directed there is a \( z' \ll z \) with \( x \ll y \ll z' \). Thus, \( K^\leq \) is directed, and so \( u = \sqcup K^\leq \) exists. Since \( K^\leq \subseteq K \) and \( K \) is closed, \( u \in K \); also, if \( x \in K \) then \( x = \sup x^\leq \leq u \), whence \( K = u^\leq \).

The property of part b is called local compactness by many authors in lattice theory; in this text the term strongly locally compact will be used. The weaker property of local compactness as defined in chapter 17 has been of interest in topology; by lemma 17.12.c, in a Hausdorff space the two notions are equivalent. For the rest of this section, “s.l.c.” is used as an abbreviation for “strongly locally compact”.

**Lemma 24.** Suppose \( P \) is a continuous poset.

a. If \( x \ll y \) then there is an open filter \( U \) with \( U \subseteq x^\gg \) and \( y \in U \).

b. If \( z \not\ll y \) then there is an open filter \( U \) with \( z \not\in U \) and \( y \in U \).

**Proof:** Let \( y_0 = y \) and inductively choose \( y_{i+1} \) with \( x \ll y_{i+1} \ll y_i \); let \( U = \sqcup y_i^\gg \). Clearly \( U \) is a filter containing \( y \), and \( U \subseteq x^\gg \) since \( y_i^\gg \subseteq x^\gg \) for each \( i \). If \( u \in U \) then \( u \geq y_i \) for some \( i \), whence \( u \in y_i^\gg \); thus, \( U \) is open. This proves part a. For part b, since \( y = \sqcup y^\leq \), there is an \( x \ll y \) with \( x \not\ll y \); the claim follows by part a.

Let \( L \) be a complete lattice. Generalizing an earlier notation, for \( S \subseteq L \) let \( \beta_S = \{p \in \text{Pt}(L) : p(x) = 1 \} \) for all \( x \in S \). For \( T \subseteq \text{Pt}(L) \) let \( \sigma_T = \{p \in L : p(x) = 1 \} \) for all \( p \in T \). Then \( \beta \) and \( \sigma \) are the maps of a Galois adjunction from \( \text{Pow}(L) \) to \( \text{Pow}(\text{Pt}(L))^{op} \). In addition, \( \beta_S \) and \( \sigma_T \) are both \( \geq \)-closed.

**Lemma 25.** Let \( L \) be a complete lattice, and \( T \subseteq \text{Pt}(L) \). Then \( T \) is compact iff \( \sigma_T \) is Scott open.

**Proof:** First, note that in a lattice, a subset \( U \) is Scott open iff it is \( \geq \)-closed, and whenever \( \sqcup S \subseteq U \) then for some finite \( F \subseteq S, \sqcup F \subseteq U \). Second, recall that \( \sqcup_{x \in S} \beta_x = \beta_{\sqcup S} \); whence \( T \subseteq \sqcup_{x \in S} \beta_x \) iff \( T \subseteq \beta_{\sqcup S} \) iff \( \sqcup S \in \sigma_T \). Suppose \( T \) is compact, and \( \sqcup S \in \sigma_T \) for \( S \subseteq L \). Then \( T \subseteq \sqcup_{x \in S} \beta_x \); whence \( T \subseteq \sqcup_{x \in S} \beta_x \) for some finite \( F \subseteq S \), whence \( \sqcup F \in \sigma_T \). This shows that \( \sigma_T \) is Scott open. Now suppose \( \sigma_T \) is Scott open, and \( T \subseteq \sqcup_{x \in S} \beta_x \) for \( S \subseteq L \). Then \( \sqcup S \in \sigma_T \), whence \( \sqcup F \in \sigma_T \) for some finite \( F \subseteq S \), whence \( T \subseteq \sqcup_{x \in F} \beta_x \). This shows that \( T \) is compact.

**Lemma 26.** Suppose \( Y \) is a distributive complete lattice, and \( F \) is a Scott open filter in \( Y \). Then \( \sigma_{\beta_F} = F \), and \( \beta_F \) is a compact subset of \( \text{Pt}(Y) \).

**Proof:** By the adjunction \( F \subseteq \sigma_{\beta_F} \). By corollary 11.17, for \( x \not\in F \) there is a \( p \in \text{Pt}(Y) \) such that \( p(w) = 1 \) for \( w \in F \) and \( p(x) = 0 \). This proves the first statement. The second statement follows by lemma 25.

277
Suppose \( L \) is a continuous lattice, and is distributive. Then the map \( x \mapsto w \cup x \) preserves binary joins, and by theorem 21 it preserves directed joins, so it preserves all joins and \( L \) is a frame. Let \( \text{ContFrm} \) denote the full subcategory of \( \text{Frm} \), whose objects are those frames which are also continuous lattices.

**Theorem 27.** If \( Y \in \text{ContFrm} \) then \( \text{Pt}(Y) \) is s.l.c.

**Proof:** Suppose \( p \in \beta_y \) where \( y \in Y \). Since \( y = \sqcup y^\preccurlyeq \), if \( p(z) = 0 \) for all \( z \preccurlyeq y \) then \( p(y) = 0 \); so for some \( z \preccurlyeq y \) \( p(z) = 1 \). Choose an open filter \( U \) with \( U \subseteq z^\preccurlyeq \) and \( y \in U \). Then \( p \in \beta_z \subseteq \beta_U \subseteq \beta y \).

In the lattice \( \Omega(X) \) of open sets of a topological space \( X \), \( U \preccurlyeq V \) holds iff for any open cover of \( V \) there is a finite subset which covers \( U \). \( X \) is said to be core-compact if \( \Omega(X) \) is a continuous lattice, that is, given any open set \( V \) and point \( x \in V \), there is an open set \( U \) with \( x \in U \subseteq V \) and \( U \preccurlyeq V \).

**Theorem 28.** Let \( X \) be a topological space.

a. For open \( U, V \), if \( U \subseteq K \subseteq V \) where \( K \) is compact then \( U \preccurlyeq V \).

b. If \( X \) is s.l.c. the converse holds (i.e., such a \( K \) exists).

c. If \( X \) is s.l.c. then \( X \) is core-compact.

d. If \( X \) is sober then \( X \) is core-compact iff \( X \) is s.l.c.

e. If \( X \) is Hausdorff then \( X \) is core-compact iff \( X \) is locally compact.

**Proof:** For part a, any open cover of \( V \) is a cover of \( K \), so there is a finite subset which is a cover of \( K \), and this covers \( U \). For part b, for each \( x \in V \) there is an open \( U_x \) and a compact \( K_x \) with \( x \in U_x \subseteq K_x \subseteq V \). Then \( V = \cup U_x \), so there is a finite set \( F \) with \( U \subseteq \cup_{x \in F} U_x \). Let \( K = \cup_{x \in F} K_x \); then \( K \) is compact and \( U \subseteq K \subseteq V \). Part c follows by part a. For part d, if \( X \) is sober then it is isomorphic to \( \text{Pt}(\Omega(X)) \) by theorem 14. If \( X \) is also core-compact then \( \Omega(X) \in \text{ContFrm} \), so by theorem 27 \( \text{Pt}(\Omega(X)) \) is s.l.c. Part e follows because a Hausdorff space is sober, and a Hausdorff locally compact space is s.l.c.

By lemma 24.b and lemma 11.16, a continuous frame is spatial. So by the theorem the adjunction of theorem 14 restricts to an equivalence of categories between the full subcategory \( \text{ContFrm} \) of \( \text{Frm} \), and the full subcategory of \( \text{Top} \) of sober and s.l.c. spaces. A further adjunction of interest may be obtained by requiring the maps between continuous frames to preserve \( \preccurlyeq \) as well as joins and finite meets. The maps between spaces are required to be such that for “saturated” \( x \in (K \ f^{-1}[K]) \) is compact; see [Gierz et al].

A relation \( \prec \) on a complete lattice \( L \) is called an auxiliary order if

1. \( x \prec y \) implies \( x \leq y \),
2. \( x \prec y \) and \( x' \leq x \) implies \( x' \prec y \),
3. \( x \prec y \) and \( y \leq y' \) implies \( x \prec y' \),
4. \( x_1 \prec y \) and \( x_2 \prec y \) imply \( x_1 \sqcup x_2 \prec y \), and
5. \( 0 \prec y \).

For example \( \ll \) is an auxiliary order. An auxiliary order is said to be approximating if for all \( x \in L \), \( x = \sup \{ w : w \prec x \} \). For example a complete lattice is continuous iff \( \ll \) is approximating.

For a topology \( T \) on a complete lattice \( L \), define \( x \prec_T y \) iff \( y \in (x^\preccurlyeq)^\text{int} \), where the interior is taken in \( T \). That is, there must be a \( U \in T \) with \( y \in U \) and \( U \subseteq x^\preccurlyeq \). If open sets are \( \geq \)-closed then \( \prec_T \) is an auxiliary order; the verification of the required properties is straightforward and left to the reader.

The \( \geq \)-closed sets comprise a topology on \( L \), called the Alexandrov topology. The open sets of a topology \( T \) are \( \geq \)-closed iff \( T \) is contained in this. One readily sees that if \( T \) is the Alexandrov topology then \( \prec_T \) is \( \leq \).

**Lemma 29.** Let \( L \) be a complete lattice, and let \( \prec \) be \( \prec_T \) where \( T \) is the Scott topology.

a. If \( x \prec y \) then \( x \ll y \).

b. \( \prec \) and \( \ll \) coincide iff \( x^\preccurlyeq \) is Scott open for all \( x \), in particular if \( L \) is continuous.

c. \( L \) is continuous iff \( \prec \) is approximating.

278
Proof: Suppose \( y \in U \) where \( U \) is Scott open and \( U \subseteq x^2 \), and suppose \( y \leq \sqcup S \). Then \( \sqcup S \subseteq U \), so \( \sqcup F \subseteq U \) for some finite \( F \subseteq S \), and \( z \leq \sqcup F \). This proves part a. Suppose \( x \ll y \) implies \( x \prec y \), and suppose \( \sqcup S \in x^\geq, \) i.e., \( x \ll \sqcup S \). The \( x \prec \sqcup S \), so for some Scott open \( U, \sqcup S \in U \subseteq x^2, \) so \( \sqcup F \subseteq U \) for some finite \( F \subseteq S \), and \( \sqcup F \in x^\geq \). This shows that \( x^\geq \) is Scott open. Suppose \( x^\geq \) is Scott open for all \( x \). Suppose \( x \ll y \); then \( y \in x^\geq \subseteq x^2 \), so \( x \ll y \). This proves part b. For part c, if \( \prec \) is approximating then \( \ll \) is, by part a; and if \( L \) is continuous then \( \prec \) is approximating by part b.

6. Exponential correspondence. Recall the correspondence \( g \mapsto \bar{g} \) where \( g(x,y) = \bar{g}(x)(y) \) between \( \text{Hom}(X \times Y, Z) \) and \( \text{Hom}(X, \text{Hom}(Y, Z)) \), of chapter 13. These are the maps of an adjunction making \( \text{Hom}(Y, -) \) a right adjoint to \( - \times Y \). Specializing the definition in section 18.2, a topological space \( Y \) is said to be exponentiable if for any topological space \( Z \) there is a topology on \( \text{Hom}_{\text{Top}}(Y, Z) \), such that these are the maps of such an adjunction in \( \text{Top} \).

In fact, it is not necessary to specialize the definition. If \( Y \) is exponentiable according to the general definition, letting \( X \) be a one point space induces a bijection between \( Z^Y \) and \( \text{Hom}(Y, Z) \). Further the induced adjunction in \( \text{Set} \) is naturally equivalent to the one above.

A topology on \( \text{Hom}_{\text{Top}}(Y, Z) \) is said to be exponential if

1. continuity of \( g \) implies continuity of \( \bar{g} \); and
2. continuity of \( \bar{g} \) implies continuity of \( g \).

\( Y \) is exponentiable iff \( \text{Hom}(Y, Z) \) has an exponential topology for all \( Z \). The two requirements are often given names; we will let \( T_i \) denote the topologies on \( \text{Hom}(Y, Z) \) satisfying requirement \( i \) for \( i = 1, 2 \). Recall that a topology with fewer (more) sets is called weaker (stronger). Recall also from section 18.2 the evaluation map \( \varepsilon_{Y,Z} \), or simple \( \varepsilon \), where \( \varepsilon(f,y) = f(y) \); note that \( \varepsilon \) is the identity function.

Lemma 30. A topology \( T \) on \( \text{Hom}(Y, Z) \) is in \( T_2 \) iff it makes \( \varepsilon_{Y,Z} \) continuous.

Proof: For any topology \( T, \varepsilon \), being the identity, is continuous from \( \text{Hom}(Y, Z) \) to \( \text{Hom}(Y, Z) \), both equipped with \( T \). If \( T \in T_2 \) it follows that \( \varepsilon \) is continuous. Conversely suppose \( T \) is such that \( \varepsilon \) is continuous, and suppose \( \bar{g} \) is continuous. Since \( g = \varepsilon \circ (\bar{g} \times \iota_Y) \), \( g \) is continuous.

Lemma 31. \( T_1 \) is closed down, \( T_2 \) is closed up, and any topology in \( T_1 \) is weaker than any topology in \( T_2 \). In particular, there is at most one exponential topology.

Proof: If \( \bar{g} \) is continuous when \( \text{Hom}(Y, Z) \) is equipped with \( T \), and \( T' \subseteq T \), then \( \text{Hom}(Y, Z) \) is continuous when equipped with \( T' \); thus, \( T_1 \) is closed down. If \( \bar{g} \) is not continuous when \( \text{Hom}(Y, Z) \) is equipped with \( T \), and \( T \subseteq T' \), then \( \text{Hom}(Y, Z) \) is not continuous when equipped with \( T' \); thus, \( T_2 \) is closed up. Suppose \( T_i \in T_i \) for \( i = 1, 2 \). Let \( \text{Hom}_{i}(Y, Z) \) denote \( \text{Hom}(Y, Z) \) equipped with \( T_i \). By lemma 30 \( \varepsilon \) is continuous on \( \text{Hom}_{2}(Y, Z) \). By the hypothesis on \( T_i \), \( \varepsilon \), that is, the identity, is a continuous map from \( \text{Hom}_{2}(Y, Z) \) to \( \text{Hom}_{1}(Y, Z) \). This shows that \( T_1 \) is weaker than \( T_2 \).

For the rest of this section let \( C_2 \) denote the two element chain \( \{0, 1\} \), with the topology \( \{\emptyset, \{1\}, \{0, 1\}\} \) (called the Sierpinski space). The map \( f \mapsto f^{-1}[1] \) is a bijection from \( \text{Hom}(Y, C_2) \) to \( \Omega(Y) \), with inverse \( U \mapsto \chi_{U} \) (the characteristic function of \( U \)). This induces a bijection between the topologies on \( \text{Hom}(Y, C_2) \) and those on \( \Omega(Y) \). A topology on \( \Omega(Y) \) is said to be exponential if the corresponding topology on \( \text{Hom}(Y, C_2) \) is; the requirements are as follows, where lemma 30 is used for the second.

1. If \( W \) is open in \( X \times Y \) then the function \( x \mapsto W_x \) is a continuous function from \( X \) to \( \Omega(Y) \), where \( W_x = \pi_2^{-1}(x) \cap W \) and \( \pi_2 \) is projection; note that \( W_x \) is open.
2. Letting \( E_Y \), or just \( E \), denote \( \{(U, y) : U \in \Omega(Y), y \in Y\} \), \( E \) is open in the product of the topologies \( T \) and \( \Omega(Y) \).

We use \( T_i \) also to denote the topologies on \( \Omega(Y) \) satisfying requirement \( i \).
If $T$ is a topology on $\Omega(Y)$, $Z$ is a topological space, $U \subseteq T$, and $V \subseteq \Omega(Z)$ let $\sigma_{UV} = \{f \in \text{Hom}(Y, Z) : f^{-1}[V] \in U\}$. Let $T^*$ be the topology on $\text{Hom}(Y, Z)$ which has $\{\sigma_{UV}\}$ as a subbase. It is readily verified that $\sigma_{U_1 \cap \cdots \cap U_k, V} = \sigma_{U_1, V} \cap \cdots \cap \sigma_{U_k, V}$, and $\sigma_{U_{\alpha}, V} = \cup_{\alpha} \sigma_{U_{\alpha}, V}$, and so if $S$ is a subbase for $T \{\sigma_{UV} : U \subseteq S\}$ is a subbase for $T^*$.

For example, let $x^u = \{U \in \Omega(Y) : x \in U\}$, and let $T$ be the topology on $\Omega(Y)$ which has $\{x^u\}$ as a subbase. Then $f \in \sigma_{x^u}$ iff $f^{-1}[V] \subseteq x^u$ iff $x \in f^{-1}[V]$ iff $f(x) \in V$. This topology on $\text{Hom}(Y, Z)$ is called the pointwise topology; note that it is just the product topology.

More generally for $D \subseteq Y$ let $D^u = \{U \in \Omega(Y) : D \subseteq U\}$. The topology $T^*$ where $S$ is the set of $\sigma_{K^u, V}$ where $K$ is compact is called the compact-open topology. This has as a subbase those sets $\{f : f[K] \subseteq V\}$ where $K$ is a compact subset of $Y$ and $V$ is an open subset of $Z$.

**Lemma 32.** $Y$ is exponentiable iff $\Omega(Y)$ has an exponential topology. If $T$ is the exponential topology on $\Omega(Y)$ then $T^*$ is the exponential topology on $\text{Hom}(Y, Z)$.

**Proof:** If $Y$ is exponentiable then $\text{Hom}(Y, C_2)$ has an exponential topology, so $\Omega(Y)$ does. Conversely suppose $T$ is an exponential topology on $\Omega(Y)$. For $V \subseteq \Omega(Y)$, $y \in (\tilde{g}(x))^{-1}[V]$ iff $g(x, y) \in V$ iff $(x, y) \in g^{-1}[V]$ iff $y \in \pi_1^{-1}(x) \cap g^{-1}[V]$. That is, $(\tilde{g}(x))^{-1}[V] = \pi_1^{-1}(x) \cap g^{-1}[V]$. So for $U \subseteq T$, $x \in g^{-1}[\sigma_{UV}]$ iff $\tilde{g}(x) \in \sigma_{UV}$ iff $(\tilde{g}(x))^{-1}[U] \subseteq \pi_1^{-1}(x) \cap g^{-1}[V]$. Since $T \subseteq T_1$, $g^{-1}[\sigma_{UV}]$ is open. This shows $T^* \subseteq T_1$.

Given $f \in \text{Hom}(Y, Z)$ and $y \in Y$ let $W$ be an open neighborhood of $f(x) = \varepsilon(f, y)$. Then $(f^{-1}[W], y)$ is in $E = \{(U, y) : U \subseteq \Omega(Y), y \in Y\}$, which is open since $T \subseteq T_2$, so take $U \subseteq T$, $V \subseteq \Omega(Y)$ with $f^{-1}[U] \in U$, $y \in V$, and $U \times V \subseteq E$. Now, $\sigma_{UV} \times V$ is an open neighborhood of $(f, y)$, and if $(\hat{f}, \hat{y})$ is in this neighborhood then it is in $E$, whence $\hat{y} \in \hat{f}^{-1}[W]$, so $\varepsilon(\hat{f}, \hat{y}) \in W$. Thus, $\varepsilon$ is continuous, and thus $T^* \subseteq T_2$.

In fact we have shown that for $i = 1, 2$, if $T \subseteq T_i$ then $T^* \subseteq T_i$ for all $\text{Hom}(Y, Z)$.

By the definitions, for a topology $T$ on $\Omega(Y)$, $U \prec T V$ iff there is an $O \subseteq T$ such that $V \subseteq W \subseteq \Omega(Y)$. In the following we will use the term “approximating” for $\prec_T$ even if $T$ is not weaker than the Alexandrov topology. We also use $U^n$ to denote $U \subseteq \Omega(Y)$.

**Lemma 33.** Let $S$ denote the Scott topology on $\Omega(Y)$, and $T$ any other topology.

a. $S \subseteq T_1$.

b. $T$ is in $T_2$ iff $\prec_T$ is approximating.

c. $S = \cap T_2$.

d. $\Omega(Y)$ has an exponential topology iff $S$ is exponential.

**Proof:** Suppose $W \subseteq \times Y$ is open, $x \in W$, and $O$ is a Scott open neighborhood of $W_x$. To prove part a, we will find an open neighborhood $U$ of $x$ such that $W_u \subseteq O$ for all $u \subseteq U$. For each $y$ with $(x, y) \in W_x$ choose open neighborhoods $U_y$ of $x$ and $V_y$ of $y$ such that $U_y \times V_y \subseteq W$. The union of the $V_y$ is in $O$, so because $O$ is Scott open there is some finite set of $V_y$ whose union $V$ is in $O$; let $U$ be the intersection of the $U_y$ of this finite set. Suppose $u \subseteq U$; if $v \in V$ then $v \in V_y$ for some $y$, whence $(u, v) \in U_y \times V_y$, whence $(u, v) \in W$. This shows that $V \subseteq W_u$, which shows that $W_u \subseteq O$ and completes the proof of part a. Suppose $T \subseteq T_2$, $V \subseteq T$, and $y \in V$. Then $(V, y) \in E$, and since $E$ is open there are $O, U$ with $O \subseteq T$, $U \subseteq \Omega(Y)$, $V \subseteq O$, $x \subseteq U$, and $O \times U \subseteq E$. Since $\hat{y} \in \hat{F}$ for $\hat{v} \subseteq O$ and $\hat{y} \in U$, $O \subseteq U^n$. Thus, $x \subseteq U$ and $U \prec_T V$, which shows that $\prec_T$ is approximating. Suppose $\prec_T$ is approximating and $y \in V$. Then there is a $U \in \Omega(Y)$ with $U \prec_T V$; choose $O \subseteq T$ with $V \subseteq O$ and $O \subseteq U^n$. Then $(V, x) \in O \times U$ and $O \times U \subseteq E$, so $E$ is open and $T \subseteq T_2$. This proves part b. $S \subseteq \cap T_2$ by lemma 31 and part a. For $R \subseteq \Omega(Y)$ let $T_R$ be those $u$-closed $O \subseteq \Omega(Y)$ such that if $u \in R$ then $\cup F^{\infty} \subseteq O$ for some finite $F \subseteq R$. Clearly $S = \cap \cap T_R$, so to prove part c it suffices to show that $T_R \subseteq T_2$, for which by part b it suffices to prove that $\prec_{T_R}$ is approximating. If $u \in \cap \cup O_\alpha$ where $O_\alpha \in T_R$ then $\cup R \subseteq O_\alpha$ for some $\alpha$, whence $\cup F \in \cup O_\alpha$, so $\cup \cup O_\alpha \in T_R$. If $R \in O_i \cap O_2$ then for some finite $F_i \subseteq R \subseteq F_i \subseteq O_i$ for $i = 1, 2$, whence $\cup (F_1 \cup F_2) \subseteq O_1 \cap O_2$, whence
O_1 \cap O_2 \in T_R. Thus, T_R is a topology. Suppose V \in \Omega(Y) and y \in V. If y \not\in \cup R then V \prec_{T_R} V, because \( V^w \in T_R \) since V \subseteq W implies \cup R \not\subseteq W. If y \in U for some U \in R then U \cap V \prec_{T_R} V since (U \cap V)^w \in T_R.

This shows that T_R is approximating, and hence T_R \in T_2 by part b. For part d, one direction is trivial, and for the other, suppose T is an exponential topology on \( \Omega(Y) \). Then \( T \in T_1 \), so \( T \subseteq \cap T_2 \), so by part c \( T \subseteq S \). Since \( T \in T_2 \), \( S \in T_2 \). By part a, \( S \in T_1 \).

**Theorem 34.** \( \Omega(Y) \) has an exponential topology iff \( \Omega(Y) \) is a continuous lattice, in which case the exponential topology is the Scott topology. In particular, a topological space is exponentiable iff it is corecompact.

**Proof:** The first claim follows by lemma 33 and lemma 29, and the second is then follows by lemma 32.

**Corollary 35.** A sober space is exponentiable iff it is strongly locally compact. A Hausdorff space is exponentiable iff it is locally compact. For a sober exponentiable space Y the exponential topology on \( \text{Hom}(Y, Z) \) is the compact-open topology.

**Proof:** The first two claims follow by lemma 28. For the third, as observed above the compact-open topology is \( T^* \) where \( T \) has the set of \( K^u \) for \( K \) compact as a subbase; so it suffices to show that these are a base for the Scott topology. Given a set \( O \) open in the Scott topology on \( \Omega(Y) \) and \( V \in O \) choose \( U \in O \) with \( U \ll V \). By theorem 28 there is a compact \( K \subseteq Y \) with \( U \subseteq K \subseteq V \), and \( V \in K^u \) and \( K^u \subseteq U^u \subseteq O \).

We have followed [EscHeck] in this section. The main result was first proved in 1970; the result for Hausdorff spaces was proved in 1945.

**Exercises.**

1. For \( L \in \text{Lat}_0 \) with a pseudo-complement, show that \( (x \cap y)^{pp} = x^{pp} \cap y^{pp} \). Hint: since \( x \cap y \leq x, y \), \( (x \cap y)^{pp} \leq x^{pp} \cap y^{pp} \). Also \( (x \cap y)^p \geq x^p \cup y^p \).

2. Suppose \( L \) is a distributive bounded lattice where every element has a pseudo-complement. Prove that the conditions stated in section 2 characterizing a Heyting algebra with 0 where \( L^p \) is a sublattice, are equivalent. Hint: Condition b follows from a because the greatest element of \( L^p \) is 1. Condition a follows from c because \( x^p \cup y^p \) is in \( L^p \). To see that c follows from b, in any such \( L \), \( (x \cup y)^p \cap (x \cap y) = 0 \); and if \( w \cup x \cap y = 0 \) then \( y \cup w \leq x^p \), so \( y \cup w \cap x^{pp} = 0 \), so \( w \cap x^{pp} \leq y^p \). Now meet \( w \) with 1 and use \( x^p \cup x^{pp} = 1 \), to obtain \( w \leq x^p \cup y^p \).

3. In \( \text{CLat}_{\uparrow,0} \), show the following.
   a. The free object generated by a set \( X \) is \( \text{Pow}(X) \), with the subset order. Hint: The universal arrow from \( X \) to the power set is \( x \mapsto \{ x \} \). A function \( k \) from \( X \) to the set underlying an object \( S \) is extended to \( \text{Pow}(X) \) by mapping a set \( s \) to \( \sqcup_{x \in x} k(x) \). The resulting map preserves joins.
   b. The coproduct of a set of objects exists. Hint: A general approach is to take the quotient of the free algebra on all the generators (the disjoint union), by the equivalence resulting from the relations which hold in each individual object (this exists by theorem 5.c). In this case, the coproduct consists of the partial functions selecting an element from each object of a subset of the objects. (In \( \text{CLat}_{\uparrow,0} \), the pairwise product is a biproduct).
   c. Coequalizers exist.

4. Show the following.
   a. If \( Y \in \text{SLat}_{\uparrow,0} \), the subsets of \( Y \) which are \( \leq \)-closed is the universal frame for \( Y \).
   b. In \( \text{Frm} \), the free object generated by a set \( X \) is those subsets of \( Y \) which are closed under subset, where \( Y \) is the finite subsets of \( X \). Hint: use part a and exercise 3.
   c. Colimits exist in \( \text{Frm} \).

5. In \( \text{PoSet}_{D,\sqcup} \), show the following.
a. The free object generated by a set $X$ is the discrete order on $X$.
b. Colimits exist.

6. Show that the Scott open sets in a poset $P$ in $\text{PoSet}_{\text{DLat}}$ comprise a topology on $P$.

7. Suppose $U$ is an open set in the upper interval topology on a poset $P$.
a. Show that $U$ is $\geq$-closed.
b. Show that if $\sqcup S \in P$ for some $S \subseteq P$, then $S \cap U \neq \emptyset$.

8. Show that in a coherent Hausdorff space, if the clopen sets form a Boolean algebra then they are closed under union. Hint: suppose $u$ is regular open. Let $s = \sqcup_B (B \cap u^c)$; then $s \geq \sqcup (B \cap u^c) = u$. If $w$ is clopen and $w \subseteq u^c$ then $s \cap w = \emptyset$; thus $s \cap u^c = \emptyset$. This shows that $s = u$, and so $u$ is clopen.

9. Show that $\text{Hom}(-, C_2)$, considered as a functor from $\text{DLat}^{\text{op}}$ to $\text{PoSet}$, is right adjoint to the functor from $\text{PoSet}$ to $\text{DLat}^{\text{op}}$, which takes a poset to its lattice of $\leq$-closed sets, and an arrow $f : P \rightarrow Q$ in $\text{PoSet}$ to the map $S \mapsto f^{-1}[S]$.

10. Show that the closed unit interval with the usual order is a continuous lattice which is not algebraic.

11. Show that axioms 3 and 4 for an orthomodular lattice imply DeMorgan’s laws. Hint: From $x, y \leq x \sqcup y$ we have $x^c \cap y^c \geq (x \sqcup y)^c$. Substitute $x^c, y^c$ for $x, y$ in the dual inequality and complement.

12. Show that axioms 1 to 4 for an orthomodular lattice are equivalent. Hint: The equivalence of 1 and 2 is as earlier arguments. Suppose 3 holds, and the hypothesis of 1; then the hypothesis of 3 holds, so the conclusion of 3 holds, so the conclusion of 1 holds. Suppose 1 holds, and the hypothesis of 3. Apply DeMorgan’s law to the hypothesis of 3, and conclude that $x^c \cap y = (x \sqcup y)^c \cap y$. Then $(x \sqcup y^c) \sqcup (x^c \cap y) = (x \sqcup y) \sqcup ((x \sqcup y)^c \cap y)$, which by 1 equals $y$. That 4 follows from 1 is immediate. Suppose 4 holds, and $x \leq y$; let $w = x \sqcup (y \cap x^c)$. Then $w \leq y$, and using DeMorgan’s laws $y \cap w^c = 0$, whence by 1 $w = y$. 282
22. Convexity.

1. Basic facts. Suppose $X$ is a vector space over $\mathbb{R}$. A subset $S \subseteq X$ is said to be convex if $rx + (1 - r)y \in S$ whenever $x, y \in S$ and $0 \leq r \leq 1$. Some authors require a convex subset to be nonempty; but we will not do so, and indicate explicitly when this is required.

The study of convex sets goes back to the nineteenth century; an excellent recent text on the subject is [Barvinok]. Convex sets in Euclidean space $\mathbb{R}^n$ of some dimension $n$ are of particular interest, but the notion may be defined in an arbitrary real vector space.

The set $\{rx + (1 - r)y : 0 \leq r \leq 1\}$ is called the “line segment” between $x$ and $y$, so that a set is convex if whenever it contains two points it contains the line segment between them. The term “convex body” is used variously in the literature to denote a convex set $S$ in $\mathbb{R}^n$ with further restrictions imposed. Typical such are that $S$ be closed, bounded, and have non-empty interior.

One readily verifies that in any real linear space $X$, the intersection of any family of convex sets is again convex; and the union of any directed family of convex sets is again convex. Thus, the convex subsets of $X$ form an algebraic closure system. In particular, given any set $S$ there is a least convex set containing it. This is called the convex hull of $S$; we use Chull($S$) to denote it.

As usual, Chull($S$) for a subset $S$ of a real linear space $X$ may be given an alternative description. A convex linear combination of elements $x_i \in X$ is defined to be a linear combination $\sum r_i x_i$, where $r_i \geq 0$ for all $i$, and $\sum r_i = 1$. One readily verifies by induction that any convex linear combination of elements of a convex set $K$ is again in $K$. Chull($S$) equals the set of convex linear combinations of $S$; indeed, the latter set is convex, so contains Chull($S$); and must be contained in any convex set containing $S$.

Other closure properties enjoyed by the convex subsets of a real linear space include the following.

- The sum $S_1 + S_2$ (also called the Minkowski sum) of convex sets $S_1$ and $S_2$ is convex.
- A translate $r + S$ or scalar multiple $rS$ of a convex set $S$ is convex.
- A subspace is convex.
- The image $f[S]$ or inverse image $f^{-1}[S]$ of a convex set $S$ under a linear map $f$ is convex.

If the vector space has a topology then relations between the topology and convexity are of interest. Rather than giving such just for $\mathbb{R}^n$, we will define the categories of real topological vector spaces, real inner product spaces, and real Hilbert spaces. These are discussed at greater length in chapter 24; here proofs of facts regarding convexity will be given directly from the definitions.

A real topological vector space is a real vector space $V$ equipped with a topology, under which addition, and scalar multiplication as a function on the product topological space $\mathbb{R} \times V$, are continuous. A morphism in the category of such is a linear map which is also continuous.

The notion of a real normed linear space $V$ has already been mentioned, in section 17.9. It is a real vector space equipped with a function $|x| : V \mapsto \mathbb{R}$, which satisfies the axioms for a norm. The latter are given in section 10.7, and are $|x| \geq 0$, $|x| = 0$ iff $x = 0$, $|ax| = |a||x|$, and $|x + y| \leq |x| + |y|$. One readily verifies that $|x - y|$ is a metric on $V$; thus a real normed linear space is a real topological vector space when equipped with the metric (or norm) topology. This category is a full subcategory of the real topological vector spaces, that is, morphisms are continuous linear maps.

A real Hilbert space is a real vector space $V$, equipped with an inner product, which as a metric space is complete. Here an inner product is a positive definite symmetric bilinear form $x \cdot y : V \times V \mapsto \mathbb{R}$. It is shown in section 10.7 that the function $|x| = \sqrt{x \cdot x}$ is a norm. This category is a full subcategory of the real normed linear spaces. $\mathbb{R}^n$ with the standard Euclidean inner product (whence the Euclidean norm) is a real Hilbert space, in fact as will be seen in chapter 24 the only $n$-dimensional real topological vector space.

In any real topological vector space, the closure of a convex set is again convex. Indeed, if $U$ is an open neighborhood of $rx + (1 - r)y$ then there are open neighborhoods $V$ and $W$ of $x$ and $y$ respectively, such
that \( rV + (1 - r)W \subseteq U \); so if \( U \) is disjoint from \( S \) one of \( V \) or \( W \) must be.

In any real topological vector space, the interior of a convex set \( S \) is again convex. Indeed, if \( x \in U \subseteq S \), \( y \in S \), and \( z = rx + (1 - r)y \) where \( 0 < r \leq 1 \), then \( w \in rU + (1 - r)y \subseteq S \).

In a real normed linear space, an open or closed ball is convex; this follows because \(|rx + (1 - r)y| \leq r|x| + (1 - r)|y| \). A subcategory of interest of the real topological spaces is the locally convex real topological vector spaces, which are those having a base of convex open sets; the preceding shows that these include the normed linear spaces.

In a real Hilbert space the parallelogram law (defined in section 10.7) holds, in the form \(|x + y|^2 + |x - y|^2 = 2(|x|^2 + |y|^2)\).

**Lemma 1.** Suppose \( X \) is a real Hilbert space, \( S \subseteq X \) is a nonempty closed convex subset, and \( x \in X - S \). Then there is a unique point \( y \in S \) such that \(|y - x| = \inf\{\|w - x\| : w \in S\}\).

**Proof:** By translating \( S \) we may suppose that \( x = 0 \). Let \( d = \inf\{|w| : w \in S\} \), and let \( w_n \in S \) be such that \(|w_n|\) converges to \( d \). Since \( S \) is convex, \(|(w_n + w_n)/2| \in S \), so \(|w_n + w_n| \geq d \). Since \(|w_n + w_n|^2 + |w_n - w_n|^2 = 2|w_n|^2 + 2|w_n|^2, |w_n - w_n|\) must approach 0, and \( w_n \) is a Cauchy sequence. Let \( w \) be the limit of \( w_n \); then \( w \in S \) and \(|w| = d \). If \( w' \in S \) and \(|w'| = d \), then \(|w - w'| = 0 \), else using the parallelogram law \(|(w + w')/2| < d \), a contradiction; thus, \( w \) is unique.

Finding the point which minimizes the distance is an optimization problem with a corresponding “variational inequality”; the following lemma gives the inequality.

**Lemma 2.** Suppose \( X \) is a real Hilbert space, \( S \subseteq X \) is a nonempty closed convex subset and \( x \in X - S \). Then \( z \in S \) is the unique point which minimizes \(|y - x|\) iff for all \( y \in S \), \((x - z) \cdot (y - z) \leq 0 \).

**Proof:** By the convexity of \( S \), \( z \) is the closest point iff for each \( y \in S \), for each \( t \in [0, 1] \) \(|x - (z + t(y - z))|^2 \geq |x - z|^2 \). By algebra the requirement for \( y \) holds iff for each \( t \in [0, 1] \) \( t(-2(x - z) \cdot (y - z) + t|y - z|^2) \geq 0 \), iff for each \( t \in [0, 1] \) \(-2(x - z) \cdot (y - z) + t|y - z|^2) \geq 0 \); the lemma follows.

For convenience we repeat some facts from basic linear algebra, and state a few further facts. In a vector space \( X \) over a field \( F \), the span of a subset \( S \subseteq X \) is the set of all linear combinations of its elements; \( S \) is said to generate its span. A subspace is called finite dimensional if it has a finite generating set, and the least size of such, which equals the size of a linearly independent generating set, is the dimension of the subspace. In \( F^n \) a subspace is a set of the form \( \{x : Mx = 0\} \) where \( M \) is a matrix; the dimension of the subspace is \( n - n' \) where \( n' \) is the rank of the matrix.

In a vector space, a hyperplane is defined to be a maximal proper subspace. If \( f \) is a linear functional and \( f(v) \neq 0 \) then for any \( x, x = w + (f(x)/f(v))v \) where \( f(w) = 0 \), which shows that \( \ker(f) \) is a hyperplane, and \( X = \ker(f) \oplus Fv \). On the other hand if \( H \) is a hyperplane and \( v \notin H \) then \( X = H \oplus Fv \), and the map \( w + rv \mapsto r \) is a linear functional.

In a real vector space, a halfspace is defined to be a set of the form \( \{x : f(x) \leq 0\} \) where \( f \) is a linear functional. A hyperplane corresponds to two halfspaces, one on “either side”, namely \( \{x : f(x) \leq 0\} \) and \( \{x : f(x) \geq 0\} \). These are also called closed halfspaces, although care must be taken, because if the space is a real normed linear space and the linear functional is not continuous then the hyperplane need not be a closed subset topologically.

In chapter 24 it will be shown that in a real Hilbert space, any linear functional is of the form \( v \cdot x \) for some vector \( v \) (in particular it is continuous). Thus, in a real Hilbert space a hyperplane is a set \( \{x : v \cdot x = 0\} \) for some vector \( v \); \( v \) is called a normal to the hyperplane. Its two halfspaces are \( \{x : v \cdot x \leq 0\} \) and \( \{x : v \cdot x \geq 0\} \).

An affine subspace (hyperplane, halfspace) is defined to be a translate of a subspace (hyperplane, halfspace). One readily verifies that a \( d \)-dimensional affine subspace is the “affine span” of \( d + 1 \) linearly
independent elements, i.e., the linear combinations $\sum r_i x_i$ where $\sum r_i = 1$. An affine subspace (hyperplane, halfspace) is a convex set. Affine subspaces are also called linear varieties or flats. These enjoy many closure properties; for example they are closed under intersection, sum, translation, and scalar multiplication.

An affine hyperplane is a set of the form $\{ x : f(x) = a \}$ where $f$ is a linear functional not identically 0 and $a \in F$. Its halfspaces are $\{ x : f(x) \leq a \}$ and $\{ x : f(x) \leq a \}$. In a real Hilbert space an affine hyperplane is a set of the form $\{ x : v \cdot x = a \}$ for some vector $v$; and its halfspaces are $\{ x : v \cdot x \leq a \}$ and $\{ x : v \cdot x \geq a \}$.

In a real vector space, an affine hyperplane $H$ is said to separate a set $T$ from a set $S$ if for one of the halfspaces $K$ of $H$, $S \subseteq K$ and $T \cap K = \emptyset$.

In a real topological vector space, a supporting hyperplane $H$ of a closed convex set $S$ is defined to be an affine hyperplane where $S$ lies in one of the halfspaces of $H$, and $S \cap H \neq \emptyset$. A proper face of $S$ is defined to be $S \cap H$ for a supporting hyperplane $H$. A face of $S$ is either a proper face, or all of $S$ (we have defined faces for any convex set, but they are of greater interest for polyhedra, as will be seen).

In a real Hilbert space $X$, using lemma 2 one verifies the following.
- If $S \subseteq X$ is a nonempty closed convex subset and $x \in X - S$, there is a supporting hyperplane separating $x$ from $S$.
- If $S$ is closed and convex, $S$ is either $X$, or the intersection of the affine halfspaces containing $S$.

**Theorem 3.** Suppose $S \subseteq \mathbb{R}^n$ is a nonempty closed convex subset, and $x$ is a boundary point of $S$. Then there is a supporting hyperplane of $S$ containing $x$.

**Proof:** There is a sequence $w_j$ in $\mathbb{R}^n - S$ converging to $x$. There is an affine hyperplane $H_j$, with unit normal $u_j$, with $S$ in, say, the nonnegative halfspace and $w_j$ not in it. The sequence $u_j$ has limit point, say $u$. One verifies that the halfspace at $x$ with normal $u$ has $S$ on the nonnegative side.

If $S$ is a convex subset of a real vector space $X$, an extreme point of $S$ is a point $x \in S$ such that if $x = \lambda u + (1 - \lambda)v$ for $0 < \lambda < 1$ and $u, v \in S$ then $x = u = v$. That is, $x$ is not in the interior of any line segment contained in $S$. Note that $\lambda$ can be replaced by $1/2$.

Following are some basic facts concerning extreme points; proofs are given in the exercises.
1. In a real linear space, if $S$ is a convex set and $F$ is a face of $S$ then the extreme points of $S$ in $F$ are the extreme points of $F$.
2. In a real linear space, if $S$ is a convex set and $\{ x \}$ is a face then $x$ is an extreme point.
3. In $\mathbb{R}^n$, a compact convex set is the convex hull of its extreme points.

We will see in theorem 6 below that the converse of fact 2 holds for polyhedra in $\mathbb{R}^n$.

A cone in a real vector space $X$ is a nonempty subset $S$ such that if $x \in S$ then $rx \in S$ for any $r \geq 0$. One readily verifies the following.
- Any cone contains 0.
- The convex cones in $X$ form an algebraic closure system.
- The join of two cones is their sum.
- A convex cone may be characterized as a nonempty subset of $X$ which is closed under nonnegative linear combinations.
- Letting $Ccone(S)$ denote the smallest convex cone containing a subset $S$ of $X$, $Ccone(S)$ is the set of nonnegative linear combinations of members of $S$.
- A subspace is a convex cone.
- A halfspace is a convex cone.
- The image $f[S]$ or inverse image $f^{-1}[S]$ of a convex cone $S$ under a linear map $f$ is a convex cone.

For readers familiar with projective space, convex cones are related to convex sets much as projective sets are to affine sets. Given any convex set $S$ in $\mathbb{R}^n$, a convex cone in $\mathbb{R}^{n+1}$ may be obtained by adding an extra component with value 1, and closing under nonnegative scalar multiplication.
By an argument similar to one given above, in a real topological vector space the closure of a cone is closed under multiplication by a nonnegative scalar and is again a cone. The closure of a convex cone is thus a convex cone. This is generally not true for the interior, since a convex cone must contain the origin.

Suppose \( S \) is a nonempty cone in a real vector space, lying in the affine halfspace \( \{x : f(x) \leq r \} \). Since \( 0 \in S, r \geq 0 \). If there is some \( x \in S \) with \( f(x) = r \), then \( r \) must equal 0, else for \( s > 1 \) \( f(sx) = sf(x) = sr > r \), a contradiction. It follows that in a real topological vector space, a supporting hyperplane of a closed convex cone must contain the origin.

A ray of a convex cone \( S \) is a set \( \{rx : r \geq 0\} \) for some nonzero \( x \in S \). An extreme ray is defined to be a ray \( L \), such if \( I \subseteq S \) is a line segment containing a point of \( L \) in its interior, then \( I \subseteq L \).

In a real Hilbert space \( X \), for a convex cone \( S \) in \( X \) let \( S^{\text{pol}} = \{x : x \cdot y \leq 0 \text{ for all } y \in S\} \). Let \( C \) be the convex cones, ordered by inclusion. \( S \mapsto S^{\text{pol}} \) is readily verified to be an order preserving map from \( C \) to \( C^{\text{op}} \). Further, \( S \subseteq T^{\text{pol}} \iff x \cdot y \leq 0 \) for all \( x \in S \) and \( y \in T \) iff \( T \subseteq S^{\text{pol}} \). Thus, the pair of maps \( S \mapsto S^{\text{pol}} \) from \( C \) to \( C^{\text{op}} \), and from \( C^{\text{op}} \) to \( C \), form a Galois adjunction. The usual facts follow, in particular - \( S \subseteq (S^{\text{pol}})^{\text{pol}} \), \( ((S^{\text{pol}})^{\text{pol}})^{\text{pol}} = S \), and \( (\bigcap_i S_i)^{\text{pol}} = \bigcap_i S_i^{\text{pol}} \).

The cone \( S^{\text{pol}} \) is called the polar of \( S \).

**Theorem 4.** If \( S \) is a closed convex cone in a real Hilbert space then \( (S^{\text{pol}})^{\text{pol}} = S \).

**Proof:** Suppose \( x_1 \notin S \), and let \( x_2 \) be the point \( y \) of lemma 1. By theorem 17.25, \( |x_1 - x_2| > 0 \). Let \( x_0 = (x_1 + x_2)/2 \), and let \( v = x_1 - x_0 \). Then \( v \cdot (x-x_0) < 0 \) for \( x \in S \), and it follows that \( v \cdot x \leq 0 \) for \( x \in S \), since otherwise \( v \cdot (rx-x_0) > 0 \) for \( r \) sufficiently large. Thus, \( v \in S^{\text{pol}} \). Since \( 0 \in S \), \( v \cdot x_0 > 0 \); since also \( v \cdot (x_1 - x_0) > 0 \), \( v \cdot x_1 > 0 \). Thus, \( x_1 \notin (S^{\text{pol}})^{\text{pol}} \). We have shown that \( (S^{\text{pol}})^{\text{pol}} \subseteq S \), proving the theorem.

2. Polyhedra and linear inequalities. From hereon we consider only Euclidean space, although some topics have infinite dimensional generalizations. Let \( \leq \) denote the product order on \( \mathbb{R}^n \), that is, \( x \leq y \) if \( x_i \leq y_i \) for all \( i \). The term “polyhedron” has various uses in mathematics. In convexity theory a polyhedron is defined to be the intersection of finitely many affine halfspaces, or equivalently as the points satisfying a system of linear inequalities \( Vx \leq a \), where each row of the matrix \( V \) is nonzero. Clearly a polyhedron is a closed convex set.

The theory of polyhedra has many relations to the theory of systems of linear inequalities. For example, corollary 8 below is a fundamental fact about polyhedra; we follow [Schröder] and prove it as a corollary of theorem 7, the “fundamental theorem of linear inequalities”, which may be proven by methods from the theory of optimization of linear systems, or “linear programming”. Before doing so, we prove some facts which may be proven without the fundamental theorem.

For the remainder of this section \( A \) will denote an \( n \times m \) matrix, with columns \( a_1, \ldots, a_m \); and \( b \) a column vector in \( \mathbb{R}^n \). For \( I \subseteq \{1, \ldots, m\} \) we suppose its elements are \( i_1 < \cdots < i_{|I|} \); and let \( A_I \) denote the matrix whose \( j \)th column is \( a_{i_j} \).

Let \( C \) be the statement that \( b \in \text{Cone}(a_1, \ldots, a_m) \); this is \( \exists x (Ax = b \wedge x \geq 0) \). Let \( S \) be the statement that there is a hyperplane separating \( b \) from \( \text{Cone}(a_1, \ldots, a_m) \); this is \( \exists y (y^T A \leq 0 \wedge y^T b > 0) \). It follows by arguments as in section 1 that \( \neg C \Rightarrow S \). The contrapositive \( \neg S \Rightarrow C \), if \( b \) is such that \( y^T b \leq 0 \) whenever \( y^T A \leq 0 \) then \( b \in \text{Cone}(a_1, \ldots, a_m) \), is known as Farkas’ lemma. One readily verifies that \( C \Rightarrow \neg S \), so that in fact \( \neg S \Rightarrow C \).

**Theorem 5.** Suppose \( A \) has rank \( n \), and \( Ax = b \) for some \( x \geq 0 \). Then there is a \( y \geq 0 \) with \( Ay = b \), and \( y_i > 0 \) for at most \( n \) components.

**Proof:** Let \( I = \{i : x_i > 0\} \). If \( |I| > n \) then there is a nontrivial linear combination \( \sum_{i \in I} \alpha_i a_i = 0 \). Let \( I' = \{i \in I : \alpha_i < 0\} \); note that \( 0 \subseteq I' \subseteq I \). Let \( j \) be such that \( -x_j/\alpha_j \) is least among \( \{-x_i/\alpha_i : i \in I'\} \).
Adding \((-x_i/\alpha_i)a_i\) to \(x_i\) for all \(i \in I\) leaves \(Ax = b\) and \(x \geq 0\), and reduces the number of nonzero components.

This theorem is known as Caratheodory’s theorem; it has a “geometric” statement, as follows. Suppose \(S\) is a subset of \(\mathbb{R}^n\), and \(x \in \text{Cone}(S)\). Then there is a subset \(I \subseteq S\) with \(x \in \text{Cone}(I)\), and \(|I| = k\) where \(k\) is the dimension of \(\text{Span}(S)\). If \(A\) does not have full rank, it may be multiplied on the left by an invertible transformation, and 0 rows ignored, yielding a full rank system with the same solutions. Thus, there is a \(y\) with \(y_i > 0\) for at most \(n'\) components, where \(n'\) is the rank of \(A\).

**Theorem 6.** Suppose \(A\) is full rank, \(c \in \mathbb{R}^m\), \(S\) is the polyhedron \(\{y : y'Ax \leq c^t\}\), and \(y \in S\). The following are equivalent.

a. \(\{y\}\) is a face of \(S\).

b. \(y\) is an extreme point of \(S\).

c. Let \(I = \{i : y' a_i = c_i\}\); then \(|I| = n\) and \(\{a_i : i \in I\}\) is linearly independent.

**Proof:** As observed above, \(a \Rightarrow b\) in any real linear space. For \(I\) as in \(c\), if \(|I| < n\) or the \(a_i\) are linearly dependent, there is a nonzero \(w\) with \(w'A_I = 0\). Since \(y' a_i < 0\) for \(i \notin I\), for \(\epsilon\) sufficiently close to 0 \(y + \epsilon w \in S\). Thus, \(b \Rightarrow c\). Suppose \(|I| = n\) and the \(a_i\) are linearly dependent. Let \(v = \sum_{i \in I} a_i\). Suppose 
\[z^tA \leq c\] and 
\[z^tv = y^tv\] then \(z^t a_i \leq c_i\) for \(i \in I\), and 
\[\sum_{i \in I} z^t a_i = \sum_{i \in I} y^t a_i = \sum_{i \in I} c_i\], so \(z^t a_i = c_i = y^t a_i\) for \(i \in I\). It follows that \(z = y\); this shows that \(c \Rightarrow a\).

A point \(y \in S\) satisfying the conditions of the theorem is called a vertex.

**Theorem 7.** Suppose \(A\) has a rank \(n\). If \(b \notin \text{Cone}(\{a_i\})\) then there is a hyperplane containing \(n - 1\) of the \(a_i\), separating \(b\) from \(\text{Cone}(\{a_i\})\).

**Proof:** Suppose \(I \subseteq \{1, \ldots, n\}\) is such that \(|I| = n\) and \(\{a_i : i \in I\}\) is linearly independent. Iterate the following steps.

1. Let \(\lambda_j\) for \(1 \leq j \leq n\) be such that \(b = \sum_j \lambda_j a_{i_j}\); by hypothesis \(\lambda_j < 0\) for some \(j\).
2. Let \(k\) be the least \(j\) such that \(\lambda_k < 0\). Let \(v\) be normal to the span of \(\{a_i : i \in I, i \neq k\}\), chosen so that \(v \cdot a_k = 1\). Then \(v \cdot b = \lambda_k < 0\).
3. If \(v \cdot a_i \geq 0\) for all \(i\) then terminate the iteration.
4. Let \(l\) be the least \(i\) such that \(v \cdot a_i < 0\). Replace \(i_k\) by \(l\) in \(I\).

To show that the procedure terminates it suffices to show that \(I\) never repeats, so suppose otherwise. We may find \(l, s, t\) such that \(l\) leaves \(I\) during iteration \(s\), \(l\) next enters \(I\) during iteration \(t\), and no \(i\) with \(i > l\) is added to or removed from \(I\) between \(s\) and \(t\). Let \(b = \sum_j \lambda_j a_{i_j}\) during \(s\). Let \(v'\) be \(v\) during \(t\). Then \(v' \cdot b < 0\), if \(i_j < r\) then \(\lambda_j \geq 0\) and \(v' \cdot a_{i_j} \geq 0\), if \(i_j = r\) then \(\lambda_j < 0\) and \(v' \cdot a_{i_j} < 0\), and if \(i_j > r\) then \(v' \cdot a_{i_j} = 0\) (because \(i_j\) is in \(I\) during \(t\)). This is a contradiction.

If \(A\) does not have full rank, it may be multiplied on the left by an invertible transformation, and 0 rows ignored, provided \(b\) lies in the span of the columns. If \(b\) does not lie in the span of the columns then the conclusion of the theorem is trivial.

A polyhedral cone is defined to be the intersection of finitely many halfspaces. It is not difficult to see that a polyhedron which is a cone is a polyhedral cone (exercise 2).

With \(A, b\) as usual, let \(C_R = \{b : \exists x(Ax = b \land x \geq 0)\}\), and let \(C_L = \{y : y' A \leq 0\}\). Using Farkas’ lemma, and the fact that \(y' A \leq 0\) iff \(y' A x \leq 0\) for all \(x \geq 0\), it is readily verified that \(C_L^{\text{pol}} = C_R\) and \(C_R^{\text{pol}} = C_L\).

**Corollary 8.** A set \(S \subseteq \mathbb{R}^n\) is a polyhedral cone iff \(S = \text{Cone}(a_1, \ldots, a_m)\) for some \(a_1, \ldots, a_m\).
3. Linear programming. A linear program is defined to be a linear optimization problem of the form, minimize (or maximize) $c^T x$ for $x \in S$, where $S$ is a polyhedron. A point $x \in S$ is called feasible. A point $x_0 \in S$ is called optimal if $c^T x_0 \leq c^T x$ for all $x \in S$.

A linear program in standard form is defined to be one where $S$ equals $\{ x : A x = b, x \geq 0 \}$. Here $A$ is an $n \times m$ matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}^m$. If $A$ is not full rank, the problem be transformed by row operations so that it is, so there is no loss of generality in requiring it to be. For the remainder of the section, $A$ is assumed to have full rank.

A solution to $A x = b$ with $x \geq 0$ is called a feasible solution. If $I \subseteq \{ 1, \ldots, m \}$ is such that $|I| = n$ and $\{ a_i : i \in I \}$ is linearly independent, then there is a unique $x$ such that $A x = b$ and $x_i = 0$ for $i \notin I$. Such a solution is called a basic solution. Caratheodory’s theorem is equivalent to the statement that, if $A x = b$ has a feasible solution then it has a basic feasible solution.

The feasible region $S$ of a linear program in standard form is a polyhedron, lying in the $m-n$-dimensional subspace $A x = b$. A point $x \in S$ is a vertex if it is a basic feasible solution. This may be proved by a variation of the proof of theorem 6. Let $I = \{ i : x_i > 0 \}$, and suppose $I$ is linearly dependent, so that there is a nonzero $w$ with $A_I w = 0$; then for $\epsilon$ sufficiently close to $0$, $x + \epsilon w \in S$. If $|I| < n$ $I$ can be enlarged to a set $J$ with $|J| = n$ and $\{ a_i : i \in J \}$ linearly independent. This shows that an extreme point is a basic feasible solution. If $|I| = n$ and $\{ a_i : i \in I \}$ is linearly independent let $v_i = 1$ if $i \notin I$, else let $v_i = 0$. If $v^T x = 0$, $A x = b$, and $x \geq 0$ then $x_i = 0$ for $i \notin I$, so $x$ is the basic feasible solution determined by $I$. This shows that a basic feasible solution is a vertex.

Suppose $x_o$ is an optimal point, and $c^T x_o = \chi$. The equation $c^T x = \chi$ may be added to the equations for the feasible set $S$. The resulting polyhedron is a face $F$ of $S$ (possibly all of $S$). By hypothesis the equations for $F$ have a feasible solution, hence they have a basic feasible solution. This is a vertex $F$, hence a vertex of $S$, hence a basic feasible solution to the linear program; also it has optimal cost. We have shown that if a linear program in standard form has an optimal solution then it has a basic optimal solution, a fact known as the fundamental theorem of linear programming.

Say that a linear program is feasible if it has a feasible point. Since there are only finitely many basic feasible solutions, it follows by the fundamental theorem of linear programming that if a problem in standard form is feasible, then either it has an optimal point (in which case it is called solvable), or there are feasible points of arbitrarily small cost (in which case the problem is called unbounded).

The simplex method is a method for searching through the basic feasible solutions for an optimal solution. An outline is as follows. Assume that a basic feasible solution $x$ is known; we will see how to remove this assumption later.

Suppose $I \subseteq \{ 1, \ldots, n \}$ is such that $|I| = n$ and $\{ a_i : i \in I \}$ is linearly independent, and let $x$ be the corresponding basic feasible solution. Iterate the following steps.

1. For each $j \notin I$ let $\xi_j$ be such that $\xi_{j j} = 1$ and $A \xi_j = 0$.
2. If $c^T \xi_j \geq 0$ for all $j$ then $x$ is optimal; terminate the iteration.
3. Let $j$ be least such that $c^T \xi_j < 0$.
4. If $\xi_{j i} \geq 0$ for all $i$ then the cost is unbounded below; terminate the iteration.
5. Let $\theta = \min \{ -x_i/\xi_{ji} : \xi_{ji} < 0 \}$. Let $i$ be least such that $-x_i/\xi_{ji} = \theta$. Replace $i$ by $j$ in $I$. 288
If $Ay = b$ then $y = x + \sum_{j \notin I} y_j \xi_j$ (indeed, $\{x_{ij} : j \notin I\}$) is a linearly independent set of size $m - n$, so a basis for the nullspace of $A$, and the coefficient of $\xi_j$ in $y - x$ is clearly $y_j$). Letting $\xi_j = 0$ for $j \in I$ the sum may be taken over all $j$.

It follows that if $c^t \xi_j \geq 0$ for all $j$ then $x$ is optimal. It remains to show that the iteration terminates. We may find $p$, $q$, $s$, and $t$ such that $q$ enters $I$ during iteration $s$, $q$ next leaves $I$ during iteration $t$, and $p$ enters $I$ during $t$, no $i$ with $i > q$ is added to or removed from $I$ between $s$ and $t$, and the cost after $t$ is the same as that before $s$. Let $x$ and $\xi_j$ denote these quantities during $s$, and $\hat{x}$ and $\hat{\xi}_j$ during $t$. Let $w_t = \hat{x} + \alpha \xi_p$ where $\alpha > 0$, and let $w_s = x + \alpha \xi_p$. Then $c^t \hat{\xi}_p = c^t (w_t - x) = c^t (w_s - x) = \sum_j w_s c^t \xi_j$ (the last equality follows because $\hat{\xi}_p$ is in the nullspace of $A$). The first quantity is negative. However, if $j < q$ then $c^t \xi_j \geq 0$, and $w_{sj} \geq 0$ also since it equals $x_j + \alpha \xi_{pq}$. Also, $c^t \xi_q < 0$, and $w_{sq} = \hat{\xi}_{pq} < 0$. This is a contradiction.

The computations during the simplex algorithm can be arranged into a “tableau”, resulting an an efficient implementation. The initial tableau may be computed by row operations, also ensuring full rank. The choice of $j$ and $i$ as above is called “Bland’s anticycling algorithm”; with other choices the simplex algorithm can enter an infinite loop. See any of numerous references for further discussion, for example [PapSteig].

A feasible $x$ to start the iteration may sometimes be provided by the formulation of the problem. If not, the “two stage” simplex algorithm may be used. In the first stage, rows are multiplied by $-1$ if necessary to ensure $b \geq 0$, $n$ “artificial” variables $x_{m+1}, \ldots, x_{m+n}$ are added, and $A$ is extended with an $n \times n$ identity matrix on the right; the cost is $\sum_{i > m} x_i$. If the cost after the first stage is not 0, the problem is not feasible. Otherwise, let $I_1 \subseteq \{1, \ldots, m + n\}$ be the final basic columns, let $I_2 = I_1 \cap \subseteq \{1, \ldots, m\}$, and let $I$ be any extension of $I_2$ with linearly independent columns. ($I$ is readily computed from the final tableau of the first stage).

In a linear programs in “general form”, there may be both equality constraints $a^t x = b$, and inequality constraints $a^t x \leq b$. Also, a component of $x$ may either be constrained by the requirement $x_i \geq 0$, or unconstrained. Linear programs in general form may be transformed to standard form, as follows.

Let $M$ be the matrix whose rows are the $a^t$ of the constraints, with the first $n'$ rows being those of the equality constraints. Here we do not require $M$ to have full rank. Suppose also that the columns are ordered so that the first $m'$ correspond to the constrained components. Let $M_C$ be the first $m'$ columns, and $M_U$ the remaining columns. The transformed problem has matrices and vectors as follows.

$$A_s = \begin{bmatrix} M_C & M_U \end{bmatrix} - M_U \begin{bmatrix} 0 \\ I \end{bmatrix}$$

$$x_s = \begin{bmatrix} x_C [x_U^t x_U | x_S]^t \\ c_s = [cc^t | uv] - cv \end{bmatrix}$$

Here an unconstrained variable $x_i$ is “split” into two constrained variables $x_i^+$ and $x_i^-$; and $x_S$ is $n - n'$ new “slack” variables for the inequality constraints.

The feasible points of the general form problem are in bijective correspondence with the feasible points of the transformed problem, and the correspondence preserves the cost. It follows that a linear program in general form is either infeasible, unbounded, or solvable.

For the following we need a fact about matrices. Suppose $A$ is a full rank $n \times m$ matrix with $m \geq n$, $J = \{i_1 < \cdots < i_n\}$ is such that $\{a_{ij} \}$ is linearly independent, and $\Lambda$ is an $m \times m$ matrix such that row $i$ of $\Lambda$ is 0 for $i \notin J$. Let $A_J$ be the matrix whose $j$th column is $a_{ij}$. Let $\Theta$ be the $m \times n$ matrix whose $i$th row is the $j$th row of $A_J^{-1}$, and whose other rows are 0. One verifies that $\Theta A \Lambda = \Lambda$.

Let $A$ be as in the simplex algorithm, with $J$ the column indices for a basic feasible solution. Let $\Xi$ be the matrix whose $j$th column is $\xi_j$. With $\Theta$ as in the preceding paragraph, $\Theta A (I - \Xi) = I - \Xi$ where $I$ is the identity matrix. Since $A \Xi = 0$, also $\Theta A (I - \Xi) = \Theta A$; thus, $\Theta A = I - \Xi$.

289
The “dual” linear program to a linear program in standard form, called the primal problem, introduces a vector \( y \) of \( n \) dual variables. The problem is, to maximize \( y' b \) subject to \( y' A \leq c' \).

**Theorem 9.**

a. If the primal problem is solvable then the dual problem is solvable, and the optimum cost for both problems is the same.

b. The dual of the dual is the primal.

c. If the dual problem is solvable then the primal problem is solvable, and the optimum cost for both problems is the same.

d. If \( x \) is feasible in the primal problem, and \( y \) in the dual problem, then \( y' b \leq c' x \).

e. If the primal and dual problems are both feasible then they are both solvable.

f. \( x \) and \( y \) are both optimal iff \( y' b = c' x \), iff \((c' - y'A)x = 0\).

These facts hold even if \( A \) is not full rank.

**Proof:** In the full rank case, if the primal problem is solvable let \( J \) and \( \Xi \) be as in the preceding paragraph for an optimal basic solution \( x \). As noted above, \( c' \Xi \geq 0 \), whence \( c' \Theta A = c'(I - \Xi) \leq c' \); that is, \( c' \Theta \) is a feasible point for the dual problem. The cost in the dual problem of this point is

\[
c' \Theta b = c' \Theta Ax = c'(I - \Xi)x = c' x - c' \Xi x = c' x.
\]

This proves part a. The proof of part b is left to exercise 3, and part c follows from parts a and b. Part d follows because \( y' b = y' Ax \leq c' x \), and part e follows. \( x \) and \( y \) are both optimal iff \( y' b = y' Ax = c' x \), proving part f. If \( A \) is rank \( n' < n \) there is an invertible matrix \( M \) so that \( MA \) has \( n - n' \) zero rows; for the primal problem to be feasible \( Mb \) must also be 0 in these rows. In the problem ignoring these rows, there is a solution to the dual problem. Adding any values in the ignored rows, and transforming by \( M^{-1} \), yields a solution to the original dual problem. The remaining parts require few if any changes.

The equation \((c' - y'A)x = 0\) is called the complementary slackness conditions; these state that at optimality, \( x_i > 0 \) and \( y\sum_i a_{ij} < c_j \) cannot both hold. The complementary slackness conditions are an instance of a more general set of equations, the “KT”, or “KKT” conditions. (KT is an abbreviation for Kuhn-Tucker; KKT adds the name of Karush, who was considered to have been omitted). These apply to more general “nonlinear” optimization problems, and have various formulations depending on the type of problem.

In one formulation, functions \( f : \mathbb{R}^n \to \mathbb{R} \), \( g : \mathbb{R}^n \to \mathbb{R}^m \), and \( h : \mathbb{R}^n \to \mathbb{R}^k \) are given. The “Lagrangian” for the problem, minimize \( f(x) \) subject to \( g(x) = 0 \) and \( h(x) \geq 0 \), is \( f + \sum_{i=1}^m \lambda_i g_i + \sum_{i=1}^k \mu_i h_i \). The KKT conditions are

- the constraints hold,
- the gradient of the Lagrangian vanishes,
- \( \mu_i \geq 0 \), and
- \( \mu_i h_i(x) = 0 \).

The last of these is the complementary slackness conditions.

For a linear program in standard form, the Lagrange multipliers \( \lambda_i \) are the “dual variables” \( y \), and the Lagrange multipliers \( \mu_i \) are the “dual slack variables” \( s = c' - y'A \). A local optimum is a global optimum, and the KKT conditions are necessary and sufficient for optimality. For further information see any of numerous textbooks on optimization theory, for example [SimmonsD].

The simplex method as presented above, for a linear program in standard form, is called the primal simplex method. There are variations of the simplex method, which use the problem in other forms. For example the “dual simplex” method uses the dual of the standard form, and “primal-dual” methods use both the standard form and its dual. Primal-dual methods have become important in “interior point”
algorithms, which do not stay on the boundary of the feasible region. Although the simplex method is quite fast in practice, no variation has been found, which runs in polynomial time on rational inputs. However there are polynomial time interior point methods. References for further information on these topics include \cite{PapSteig}, \cite{Schriver}, and \cite{Wright}.

4. Polytopes.

**Theorem 10.** $S \subseteq \mathbb{R}^n$ is a polyhedron iff $S = C + P$ where $C$ is a polyhedral cone and $P = \text{Chull}(P_0)$ for a finite set $P_0$.

**Proof:** Suppose $S = \{ y : y^t A \leq b \}$. Let $K = \{ \langle x, \lambda \rangle : Ax \leq \lambda b \}$ be the cone in $\mathbb{R}^{n+1}$ corresponding to $S$. By corollary 8 $K = \text{Ccone}(a_1, \ldots, a_m)$ for some $a_1, \ldots, a_m$, where we may assume $a_i$ is $\langle a'_i, \sigma_i \rangle$ and $\sigma_i$ is 0 or 1. Let $P_0 = \{ a'_i : \sigma_i = 1 \}$, and let $C = \text{Ccone}(a'_i : \sigma_i = 0)$. Then $x \in P$ iff $\langle x, 1 \rangle \in K$, iff $\langle x, 1 \rangle = \sum_i \alpha_i a_i$ where $\alpha_i \geq 0$. The linear combination yields a decomposition of $x$ as $x_1 + x_2$ where $x_1 \in C$ and $x_2 \in P$, and conversely. Conversely, suppose $S = C + P$ as in the theorem, say $C = \text{Ccone}(a_1, \ldots, a_r)$ and $P = \text{Chull}(a_{r+1}, \ldots, a_m)$. Then $x \in S$ iff $\langle x, 1 \rangle \in K$ where $K = \text{Ccone}(\{ a_i, \sigma_i \})$, where $\sigma_i = 0$ for $i \leq r$ and $\sigma_i = 1$ for $i > r$. Again by corollary 8, $K = \{ \langle x, \lambda \rangle : Ax \leq \lambda b \}$ for some $A$ and $b$; whence $S = \{ x : Ax \leq b \}$.

**Corollary 11.** A polyhedron is bounded iff it is the convex hull of a finite set of points.

**Proof:** If a polyhedron $S$ is bounded, the cone as in the theorem must be $\emptyset$. Conversely by the theorem the convex hull of a finite set of points is a polyhedron, and it is clearly bounded.

A polyhedron as in the corollary is called a polytope. By exercise 1 and theorem 6 a polytope $P$ is the convex hull of its vertices $V$.

The faces of a polytope are a subject of diverse interest. First we prove some basic facts about the faces of a polyhedron.

**Theorem 12.** Let $P$ be the polyhedron $\{ y : y^t A \leq c \}$ where $A$ is an $n \times m$ matrix. Then $F$ is a face of $P$ iff $F = \{ y \in P : y^t a_i = c_i \text{ for } i \in I \}$, where $I \subseteq \{1, \ldots, m\}$ and $F$ is nonempty.

**Proof:** For $P$ itself let $I = \emptyset$. The proper faces of $P$ are the sets of the form $F = \{ y \in P : y^t b = v \}$ for some $b \in \cap \mathbb{R}^n$ and $v \in \mathbb{R}$, where $y^t b \leq v$ for all $y \in P$ and $y^t b = v$ for some $y \in P$. Given $I$, let $x_i$ be 1 in components $i \in I$ and 0 otherwise. Suppose $F = \{ y : y^t Ax_i = cx_i \}$, and let $b = Ax_i$. Then $y^t b \leq c x_i$, so the dual problem is bounded; let $v = \max(y^t b)$. Then for $y \in P$, $y \in F$ iff $y^t Ax_i = cx_i$ iff $y^t b = v$, using theorem 9. Conversely, given $b$ and $v$, the dual problem with these values is solvable, so by theorem 9 there is an $x \geq 0$ with $Ax = b$ and $c x = v$. Let $I = \{ i : x_i > 0 \}$; then $y^t b = v$ iff $y^t Ax = c x$ iff $y^t a_i = c_i$ for $i \in I$.

From the theorem we may conclude the following.

- There are only finitely many faces.
- The intersection of two faces is a face.
- The faces, together with the empty set, partially ordered by inclusion, are a lattice.
- If $F$ is a face then the subfaces of $F$ are the subfaces of $P$ contained in $F$.

It might seem more symmetric to call $\emptyset$ a face; but the terminology as stated is fairly standard.

By theorem 6 if $A$ has full rank then the minimal faces of $P$ are vertices, and $A$ has vertices. If $A$ does not have full rank, it may be multiplied on the left by an invertible transformation. The minimal faces with 0 rows ignored are vertices, and the minimal faces of $P$ are affine subspaces of dimension $n - n'$ where $n'$ is the rank of $A$. In particular, $P$ has no proper faces iff it is an affine subspace.

If the restriction $y^t a_i = c_i$ has no effect, then $P$ is already contained in this affine subspace; let $I_\subset$ be the set of such $i$. For $i \notin I_\subset$ there is a $y_i \in P$ with $y_i^t a_i < c_i$; $\sum_{i \notin I_\subset} y_i / (m - |I_\subset|)$ is in the interior of $P$, and so the affine span of $P$ is the intersection of the affine subspaces $y_i^t a_i = c_i$ for $i \in I_\subset$.
The dimension of a face is defined to be the dimension of its affine span. A maximal proper face is called a facet. The facets of $P$ are exactly the faces $P \cap \{ y : y^t a_i = c_i \}$ for $i \notin I_\omega$. Further, the dimension of a facet is $d - 1$ where $d$ is the dimension of $P$, since the affine span is intersected with a hyperplane not containing it.

A poset is said to be ranked if all its maximal chains have the same length. By the foregoing, the poset of faces of a polyhedron ordered by inclusion is ranked, with the length of a maximal chain being $d - d'$ where $d$ is the dimension of $P$, and $d'$ is the dimension of the minimal faces.

Suppose $H$ is a supporting hyperplane of a polytope $P$, and $F = P \cap H$ is a face. Clearly $\text{Chull}(V \cap H) \subseteq F$. On the other hand $F$ is the convex hull of its vertices, which are vertices of $P$, whence in $V \cap H$, whence $F = \text{Chull}(V \cap H)$. In particular $F$ is a polytope.

For a convex set $S$ in $\mathbb{R}^n$ (although more general spaces could be considered) let $S_{\text{pol}} = \{ y : y^t x \leq 1 \}$ for all $x \in S$. Let $C$ be the convex sets, ordered by inclusion. One verifies that $S \mapsto S_{\text{pol}}$ is an order preserving map from $C$ to $C_{\text{op}}$. Further, $S \subseteq T_{\text{pol}}$ iff $y^t x \leq 1$ for all $x \in S$ and $y \in T$ iff $T \subseteq S_{\text{pol}}$. Thus, the pair of maps $S \mapsto S_{\text{pol}}$ from $C$ to $C_{\text{op}}$, and from $C_{\text{op}}$ to $C$, form a Galois adjunction. By exercise 2 this operation applied to a convex cone produces the same result as that of the earlier defined operation.

**Theorem 13.** If $S$ is a closed convex set in $\mathbb{R}^n$ containing the origin then $(S_{\text{pol}})^{\text{pol}} = S$.

**Proof:** Let $x_1, x_0$ and $v$ be as in the proof of theorem 4. Then $v^t(x - x_0) < 0$ for $x \in S$; since $0 \in S$, $v^t x_0 > 0$. Letting $v_1 = v/(v^t x_0)$, $v_1^t x < 1$ for $x \in S$, showing $v_1 \in S_{\text{pol}}$. From $v^t(x - x_0) > 0$ it follows that $x_1^t v_1 > 1$, showing that $x_1 \notin (S_{\text{pol}})^{\text{pol}}$.

**Theorem 14.** If $P$ is a polyhedron then $P_{\text{pol}}$ is a polyhedron. $P_{\text{pol}}$ is a polytope iff $0 \in P_{\text{int}}$.

**Proof:** Using theorem 10 write $P$ as $\text{Cone}(a_1, \ldots, a_m) + \text{Cone}(b_1, \ldots, b_l)$. Let $Q = \{ y : y^t a_i \leq 0 \text{ for all } i \} \cap \{ y : y^t b_i \leq 1 \text{ for all } i \}$. If $y \in P_{\text{pol}}$ then $y^t a_i \leq 1/r$ for any $i$, and $y^t b_i \leq 1$ for any $i$; thus, $y \in Q$. If $y \in Q$ and $x \in P$ write $x$ as $\sum_i \alpha_i a_i + \sum_i \beta_i b_i$ where $\alpha_i, \beta_i \geq 0$ and $\sum_i \beta_i = 1$. It follows that $y^t x \leq 1$. We prove the second claim for $P_{\text{pol}}$; $0 \in (P_{\text{pol}})^{\text{int}}$ iff $l$ may be taken as 0, iff $P$ is a polytope.

**Theorem 15.** Suppose $P$ is a polytope with $0 \in P_{\text{int}}$. For $F$ a proper face of $P$ let $\hat{F} = \{ y \in P_{\text{pol}} : y^t x = 1 \text{ for all } x \in F \}$. The map $F \mapsto \hat{F}$ is an order reversing bijection from the poset of proper faces of $P$ to the poset of proper faces of $P_{\text{pol}}$.

**Proof:** If $F = \text{Chull}(x_1, \ldots, x_r)$ let $x_0 = (1/r) \sum_i x_i$, and let $G$ be the face $\{ y \in P_{\text{pol}} : y^t x_0 = 1 \}$ of $P_{\text{pol}}$. If $y \in \hat{F}$ then $y^t x_i = 1$ for $i \geq 1$, and $y \in G$ follows. If $y \in G$ then $y^t x_i \leq 1$ for $i \geq 1$, whence $y^t x_i = 1$ for $i \geq 1$, whence $y \in \hat{F}$. That $F \mapsto \hat{F}$ is order reversing is straightforward. If $G$ is a face of $P_{\text{pol}}$ let $F = \{ x \in P : y^t x = 1 \text{ for all } y \in G \}$. By what has already been shown, $F$ is a face of $P$. Clearly $G \subseteq \hat{F}$. There is an $x \in \mathbb{R}^n$ such that $y^t x \leq x$ for all $y \in P_{\text{pol}}$, and $y^t x = 1$ iff $y \in G$. We have $x \in F$, whence $y^t x = 1$ for all $y \in \hat{F}$, whence $y \in G$ for all $y \in F$.

### 5. Krein-Rutman theorem in Euclidean space

If $K \subseteq \mathbb{R}^n$ is a convex cone then as is easily seen $K \cap -K$ is the largest subspace contained in $K$. In particular $K \cap -K = \{ 0 \}$ iff $K$ contains no line through the origin, iff $0$ is an extreme point. In this case $K$ is said to be pointed. By general facts about convex sets, if $\{ 0 \}$ is a face then $K$ is pointed. The converse is true if $K$ is closed (exercise 4).

A cone is said to be solid if its interior is nonempty.

We need some preliminary facts for the proof of the next theorem. Suppose $z$ is a complex number which is not a nonnegative real number. Then $\sum_{i=1}^n q_i z^i = 0$ for some $n$ and positive real numbers $q_i$. We may suppose $z = e^{i \theta}$. If $\theta = 2\pi j/k$ where $j/k$ is in lowest terms and $k \leq 4$ the claim is clear. Otherwise $n$ may be taken large enough that there is a $z^i$ in the interior of each quadrant. Remaining details are left to the reader.
If \( t_i = c_i(r_m)\lambda_i^r \), \( i = 1, 2 \), are two terms with \( m_i, r \in N \) and \( c_i, \lambda_i \in C \), then if \( |\lambda_2| < |\lambda_1| \), or \( |\lambda_2| = |\lambda_1| \) and \( m_2 < m_1 \), then \( t_1 \) “dominates” \( t_2 \). That is, \( \lim_{t \to \infty} t_2/t_1 = 0 \) (for those familiar with the “little oh” notation, \( t_2 = o(t_1) \)). Again, details are left to the reader.

Let \( M \) be an \( n \times n \) real matrix. Recall from section 10.4 that \( M \) is similar over \( R \) to its rational canonical form. The elementary divisors are clearly of the form \( p^r \) where \( p \) is linear, or real quadratic with conjugate roots; and \( e \) is an integer. From section 10.5, \( M \) is similar over \( C \) to its Jordan canonical form. Given a block \( B \) of either form let \( \alpha \) be the eigenvalue, let \( m \) be the dimension, and let \( v \) be a vector \( v \) such that \( \{x^i v: 0 \leq i < m\} \) is a basis for the invariant subspace on which \( B \) acts.

If \( \alpha \) is real then a Jordan block \( B \) over \( R \), with real vector \( v \), is a Jordan block over \( C \), and \( v \) again generates the invariant subspace, which may be taken as having a real basis.

If \( \alpha \) is complex write the dimension of the rational canonical form block \( B \) over \( R \) as \( 2m \), so that the period is \( (x - \alpha)^m(x - \bar{\alpha})^m \). The blocks of the Jordan canonical form over \( C \) of \( B \) may be grouped into “conjugate” pairs, with generating vectors \( w \) and \( \bar{w} \), and hence ordered bases whose corresponding elements are complex conjugate. In fact, there are only two blocks; if \( k \) is the dimension of the largest, and \( k < m \), then \( B \) would satisfy the real polynomial \( (x - \alpha)^k(x - \bar{\alpha})^k \), a contradiction.

Let \( \alpha_i \) denote the eigenvalue of block \( i \) of the Jordan canonical form, and \( m_i \) the dimension. We assume the blocks are numbered \( 1, \ldots, k \) so that \( |\alpha_i| \) is nondecreasing, and if \( |\alpha_{i+1}| = |\alpha_i| \) then \( m_i \geq m_{i+1} \). The spectral radius of a matrix \( M \) over \( C \) is defined to be the maximum of \( |\alpha| \) over the eigenvalues \( \alpha \).

If \( M \) is an \( n \times n \) real matrix and \( S \subseteq R^n \) let \( MK \) denote as usual \( \{Mx: x \in S\} \).

**Theorem 16.** Suppose \( K \subseteq R^n \) is a solid, closed, and pointed cone. Suppose \( M \) is a real \( n \times n \) matrix such that \( MK \subseteq K \). Then

a. the spectral radius \( \rho \) of \( M \) is an eigenvalue of \( M \);

b. \( \alpha_1 \) in the Jordan canonical form ordered as above may be taken as \( \rho \); and

c. there is an eigenvector of \( \rho \) in \( K \).

**Proof:** Suppose \( T^{-1}MT = J \) where \( J \) is the Jordan canonical form. Let \( t_{ij} \) be the \( j \)th column of \( T \) associated with block \( i \) for \( 1 \leq j \leq m_i \), and \( t_{ij} = 0 \) otherwise. Since \( K \) is solid there is a \( z_1 \in K^{\text{int}} \), say \( z_1 = \sum ij c_{ij} t_{ij} \). We assume that the \( t_{ij} \) are taken so that they are real or occur in conjugate pairs; then \( \delta \sum t_{ij} \) is real, and for an appropriate \( \delta, z \in K^{\text{int}} \) where \( z = \sum ij c_{ij} t_{ij} \), \( c_{ij} = c_{ij}^1 + \delta \), and \( c_{ij} \neq 0 \). Now,

\[
Mt_{ij} = \alpha_i t_{ij} + t_{i,j-1},
\]

so by induction

\[
M^r t_{ij} = \sum_{s=0}^{j-1} \left( \begin{array}{c} r \\ s \end{array} \right) \alpha_i^r t_{i,j-s},
\]

whence

\[
M^r z = \sum_{ij} c_{ij} \sum_{s=0}^{j-1} \left( \begin{array}{c} r \\ s \end{array} \right) \alpha_i^r t_{i,j-s}.
\]

By remarks preceding the theorem, as \( r \to \infty \)

\[
M^r z \to \sum_{i=1}^{h} c_{im} \left( \begin{array}{c} r \\ m-1 \end{array} \right) \alpha_i^{r-m+1} t_{i1} = \left( \begin{array}{c} r \\ m-1 \end{array} \right) \rho^{r-m+1} \sum_{i=1}^{h} \beta_i t_{i1},
\]

where \( B_1, \ldots, B_h \) are the blocks with \( |\alpha_i| = \rho \); \( m_i = m_i = m \); and \( \beta_i = c_{im} e^{i\theta_i} \) for some \( \theta_i \). Since \( \beta_i \neq 0 \) and the \( t_{ij} \) are linearly independent, \( \sum_{i=1}^{h} \beta_i t_{i1} \neq 0 \); let this quantity be denoted \( v_h \). Since \( MK \subseteq K \), \( M^r z \in K \) for all \( r \). Since \( K \) is closed its intersection with the unit sphere \( S^{n-1} \) is compact. It follows that
$v_h \in K$. If $\alpha_h \neq \rho$, by remarks preceding the theorem $\sum_l q_l \alpha^l_h = 0$ for some positive real numbers $q_l$. Let $v_{h-1} = \sum_l q_l M^l \alpha^l_h$. Since $q_l M^l x_h \in K$, $v_h \neq 0$, and $K$ is pointed, $v_{h-1}$ is a nonzero vector in $K$. Also,  

$$v_{h-1} = \sum_l q_l \sum_{i=1}^h \beta_{ih} \alpha^l_i x_{i1} = \sum_{i=1}^h \beta_{ih} \sum_l q_l \alpha^l_i x_{i1} = \sum_{i=1}^{h-1} \beta_{i,h-1} x_{i1}$$

where $\beta_{i,h-1} = \beta_{ih} \sum_l q_l \alpha^l_i$. Continuing, a value $f \geq 1$ must be reached which is the largest $f \leq h$ such that $\alpha_f = \rho$; for if $v_1 \neq \rho$ then $v_0 = 0$ and $v_0 \neq 0$ both hold, a contradiction. This proves parts a and b; further,  

$$M v_f = \sum_{i=1}^f \beta_{if} M x_{i1} = \rho v_f,$$

proving part c.

In an important special case, $K$ is the “nonnegative orthant” $\{ x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i \}$. A matrix $M$ leaves this invariant iff its entries are all nonnegative, in which case $M$ is said to be nonnegative. For further information see [BerPlem].

**Exercises.**

1. Prove the facts about extreme points given in the text. Hint: For fact 1, if $x$ is not an extreme point of $F$ then clearly $x$ is not an extreme point of $S$. Suppose $x$ is an extreme point of $F$, and $F = S \cap H$. Suppose $x$ is in the interior of the line segment from $u$ to $v$ in $S$. Since at least one of $u,v$ is not in $H$ they must lie on opposite sides of $H$, a contradiction. For fact 2, if $\{ x \} = S \cap H$ then a line segment containing $x$ in its interior either lies in $H$, or has its ends on opposite sides. Fact 3 is proved by induction on $n$, the basis $n = 0$ being immediate. If $x$ is in the boundary of $S$ then $x$ lies in a face of $S$ and the induction hypothesis applies. Otherwise any line through $x$ intersects the boundary in two points, both of which by induction are in the convex hull of the extreme points.

2. Show that if a cone is contained in the halfspace $v^t x \leq b$ then it is contained in the halfspace $v^t x \leq 0$.

3. Show that the dual of the dual of a standard form (primal) linear program is the primal program. Hint: Transform the dual to standard form, first transposing it and minimizing $(-b)^t y$. Take the dual of this, and show it is equivalent to the primal problem.

4. Show that if $K$ is a pointed closed convex cone then $\{0\}$ is a face. Hint: If $K$ is a ray the claim is obvious. Otherwise there are points $P_1 \in K, P_2 \in -K$ determining a line $L$ which does not contain 0. This line contains a point $P$ which is not in $K$ or $-K$. Let $H_1$ be any hyperplane through $L$ not containing 0. In $H_1$, let $Q$ be the closest point of $K$ to $P$, and let $H_2$ be the subspace of $H_1$ normal to the line through $P$ and $Q$, containing $P$. Let $H$ be the hyperplane determined by $H_2$ and 0; this intersects $K$ in $\{0\}$.  

294
23. Point lattices.

1. Basic facts. The word “lattice” is used in two senses in mathematics, to denote a partial order with meets and joins (lattice order), and to denote a discrete additive subgroup (or sub-\(\mathbb{Z}\)-module) of Euclidean space (point lattice). In this chapter the term without qualification will be used in the latter sense, whereas in most of the text it is used in the former.

A lattice \(L\) in \(\mathbb{R}^n\) is defined to be a subgroup of the additive group, which is generated by a set \(\{g_1, \ldots, g_r\}\) of vectors, which are linearly independent over \(\mathbb{R}\). The matrix \(G\) whose columns are the \(g_i\) may be called a generator matrix; \(L = G\mathbb{Z}^r\), the integral linear combinations of the columns.

Often a lattice in \(\mathbb{R}^n\) is required to have rank \(n\); there are situations where the more general definition is useful, though. The restriction can always be met by considering a subspace. Indeed, if \(G\) is a full rank \(n \times r\) matrix where \(r \leq n\) then \(G\) has a factorization \(QG'\) where \(Q \in SO(n)\) and rows \(r + 1\) to \(n\) of \(G'\) are 0. \((SO(n)\) is defined in section 20.13; a QR factorization, defined in section 12, may be modified to yield the desired factorization.) The lattice in \(\mathbb{R}^n\) is thus geometrically congruent to a full rank lattice in \(\mathbb{R}^r\).

An additive subgroup of \(\mathbb{R}^n\) is readily seen to be discrete as a topological subspace iff for every \(x \in L\) there is an open ball \(B_{2\epsilon}\) which contains no other element of \(L\), iff there is an open ball \(B_0\) which contains no nonzero element of \(L\).

**Theorem 1.** A lattice \(L\) is discrete.

**Proof:** Reducing to full rank as noted above, the generator matrix \(G\) may be assumed to have an inverse \(G^{-1}\). \(|G^{-1}(x)|\) is a continuous function from \(\mathbb{R}^n\) to \(\mathbb{R}^n\). It has a maximum value on the closed unit ball \(B_{01}\). This shows that \(L\) has only finitely many points inside \(B_{01}\), and the theorem follows.

The converse, that a discrete additive subgroup is a lattice, will be proved below. There are direct proofs (see [Barvinok] for example), but we prefer to follow [Newman] and prove some facts of independent interest first.

A finitely generated subgroup of \(\mathbb{R}^n\) is not necessarily a lattice. Consider the subgroup of \(\mathbb{R}\) generated by 1 and an irrational \(\zeta\); it is readily seen that this is dense in the unit interval, so cannot be a lattice.

The lattices are partially ordered by inclusion. By the proof of theorem 8.6, a subgroup of a lattice is a lattice. The meet (sum) of two lattices need not be a lattice (consider the example in the previous paragraph). The meet (intersection) is a lattice, provided \(\{0\}\) is considered one.

We have used the term “integer lattice” already in chapter 8 for a sublattice of \(\mathbb{Z}^n\), although usage varies. In [Lagarias] the term “integral lattice” is used for this, whereas in [ConSl] an integral lattice is defined to be one where every vector has integer length. Another class of lattices of interest is those which are contained in \(\mathbb{Q}^n\), which might be called rational lattices. By theorem 8.6 any subgroup of \(\mathbb{Z}^n\) is a lattice. It follows from this that any finitely generated subgroup of \(\mathbb{Q}^n\) is a lattice; the generators can all be multiplied by a common denominator of all the rationals that occur.

It was observed in section 10.1 that a matrix over a ring \(R\) is invertible iff its determinant is a unit of \(R\). A matrix over \(\mathbb{Z}\) is called integral, and if its determinant is \(\pm 1\) it is called unimodular. The terminology is applied in more general settings; and as will be seen in section 4 theorems of interest for integral matrices may be proved over any principal ideal domain. We also call a matrix with entries in \(\mathbb{Z}\) an integer matrix.

**Theorem 2.** Two rank \(n \times r\) matrices \(G_1\) and \(G_2\) generate the same lattice iff there is an \(r \times r\) unimodular matrix \(U\) such that \(G_2 = G_1U\).

**Proof:** If \(U\) is unimodular then \(U^{-1}\) is unimodular; and \(U\mathbb{Z}^r = \mathbb{Z}^r\). Thus if \(U\) exists then \(G_2\mathbb{Z}^r = G_3UU^{-1}\mathbb{Z}^r = G_1\mathbb{Z}^r\). Conversely let \(g_i\) be column \(i\) of \(G_1\); then \(g_i = G_2u_i\) for some \(u_i \in \mathbb{Z}^r\). Let \(U\) be the matrix whose columns are the \(u_i\), so that \(G_1 = G_2U\). By symmetry \(G_2 = G_1V\) for some \(V\). Since \(G_i\) is full rank, \(UV\) is the identity.
By the theorem, in the full rank case \( r = n \), \(|\det(G)|\) depends only on the lattice \( L = G\mathbb{Z}^n \), and not on \( G \). This value is called the determinant of the lattice, and denoted \( \det(L) \). The fundamental region, or fundamental parallelepiped, of a generator matrix \( G \) is defined to be \( \{\sum \alpha_ig_i : 0 \leq \alpha_i < 1\text{ for all } i\} \). By theorem 6 below, in the full rank case \( \det(L) \) equals the volume of the fundamental region, regardless of the generator matrix.

The determinant can be defined in the rank deficient case \( r < n \), as follows. Let \( G_i = Q_iG_i' \) as above, for \( i = 1, 2 \). Then \( \det(G_iG_i') = \det(G_i'G_i) \). By the full rank case, the right side is the square of the determinant of the lattice, as a subspace of its span over \( \mathbb{R} \). Thus, define the determinant to be \( \sqrt{\det G_i'G_i} \), where recall \( G_i'G_i \) is the Gram matrix of \( G \).

Suppose \( G \) is a full rank generator matrix; then \( x^tG^tGx \) is a positive definite quadratic form, called the quadratic form of the lattice. If \( G_2 = QG_1 \) where \( Q \in SO(n) \) then the Gram matrices of \( G_2 \) and \( G_1 \) are the same, and so are the quadratic forms. On the other hand if \( G_1'G_1 = G_2'G_2 \) then \( (G_2^{-1})^{-1}G_1'G_1G_2^{-1} = 1 \), from which \( G_1G_2^{-1} \in SO(n) \) follows.

If \( M \) is a positive definite symmetric matrix, then there is a lattice which has \( M \) as its Gram matrix. Indeed, recall from chapter 10 the Cholesky decomposition \( G^tG \) of \( M \). \( G \) may be taken as the generator matrix of a lattice. Applications of this relation between lattices and quadratic forms may be found in [ConSi] for example.

2. Lebesgue measure. In various applications in Euclidean space it is convenient to have a definition of the volume of a subset \( S \subset \mathbb{R}^n \). The most comprehensive notion for such is the Lebesgue measure. There are less general notions; but a treatment of these is not much shorter than a treatment of the Lebesgue measure, so we will give the latter.

In \( \mathbb{R}^n \) define a cell to be a set of the form \( \{x : l_i \leq x_i < u_i \text{ for all } i\} \) for real numbers \( l_1 < u_1, \ldots, l_n < u_n \). Define \( \nu(C) \) for a cell \( C \) to be \((u_1 - l_1) \cdots (u_n - l_n)\).

Call a set of the form \( \{x \in \mathbb{R}^n : x_i = c\} \) for some \( i \) and \( c \) a cutting hyperplane. Such a hyperplane partitions \( \mathbb{R}^n \) into the sets \( \{x \in \mathbb{R}^n : x_i < c\} \) and \( \{x \in \mathbb{R}^n : x_i \geq c\} \). A cell \( C \) is partitioned into two parts \( C_1 \) and \( C_2 \) consisting of its intersection with these two sets (it is possible that one part is the whole cell and the other empty). Both parts are cells; it is clear that \( \nu(C) = \nu(C_1) + \nu(C_2) \). A bounding hyperplane of a cell \( \{x : l_i \leq x_i < u_i\} \) is defined to be any cutting hyperplane \( \{x : x_i = l_i\} \) or \( \{x : x_i = u_i\} \).

A collection of subsets of a set \( X \) which is closed under difference and pairwise union is called a ring of sets. Define a subset of \( \mathbb{R}^n \) to be elementary if it is the union of a finite disjoint collection of cells.

Lemma 3.

a. The elementary sets form a ring of subsets of \( \mathbb{R}^n \).

b. \( \nu \) can be extended to the elementary sets so that if \( E \) and is an elementary set then \( \nu(E) = \sum_i \nu(C_i) \)

   where \( \{C_i\} \) is any decomposition of \( E \) into disjoint cells.

c. If \( E \) and \( F \) are disjoint then \( \nu(E \cup F) = \nu(E) + \nu(F) \).

Proof: For part a, partition all the cells in \( E \) and \( F \) by the bounding hyperplanes of the cells in \( E \) and \( F \), and let \( H \) be the resulting collection of cells. It is readily seen that \( E \cup F \) or \( E - F \) is the union of a disjoint collection of cells from \( H \). It is also readily seen that if \( C = C_1 \cup \ldots \cup C_n \) for the \( C_i \) disjoint then \( \nu(C) = \nu(C_1) + \ldots + \nu(C_n) \). Part b follows, and also part c.

Extend the reals as in section 1.4 with \( \pm \infty \). Let \( \sum_i x_i \) be an infinite series of nonnegative extended reals. Say that the series converges to \( \infty \) if the partial sums \( s_i \) eventually become greater than any finite positive real. By modifying exercise 17.9, it follows that \( \sum_i x_i \) converges to a finite value or to \( \infty \), and that this value depends only on the \( x_i \) and not their order. From this, if \( x_{ij} \) are nonnegative extended reals, \( \sum_j x_{ij} = x_i \), and \( \sum_i x_i = x \), then \( \sum_i \sum_j x_{ij} = x \), where the order of the sum is irrelevant.
In particular, given a (finite or) countable disjoint collection \( F = \{C_i\} \) of cells, \( \sum \nu(C_i) \) converges to a finite value or to \( \infty \), and this value depends only on the set of cells and not the order in which they are listed. We denote it as \( \nu(F) \).

In this section by a cover \( K \) of a set \( S \subseteq \mathcal{R}^n \) is meant a countable disjoint collection of cells with \( S \subseteq \cup K \). Let \( \mu^*(S) \) be the greatest lower bound of the set of values \( \nu(K) \) over all covers \( K \) of \( S \). This value, a nonnegative real or \( \infty \), is called the outer measure of \( S \).

Define a packing \( P \) of a set \( S \subseteq \mathcal{R}^n \) to be a countable disjoint collection of cells \( C \subseteq S \). Given \( S \subseteq \mathcal{R}^n \) let \( \mu_+(S) \) be the least upper bound of the set of values \( \nu(P) \) over all packings \( P \) of \( S \). (We will see later why we don’t call this \( \mu_+(S) \)).

**Lemma 4.** Let \( S \subseteq \mathcal{R}^n \), and let \( \{S_i\} \) be a (finite or) countable disjoint collection of sets with \( S = \cup S_i \). Then

a. \( \sum \mu_+(S_i) \leq \mu_+(S) \),

b. \( \sum \mu^*(S_i) \geq \mu^*(S) \), and

c. \( \mu_+(S) \leq \mu^*(S) \).

**Proof:** For part a it suffices to show that if \( c < \mu_1 = \sum \mu_+(S_i) \) then \( c < \mu_+(S) \). If \( \mu_1 \) is finite there are packings \( P_i \) of \( S_i \) with \( \nu(P_i) \geq (c/\mu_1)\mu_+(S_i) \); if \( \mu_1 \) is infinite there are packings \( P_i \) with \( \sum \nu(P_i) \geq c \). In either case \( \cup P_i \) is a packing of \( S \), and the claim follows. For part b it suffices to show that if \( c > \mu_2 = \sum \mu^*(S_i) \) then \( c > \mu^*(S) \), whence \( \sum \mu^*(S_i) \geq \mu^*(S) \). If \( \mu_2 \) is infinite the claim follows vacuously. Otherwise there are covers \( K_i \) of \( S_i \) such that \( \nu(K_i) \leq (c/\mu_2)\mu^*(S_i) \). \( \cup K_i = \{E_j\} \) is not necessarily disjoint; but a disjoint collection \( \{D_j\} \) of elementary sets such that \( \cup E_i = \cup D_j \) and \( \sum \nu(E_i) \leq \sum \nu(D_j) \) can be constructed as follows. \( D_0 = E_0 \); at succeeding stages \( E_j \) is replaced by its difference from the union of the elementary sets defined so far. The claim has been proved when \( \mu_2 \) is finite. Finally for part c it suffices to show that if \( P \) and \( K \) are countable collections of disjoint cells and \( \cup P \subseteq \cup K \) then \( \nu(P) \leq \nu(K) \). Let \( Q \) be the collection of disjoint cells derived from \( P \) as follows. At stage \( n \), let \( F \) be the cells of \( P \) which intersect \( K_n \); partition these by the bounding hyperplanes of \( K_n \). Add to \( Q \) those parts which are contained in \( K_n \), and replace \( F \) in \( P \) by the remaining parts. \( Q \) is the union of families \( Q_n \) where each cell in \( Q_n \) is contained in \( K_n \). From facts for finite collections, \( \sum_{i \leq m} \nu(Q_{ni}) \leq \nu(K_n) \) for all \( m \), and so \( \sum \nu(Q_{ni}) \leq \nu(K_n) \) and the claim follows.

From the lemma it follows that \( \mu_+(S) = \mu^*(S) = \nu(S) \) for a cell \( S \), and in fact for \( S \) a union of a countable disjoint family of cells. Part b is referred to as countable subadditivity of the outer measure.

A collection of subsets of a set \( X \) closed under difference and countable union is called a \( \sigma \)-ring. Using a technical device attributed to Caratheodory, a \( \sigma \)-ring in \( \mathcal{R}^n \) and a measure on its sets may be defined. Define a set \( S \subseteq \mathcal{R}^n \) to be Lebesgue measurable if for any set \( T \subseteq \mathcal{R}^n \),

\[
\mu^*(T) = \mu^*(S \cap T) + \mu^*(S \cap T^c).
\]

For a Lebesgue measurable set \( S \), define its Lebesgue measure \( \mu(S) \) to be \( \mu^*(S) \). From hereon we use “measurable” for “Lebesgue measurable”.

**Theorem 5.**

a. The measurable sets are a \( \sigma \)-ring.

b. The cells are measurable.

c. If \( \{S_i : i \leq N\} \) is a countable collection of disjoint measurable sets then \( \mu(\cup S_i) = \sum \mu(S_i) \).

d. If \( S \) is measurable and \( x \in \mathcal{R}^n \) then \( S + x \) is measurable and \( \mu(S + x) = \mu(S) \).

**Proof:** Part a-c are proved in exercises 3-5. For part d, it is readily verified that \( S + x \) is measurable. Also, \( \nu(C + x) = \nu(C) \) for a cell \( C \), whence \( \mu^*(S + x) = \mu(S) \) for any \( S \).
By part b and remarks above, for a cell \( C \), \( \mu(C) = \nu(C) \). Part c is known as countable additivity of the Lebesgue measure; Part d is known as translation invariance. There are sets which are not Lebesgue measurable. Define two real numbers in \([0,1]\) to be equivalent if their difference is rational. Any set \( S \) consisting of one member of each equivalence class is unmeasurable. This is so because the interval \([0,1]\) is the union of countably many (one for each rational \( q \in [0,1] \)) sets whose measure equals \( \mu(S) \); \( \mu(S) \) can hence neither be 0 nor any nonzero value.

We leave it as exercise 6 to show that a set of outer measure 0 is measurable. Thus, if \( S_1 \subseteq S \subset S_2 \), and \( \mu(S_1) = \mu(S_2) = \mu \) is finite, then \( S \) is measurable and \( \mu(S) = \mu \). In particular if \( \mu_1(S) = \mu^*(S) \) and the value is finite then \( S \) is measurable. The outer measure of \( S \) is clearly the inf of \( \mu(T) \) for measurable supersets \( T \supseteq S \). The inner measure \( \mu_*(S) \) is defined to be the sup of \( \mu(T) \) for measurable subsets \( T \subseteq S \). This can be larger than \( \mu_1(S) \) (consider the irrationals in \([0,1]\)).

Define an open cell to be a set of the form \( \{ x : l_i < x_i < u_i \} \) for real numbers \( l_1 < u_1, \ldots, l_n < u_n \). It is easy to show that an open cell is a countable union of cells, by approaching the least corner in the product order with points in \( \mathbb{Q}^n \). Thus, an open cell is measurable; in fact its measure is clearly \( (u_1 - l_1) \cdots (u_n - l_n) \).

Any open set is a countable union of open cells, so any open set is measurable. It follows that any closed set is measurable. Define a closed cell to be a set of the form \( \{ x : l_i \leq x_i \leq u_i \} \); again clearly its measure is \( (u_1 - l_1) \cdots (u_n - l_n) \). By the preceding paragraph, any cell \( \{ x : l_i R_i x_i S_i u_i \} \) where \( R_i \) and \( S_i \) are \( \leq \) or \( < \), has measure \( (u_1 - l_1) \cdots (u_n - l_n) \).

By an “elementary matrix” will be meant the matrix of an elementary operation, as specified in chapter 10. By results of chapter 10, any square matrix of nonzero determinant can be written as a product of permutation matrices, elementary matrices, and a diagonal matrix.

**Theorem 6.** Suppose \( T \) is a linear transformation of determinant \( \lambda \neq 0 \), and \( S \subseteq \mathbb{R}^n \) is Lebesgue measurable. Then \( T[S] \) is Lebesgue measurable, and \( \mu(T[S]) = |\lambda| \mu(S) \).

**Proof:** The theorem is clear for \( T \) a diagonal or permutation matrix and \( S \) a cell, whence for any measurable \( S \). It thus suffices to show the theorem for a matrix of an elementary operation. For such a matrix in \( \mathbb{R}^2 \), a cell \( C \) may be dissected into rectangles and triangles whose union is \( T[C] \). In \( \mathbb{R}^n \), the dissection may be carried out in a two dimensional subspace, and extended to \( C \) by taking the product with the orthogonal subspace. It is readily seen that such a parallelootope is measurable, etc.

For the next theorem, we make some preliminary observations, of interest in themselves. If a bounded convex set lies in an affine subspace then it has measure 0, since it is contained in rotated cells of arbitrarily small measure. If \( S \subseteq \mathbb{R}^n \) does not lie in an affine subspace, \( x \in S \), and \( \epsilon > 0 \) then \( B_{2\epsilon} \) contains \( y_i, 1 \leq i \leq n \), such that \( \{ y_i - x \} \) is linearly independent, and so \( B_{2\epsilon} \) contains points in \( S^{\text{int}} \).

Define an open orthant in \( \mathbb{R}^n \) to be \( \{ x : x_i R \text{ for all } i \} \), where \( R \) is \( < \) or \( \geq \).

**Lemma 7.** Let \( S \) be a set of points in \( \mathbb{R}^n \), one from each open orthant. Then 0 is in the convex hull of \( S \).

**Proof:** This follows by induction on \( n \); the basis \( n = 1 \) is straightforward. For \( n > 1 \) consider the negative halfspace \( H_\prec \), projected onto the hyperplane \( H_0 \) where \( x_1 = 0 \). Inductively, 0 is in the convex hull of the projection, whence there is a point in the convex hull of \( S \cap H_\prec \) on the negative \( x_1 \) axis. Similarly there is a point in the convex hull of \( S \cap H_\succ \) on the positive \( x_1 \) axis, and the lemma follows.

**Theorem 8.** A bounded convex set in \( S \subseteq \mathbb{R}^n \) is measurable.

**Proof:** (\cite{Szabo}) We suppose w.l.o.g. that \( S \) does not lie in an affine subspace, and \( S \subseteq C \) where \( C \) is the closed cube \( \{ x : -1/2 \leq x_i \leq 1/2 \} \). Let \( P_j \) be the set of \( 2^n \) closed cubes arising from subdividing each edge of \( C \) into \( 2^j \) equal parts. Let \( P^I_j \) be the cubes of \( P_j \) which contain an interior point of \( K \), and let \( P^B_j \) be the cubes of \( P^I_j \) which contain a boundary point of \( K \). Then \( \cup(P^I_j - P^B_j) \subseteq S \subseteq \cup P_j \). Fixing \( j \), for each corner
$v$ of $C$ let $Q_v$ be the cubes of $P_j$ in a facet incident to $v$. Let $R_v$ be the rays from the centers of the cubes in $Q_v$, parallel to the ray from $v$ through 0. Each center of a cube in $P_j$ lies in one ray from each $R_v$, and $|R_v| \leq n2^{2n-1}$. If $c \in P_i^B$ then one of the rays of $\cup_v R_v$ intersecting $c$ intersects it “first”. If not, let $c_v$ be the cube that it does intersect first. Choose $w_v \in c_v \cap S^{\text{int}}$. Using lemma 7, $c \not\subseteq S^{\text{int}}$, a contradiction. It follows that $\mu(\cup P_i^B) \leq n2^{n-2}$, proving the theorem.

Define a set $S$ to be Jordan measurable if the sup of $\nu(E)$ for elementary subsets $E$ equals the inf of $\nu(F)$ for elementary supersets $F$. The proof of the theorem can be modified to show that in fact a convex set is Jordan measurable.

3. Minkowski’s theorem.

**Theorem 9.** Suppose $S \subseteq \mathbb{R}^n$ is bounded and convex, $S = -S$, and $\mu(S) > 2^n$; then $S$ contains a nonzero point of $\mathbb{Z}^n$.

**Proof:** Suppose $S$ contains no such point. For $z \in \mathbb{Z}^n$ let $T_z = z + (1/2)S$; we claim that the $T_z$ are disjoint. Indeed, if $w \in T_z \cap T_{z'}$ then $w = y + y'/2 = z + z'/2$ where $y', z' \in S$, whence $y - z = (y' - z')/2$ is in $S$, whence $y = z$. Let $N$ be an integer, and let $U_N = \cup_z \{T_z : 0 \leq z_i < N \text{ for all } i\}$. Then $\mu(U_N) = (N/2)^n \nu(S)$. On the other hand there is a constant $b$ such that $U_N$ is contained in a box of side $N + b$, whence $\mu(S) \leq (1+b/N)^n 2^n$. Thus, $\mu(S) \leq 2^n$, proving the contrapositive.

**Corollary 10.** Suppose $S \subseteq \mathbb{R}^n$ is bounded and convex, $S = -S$, $L$ is a full rank lattice, and $\mu(S) > 2^n \det(L)$; then $S$ contains a nonzero point of $L$.

**Proof:** Let $S = G[T]$ where $G$ is a generator matrix for $L$. Then $T$ satisfies the hypotheses of the theorem, and the image of a nonzero point of $\mathbb{Z}^n$ is a nonzero point of $L$.

**Corollary 11.** Let $G$ be an $n \times n$ real matrix with entries $g_{ij}$ and $\det(G) \neq 0$, and let $\zeta_i(x)$ be the linear form $\sum_j g_{ij}x_j$. Suppose $\lambda_i$ for $1 \leq i \leq n$ are positive real numbers with $\lambda_1 \cdots \lambda_n > |\det(G)|$. Then there is a nonzero $x \in \mathbb{Z}^n$ such that $|\zeta_i(x)| < \lambda_i$ for $1 \leq i \leq n$.

**Proof:** Let $S = \{x : |x_i| < \lambda_i \text{ for } 1 \leq i \leq n\}$ and apply corollary 10.

Theorem 9 is called Minkowski’s fundamental theorem or Minkowski’s first theorem. A set such that $S = -S$ is called symmetric, or said to have center 0. Corollary 11 is called Minkowski’s theorem on linear forms.

There is a “limiting” version of theorem 9 and its corollaries. If $S \subseteq \mathbb{R}^n$ is bounded and convex and $S = -S$ then these hypotheses also hold for $(1 + 2^{-i})S$ for any $i > 0$. Since $S^{\text{cl}}$ contains only finitely many points of $\mathbb{Z}^n$, it follows that if $\mu(S) \geq 2^n$ then there is a nonzero point of $\mathbb{Z}^n$ in $S^{\text{cl}}$. Similarly if $\mu(S) \geq 2^n |\lambda|$ then $S^{\text{cl}}$ contains a nonzero point of $L$, etc.

**Theorem 12.** Let $r_i$ for $1 \leq i \leq n$ be real numbers, and $\epsilon$ with $1 > \epsilon > 0$, be real numbers. Then there are integers $p_i$, $1 \leq i \leq n$, and $q > 0$, such that $|r_i - p_i/q| \leq \epsilon/q$ and $q < (1/\epsilon)^n$.

**Proof:** Suppose $1 > \epsilon' > \epsilon > 0$, and consider the linear forms $p_i - r_i q$ and $(1/\epsilon)q$. By Corollary 11 (with $\lambda_i = \epsilon'$ for $1 \leq i \leq n$ and $\lambda_{n+1} = 1$), there are integers $p_i$ for $1 \leq i \leq n$, and $q$, not all 0, such that $|p_i - r_i q| < \epsilon'$ and $|q| < (1/\epsilon)^n$. Negating the $p_i$ and $q$ if necessary ensures $q \geq 0$, and since $\epsilon' < 1$ clearly $q \neq 0$. Since the $p_i$ and $q$ exist for any $\epsilon' > \epsilon$, and there are only finitely many possibilities for them, the theorem is proved.

It is now easy to prove the converse of theorem 1.

**Theorem 13.** If $L \subseteq \mathbb{R}^n$ is a discrete additive subgroup then $L$ is a lattice.
Proof: By applying a transformation and ignoring 0 rows we may suppose that \( L \) is full rank. Let \( g_1, \ldots, g_n \) be linearly independent elements of \( L \), and let \( L_1 \) be the lattice generated by them. Let \( \epsilon > 0 \) be such that if \( x \in L \) and \( |x| < \epsilon \) then \( x = 0 \), and choose \( \epsilon_1 > 0 \) with \( \epsilon_1 < \epsilon/\sum_i |g_i| \). Suppose \( x \in L \), and suppose \( x = \sum_i r_i g_i \) for reals \( r_i \). By theorem 12 choose \( p_i \) and \( q \) with \( |r_i - p_i/q| \leq \epsilon_1 \) and \( q < (1/\epsilon_1)^n \). Then

\[
|x'| = \left| \sum_i (qr_i - p_i)g_i \right| \leq \sum_i |qr_i - p_i||g_i| \leq \epsilon_1 \sum_i |g_i| < \epsilon.
\]

Thus, \( x' = 0 \), whence \( qx \in L_1 \). Since \( q \) is bounded, \( L \subseteq (1/m)L_1 \) for some integer \( m \), and it follows that \( L \) is a lattice.

Approximation of reals by rationals is called Diophantine approximation. Theorem 12 is an example of a theorem on simultaneous Diophantine approximation. It follows from the theorem that there are \( p_i \) and \( q \) with \( |r_i - p_i/q| \leq 1/q^{1+1/n} \). This may be proved using “Dirichlet’s pigeonhole principle” rather than Minkowski’s theorem (exercise 7). It may be shown that there are \( p_i \) and \( q \) with \( |r_i - p_i/q| \leq (n/(n+1))/q^{1+1/n} \) ([Hua], theorem 20.6.1).

Let \( S_n \) be the closed ball of radius 1 in \( R^n \). We appeal to basic facts in the theory of integration to conclude the following.

- \( \mu(rS_n) = r^n \mu(S_n) \) (\( rS_n \) is the closed ball of radius \( r \)).
- \( \mu(S_1) = 2 \), and \( \mu(S_{n+1}) = \mu(S_n) \int_{-1}^1(1-x^2)^{n/2}dx \).

Letting \( \mu_n \) denote \( \mu(S_n) \), it follows (exercise 8) that

\[
\mu_n = \begin{cases} 
\frac{\pi^{n/2}}{(n/2)!} & n \text{ even} \\
\frac{2^n \pi^{(n-1)/2}((n-1)/2)!}{n!} & n \text{ odd}
\end{cases}
\]

Taking \( S = S_n \) in corollary 10 yields another important application of Minkowski’s fundamental theorem. Let \( \lambda_1 \) denote the length of the shortest nonzero vector in \( L \). Then \( \lambda_1 \leq r \) where \( r^n \mu_n = 2^n \det(L) \); thus \( \lambda_1 \leq (2/\mu_n)^{1/n} \det(L)^{1/n} \).

The notation \( \gamma_n \) is used for the so-called Hermite constant, the infimum of the values \( \gamma \) such that \( \lambda_1^2 \leq \gamma \det(L)^{2/n} \) for all \( L \). The inequality \( \lambda_1 \leq (4/3)^{(n-1)/4} \det(L)^{1/n} \) (q.v. see [Newman]) yields a better bound for small \( n \), indeed the exact value for \( n = 2 \). The exact value in known for \( 1 \leq n \leq 8 \); see [Lagarias] for a list.

The Hermite constant gives the density for the best “lattice packing”. The density is the ratio of the volume of a ball of radius \( \lambda_1/2 \) to \( \det(L) \); writing \( \Delta \) for this, \( \Delta = (\lambda_1^n/\det(L))(\mu_n/2^n) \) where \( \mu_n \) is as above. Thus, \( \Delta \leq (\gamma_n^{n/2})(\mu_n/2^n) \).

4. Hermite and Smith normal form. Suppose in this section that \( R \) is a principal ideal domain. Recall from section 6.1 that a gcd of \( x_1, \ldots, x_n \) exists and can be written as \( a_1x_1 + \cdots + a_n x_n \). As noted above, a square matrix over \( R \) is called unimodular if its determinant is a unit.

Lemma 14. Suppose \( x \) is a row vector over \( R \). Then there is a unimodular matrix \( U \) such that \( xU = [w, 0, \ldots, 0] \) where \( w = \gcd(x_1, \ldots, x_n) \).

Proof: For \( n = 1 \) the lemma is trivial. For \( n = 2 \) let \( w = a_1x_1 + a_2 x_2 \); let

\[
U = \begin{bmatrix} a_1 & -x_2/w \\ a_2 & x_1/w \end{bmatrix}.
\]
For $n > 2$, choose a $2 \times 2$ unimodular matrix $U'$ such that

$$\begin{bmatrix} x_1, \ldots, x_{n-1}, x_n \end{bmatrix} \begin{bmatrix} I & Z \\ Z' & U' \end{bmatrix} = [x_1, \ldots, d, 0],$$

where $I$ is an $(n-2) \times (n-2)$ identity matrix, $Z$ is an $(n-2) \times 2$ zero matrix, and $d = \gcd(x_{n-1}, x_n)$. Proceeding inductively, the lemma follows.

When $R$ is a Euclidean domain, the matrix $U$ of the lemma may be found by the extended Euclidean algorithm, and expressed as a product of elementary matrices and a permutation matrix (see exercise 6.10; in fact, the permutation may be a transposition).

**Theorem 15.** Let $M$ be $n \times n$ matrix of rank $n$ over $R$. Then $MU = H$ where $H$ is lower triangular and $U$ is unimodular. Further, $a_{ii}$ is determined up to a unit, and for $j < i$ the coset of $h_{ij}$ in $h_{ii}R$ is determined.

**Proof:** By the lemma, the first row can be put in the form $[w, 0, \ldots, 0]$; proceeding inductively, the existence of $M$ and $U$ follows. Choose an element from each class of associates. and replace each $h_{ii}$ by its chosen associate. This can clearly be accomplished by right multiplication by a unimodular matrix. Choose a system of coset representatives for each ideal of $R$, and replace each $h_{ij}$ by its representative for $h_{ii}R$. This may be accomplished by successive right multiplication by elementary matrices, as is readily verified. If $H$ and $H'$ are two resulting matrices then $HV = H'$ where $V$ is unimodular; we claim that $H = H'$ and $V = I$.

Write $HV = H'$ as

$$\begin{bmatrix} a & 0 \\ c & d \end{bmatrix} \begin{bmatrix} u & v \\ w & x \end{bmatrix} = \begin{bmatrix} a' & 0 \\ c' & d' \end{bmatrix},$$

where the top row (left column) has height (width) 1. Since $a \neq 0$ and $av = 0$, $v = 0$; whence $u$ is a unit and $x$ is unimodular. Since $u$ must be 1, $a = a'$. Further, $dx = d'$ and $x$ is unimodular, so $d = d'$ and $x = I$ inductively. Thus, $c' = c + dw$, so $c'_1 = c_1 + d_1 w_1$, so $c'_1 = c_1$ and $w_1 = 0$. Proceeding inductively, $c' = c$ and $w = 0$.

The matrix $H$ of the theorem is called a Hermite normal form for $M$. If choices are made as in the proof of the theorem, it is uniquely specified, and $U$ as well.

If $M$ is $m \times n$ where $m > n$, and has rank $n$, then the theorem may be applied to the first $n$ rows $M'$ of $M$, yielding a unique (given a choice of associates and coset representatives) top $n$ rows $H'$, and a unique $n \times n U$. Right multiplying $M$ by $U$ gives a matrix $H$, which is unique.

If $M$ is $m \times n$ where $m < n$, and has rank $m$, then the proof of the theorem may be modified. First, $U$ and $H$ are obtained as before; $H$, being lower triangular, is 0 in the last $n - m$ columns. The uniqueness proof may be modified, claiming only that $U$ is the identity in the upper left $m \times m$ submatrix, and 0 in the upper right submatrix. Thus, $H$ is still unique, but $U$ may not be.

Define the $k$th determinental divisor $\delta_k(M)$ of a matrix $M$ to be the gcd of the entries in its $k$th compound $M^{(k)}$ (see section 10.13). Also, the rank of $M$ is the largest $k$ such that $\delta_k(M) \neq 0$ (see appendix 2). By the Binet-Cauchy theorem (theorem 10.21.c), if $L = MN$ then $\delta_k(L)|\delta_k(M)$ and $\delta_k(L)|\delta_k(N)$. It follows that if $VMU = N$ where $U$ and $V$ are unimodular then $M$ and $N$ have the same determinental divisors, up to units of $R$.

**Theorem 16.** Let $M$ be $n \times n$ matrix $R$ of rank $r$. There are unimodular matrices $V$ and $U$ such that $VMU = S$, where $S$ is a diagonal matrix with $s_{ii} \neq 0$ iff $i \leq r$, and for $i < r$ $s_{ii}|s_{i+1,i+1}$. The $s_{ii}$ are determined, up to units.

**Proof:** If $M$ is all 0 there is nothing to prove. Otherwise, $m_{11}$ may be made nonzero by row and column permutations. By lemma 14, the top row may be made 0, except for $m_{11}$. This may be assumed to strictly decrease the number of factors in the prime factorization of $m_{11}$, by first setting multiples of $m_{11}$ in the first
row to 0. Then, by the column version of lemma 14, the left column may be made nonzero, except for \(m_{11}\). If \(m_{11}\) does not divide some \(m_{ij}\) with \(i, j > 1\) then add row \(i\) to row 0 and continue. Once \(m_{11}\) does divide all \(m_{ij}\) with \(i, j > 1\), carry out the procedure recursively on the minor with row and column 1 deleted. Clearly the remaining matrix will be all 0 after \(r\) stages, else the rank would be greater than \(r\). Also, \(\delta_k(S) = s_{11} \cdots s_{kk}\), so \(s_{11} \cdots s_{kk} = \delta_k(M)\), from which the last claim follows.

The matrix \(S\) is called the Smith normal form of \(M\). Writing \(s_i\) for \(s_{ii}\), these values are called the invariant factors. Clearly \(VMU = N\) for some unimodular \(U, V\) iff \(M\) and \(N\) have the same invariant factors, if they have the same determinantal divisors.

For the remainder of the section \(R\) will be taken as \(Z\), a case is of special interest, in particular to point lattices. The associate chosen from a class may be taken as the positive integer. The coset representatives for a positive integer \(k\) may be taken as \(0, \ldots, k - 1\).

Suppose \(L\) is a sublattice of \(Z^n\), with generator matrix \(G\). Any element of \(Z^n\) can be translated by an element of \(L\), so that the translate lies in the fundamental region. Further, two elements in the fundamental region cannot differ by a vector of \(L\). Thus, the elements of \(Z^n\) which lie in the fundamental region comprise a system of coset representatives of \(L\).

\(G\) may be written as \(HU\) where \(H\) is the Hermite normal form of \(G\). The matrix \(H\) is also a generator matrix for \(G\). From this it is easy to see that if \(L\) is full rank, the number of elements of \(L\) lying in the fundamental region equals \(\det(L)\). Indeed, let \(L_1\) be the the lattice by columns 2 to \(n\) of \(H\). Let \(h\) be the 1,1 entry, and \(d_1\) the product of the remaining diagonal entries. Clearly \(\det(L_1) = d_1\) and \(\det(L) = hd_1\). Inductively, there are \(d_1\) entries in the set \(S\) of coset representatives of \(L_1\) lying in the fundamental region. For \(0 \leq i < h\) let \(p_i\) be the point with \(x_1 = i\) lying in the fundamental region, whose remaining coordinates are least in the product order on \(Z^{n-1}\). The translate by \(p_i\) of \(S\) comprises the coset representatives of \(L\) in the fundamental region with \(x_1 = i\).

If \(L_2 \subseteq L_1 \subseteq Z^n\) are full rank lattices, then the index of \(L_2\) in \(L_1\) equals \(\det(L_2)/\det(L_1)\). This follows from the foregoing and the fact that \(Z^n/L_2\) is isomorphic to \((Z^n/L_1)/(L_2/L_1)\).

\(Z^n\) may be replaced by any lattice \(L_0\) in the foregoing, by using the linear isomorphism between \(L_0\) and \(Z^n\), provided \(\det(L_0)\) is taken in to account. In particular, if \(L_1 \subseteq L_0\) where both are full rank than the index of \(L_1\) in \(L_0\) equals \(\det(L_1)/\det(L_0)\).

A linearly independent generating set for a lattice is also called a basis. A linearly independent subset \(\{h_1, \ldots, h_i\}\) of a lattice \(L\) is said to be extensible if it can be extended to a basis for \(L\). If \(\{h_1, \ldots, h_i\}\) is extensible then

\[
(*) \text{ for any } x \in L \text{ with } x \in \sum_{j=1}^i h_j R, \; x \in \sum_{j=1}^i h_j Z.
\]

Indeed, if \(\{h_1, \ldots, h_n\}\) is a basis and \(x \in L\) then \(x\) has a unique expression as an integer linear combination of all the \(h_j\), and as a real linear combination of the first \(i\) of them.

The Hermite normal form can be used to prove that conversely, property (\(*\)) implies that \(\{h_1, \ldots, h_i\}\) is extensible. Let \(H\) be the \(n \times i\) matrix whose columns are the \(h_j\), and let \(G\) be a generator matrix; then \(H = GH\) for some \(n \times i\) integer matrix \(H\). Further, \(H = UH'\) where \(U\) is unimodular and \(H'\) is in (transposed) Hermite normal form. Let \(G' = GU^{-1}\), so that \(H = G'H'\). Letting \(g'_{ij}\) be the \(i\)th column of \(G'\), for \(i \leq j\) it follows that \(g'_{ij} = \sum_{j=1}^i h_j R\); indeed the coefficient of \(h_i\) is 1. Thus, \(H'_{ii} = 1\), whence \(H'_{ji} = 0\) for \(j < i\), and so \(g'_{ij} = h_j\) for \(j \leq i\).

As another consequence of the existence of the Hermite normal form, there are only finitely many integer lattices of a given determinant.

It was shown in the late 1970’s that the Hermite and Smith normal forms of an integer matrix are computable in polynomial time. See [MiccWar] for a recent reference for Hermite normal form. Other
discussions can be found in [Schr] and [Cohen].

5. Reduced bases. The Hermite normal form yields a useful basis for an integer lattice. Other useful bases include bases whose vectors are short and approximately orthogonal. It is easy to see from two dimensional examples that an orthogonal basis cannot in general be found. However, an orthogonalized version of a basis is a useful tool for measuring how close to orthogonal the basis is.

Such can be readily obtained by skipping the normalization in the the Gram-Schmidt orthonormalization procedure of section 10.8. Given linearly independent column vectors \( g_1, \ldots, g_r \) in \( \mathbb{R}^n \), define \( \hat{g}_i \) and \( \mu_{ij} \) as follows.

\[
\hat{g}_1 = g_1; \quad \text{for } 1 \leq j < i, \quad \mu_{ij} = \frac{g_i \cdot \hat{g}_j}{\|g_j\|^2}; \quad \text{for } 1 < i \leq r, \quad \hat{g}_i = g_i - \sum_{j<i} \mu_{ij} \hat{g}_j.
\]

In particular, \( g_i \) is made normal to the subspace spanned by the preceding vectors by subtracting the projection onto the subspace. We will use the notation \( \hat{g}_i \) and \( \mu_{ij} \) in this section for these quantities.

Note that \( g_i = \hat{g}_i + \sum_{j<i} \mu_{ij} \hat{g}_j \), whence \( G = GU \) where \( G \) is the matrix with columns \( g_i \) (\( \hat{g}_i \)), and \( U \) is an \( r \times r \) upper triangular matrix with 1’s on the diagonal, and \( j, i \) entry \( \mu_{ij} \) for \( j < i \). Also, \( |\hat{g}_i| \leq |g_i| \).

From hereon in this section let \( L \subseteq \mathbb{R}^n \) be a full rank lattice with basis \( g_1, \ldots, g_n \). Then \( \det(G) = \det((G^\dagger G)^{1/2}) = \Pi_i |\hat{g}_i| \), whence \( \det(G) \leq \Pi_i |g_i| \) (Hadamard’s inequality). Informally, the basis is said to be reduced if \( \Pi_i |g_i| / \det(G) \) is small. Various specific notions have been considered.

An ordered lattice basis \( \langle g_1, \ldots, g_n \rangle \) is said to be Minkowski reduced if \( g_1 \) is a shortest vector, and inductively \( g_i \) is a shortest vector such that \( \langle g_1, \ldots, g_i \rangle \) is linearly independent and extensible. This notion has been of interest in the theory of quadratic forms.

If instead, \( g_1 \) is required to be a vector whose projection onto Span\( \langle g_1, \ldots, g_{i-1} \rangle \) is of least length among \( g_i \) such that \( \langle g_1, \ldots, g_i \rangle \) is linearly independent and extensible, the basis is said to be Korkin-Zolotarev (abbreviated KZ) reduced. The requirement that \( |\mu_{ij}| \leq 1/2 \) for \( 1 \leq j < i \leq n \) is often added.

This type of reduced basis has found numerous applications in the “algorithmic geometry of numbers”; see the literature for further discussion, for example [Kannan].

In 1982 a type of reduced basis was discovered which has been of considerable interest since. Say that a basis \( g_1, \ldots, g_n \) is LLL-reduced if

1. \( |\mu_{ij}| \leq 1/2 \) for \( 1 \leq j < i \leq n \), and
2. \( |\hat{g}_i + \mu_{i,i-1} \hat{g}_{i-1}|^2 \geq (3/4)|\hat{g}_{i-1}|^2 \) for \( 1 < i \leq n \).

As will be seen, in achieving requirement 2 we allow the \( g_i \) to be reordered.

**Theorem 17.** A rank \( n \) lattice in \( \mathbb{R}^n \) has an LLL-reduced basis.

**Proof:** Let \( G_k \) (\( \hat{G}_k \)) denotes the first \( k \) columns of \( G \) (\( \hat{G} \)); then \( \hat{G}_kU_k = G_k \) where \( U_k \) is upper triangular and has diagonal 1. Supposing \( g_1, \ldots, g_{k-1} \) satisfy the requirements, we may satisfy requirement 1 for \( \mu_{kl} \) where \( l < k \) by “reducing \( \mu_{kl} \)” as follows. Let \( q \) be the integer nearest to \( \mu_{kl} \) then replace \( g_k \) by \( g_k - qg_l \). Both sides of \( \hat{G}_kU_k = G_k \) are right multiplied by an elementary matrix which only changes \( \mu_{kj} \) for \( j \leq l \).

Consider the following procedure.

1. Set \( k = 2 \).
2. Reduce \( \mu_{k,k-1} \).
3. If \( |\hat{g}_k|^2 < (3/4 - \mu_{k,k-1}^2)|\hat{g}_{k-1}|^2 \), swap \( g_k \) and \( g_{k-1} \), set \( k = \max(2, k-1) \), and go to 2.
4. Reduce \( \mu_{k,l} \) for \( l = k - 2, \ldots, 1 \) successively.
5. If \( k = n \) terminate the procedure; otherwise set \( k = k + 1 \) and go to 2.

Clearly it suffices to show that only finitely many swaps are performed. Let \( D_l = \det(G_l^\dagger G_l) \); in a swap only \( D_{k-1} \) changes. Also, \( D_l = \det(G_l^\dagger \hat{G}_l) \), whence a reduction leaves \( D_l \) unchanged. A swap leaves \( \hat{g}_j \) unchanged for \( j < k - 1 \), and replaces \( \hat{g}_{k-1} \) by \( \hat{g}_{k-1}' = g_{k-1} - \sum_{j<k-1} \mu_{k-1,j} \hat{g}_j \). Thus, it replaces \( \hat{g}_{k-1} \) by \( D_{k-1}' < (3/4)D_{k-1} \). The theorem follows because \( D_k \geq c\lambda_1 \) where \( \lambda_1 \) is the minimum length of a vector.
in the lattice and $c$ is a constant depending on $n$. This is noted for the full rank case in section 3, and the claim follows readily.

The argument shows that for a lattice in $\mathbb{Z}^n$, only polynomially many swaps are performed, whence the procedure runs in polynomial time. Note that only the values $g_i$, $\mu_{ij}$ and $|\hat{g}_i|^2$ need be maintained in an implementation. The original paper [LLL] shows how to do this incrementally, and gives a more detailed analysis of the running time.

**Theorem 18.** Suppose \{${g_1, \ldots, g_n}$\} is an LLL-reduced basis for a lattice $L \subseteq \mathbb{R}^n$.

a. $|\hat{g}_j|^2 \leq 2^{i-j} |\hat{g}_i|^2$ for $1 \leq j \leq i \leq n$.

b. $|g_j|^2 \leq 2^{i-1} |\hat{g}_i|^2$ for $1 \leq j \leq i \leq n$.

c. $|\Pi |g_i| \leq 2^{n(n-1)/4} \det(L)$.

d. $|g_1| \leq 2^{(n-1)/4} \det(L)^{1/n}$.

**Proof:** For $1 < i \leq n$ clearly $|\hat{g}_i|^2 \geq (3/4 - \mu_{i,i-1}) |\hat{g}_{i-1}|^2 \geq (1/2) |\hat{g}_{i-1}|^2$. Part a follows by induction. For part b,

$$|g_i|^2 = |\hat{g}_i|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 |\hat{g}_j|^2 \leq |\hat{g}_i|^2 + \sum_{j=1}^{i-1} (2^{i-j}/4) |\hat{g}_j|^2 \leq 2^{i-1} |\hat{g}_i|^2.$$ 

Part b follows from this and part a. For part c, $x = \sum_i r_i g_i = \sum_i \hat{r}_i \hat{g}_i$ where $r_i \in \mathbb{Z}$. If $i$ is largest so that $r_i \neq 0$ then $\hat{r}_i = r_i$ (consider the factorization $G = GU$). Thus, $|x|^2 \geq |\hat{g}_i|^2$, the claim follows by part b. Part d follows from part c and the fact that $\det(L) = \Pi \hat{g}_i$. For part e, multiply the inequalities of part b for $j = 1$ and $i$ from 1 to $n$, to obtain $|g_1|^{2n} \leq 2^{(n-1)/2} \det(L)^n$.

LLL-reduced bases have been used to obtain various polynomial time algorithms; we give a few examples. By theorem 18.c a vector in an integer lattice which is within a factor of 2 of the shortest vector may be obtained in polynomial time. The factor has been slightly improved. On the other hand, obtaining a vector within a constant factor has been shown to be “hard”, in a sense less stringent than NP-hard. The first result in this subject was obtained in 1998 by M. Ajtai, who showed that obtaining the shortest vector was hard in the less stringent sense (it is still open whether this problem is NP hard). See for example [Micc] for further discussion.

There is a polynomial time algorithm for finding the Hermite normal form which uses an LLL-reduced basis; this may be found in [Schrijver].

Given rational numbers $r_i$ for $1 \leq i \leq n$, and $\epsilon$ with $1 > \epsilon > 0$, integers $p_i$ for $1 \leq i \leq n$, and $q > 0$, such that $|r_i - p_i/q| \leq \epsilon/q$ and $q \leq 2^{O(n)/4}(1/\epsilon)^n$, may be found in polynomial time (cf. theorem 12). This was observed in the original paper [LLL], and may also be found in [Schrijver].

One form of the integer linear programming problem (ILP) asks whether a system of inequalities $Ax \leq b$, where $A$ and $b$ are rational, has an integer solution. This problem is NP-complete. LLL-reduced bases were used in 1983 to show that if the dimension of $x$ is fixed, the problem can be solved in polynomial time; see [Schrijver].

Recall from theorem 7.4 that a primitive polynomial in $\mathbb{Z}[x]$ may be written as a product of factors in $\mathbb{Z}[x]$, which are irreducible in $\mathbb{Q}[x]$. Such a factorization can be obtained in polynomial time, by an algorithm which uses LLL basis reduction, as demonstrated in [LLL]; a description of this algorithm is given in section 28.6. In [Cohen] it is stated that an older method, which can be found there or in [Knuth], performs better in practice.

**6. Root systems.** Root systems arise in the theory of reflection groups, and are consequently relevant in various topics in algebra, such as linear algebraic groups and Lie algebras. They are useful also in the
theory of point lattices, in particular giving rise to some basic families of such. There are various “flavors” of root systems, depending on the exact application.

We begin with a discussion of reflection groups, following [Vara]; see also [Howlett]. If \( v \in \mathbb{R}^n \), the projection onto the hyperplane orthogonal to \( v \) is the map \( x \mapsto x - \frac{2x \cdot v}{v \cdot v} v \); the reflection in the hyperplane is the map \( x \mapsto x - 2 \frac{x \cdot v}{v \cdot v} v \). Letting \( r = v/|v| \), one sees that the matrix for the reflection is \( I - 2rt^t \). This is an element of the orthogonal group \( O(n) \) defined in section 20.14 (in fact, it is symmetric and idempotent). It maps \( r \) to \(-r\), and any vector \( w \) with \( r \cdot w = 0 \) to itself. From this, it follows that the determinant is \(-1\).

For a unit vector \( r \), let \( \rho_r \) denote the reflection just defined. It is readily verified to be the unique linear map which fixes points in \( r^\perp \) and maps \( r \) to \(-r \) (consider \( f_{\rho_r} \)). From this, if \( r \) and \( s \) are unit vectors, and \( g \in O(n) \) is a transformation with \( g(r) = s \), then \( \rho_s = g \rho_r g^{-1} \).

A reflection in \( \mathbb{R}^n \) is defined to be a subgroup of \( O(n) \), which is generated by the reflections it contains. We define a unit root system in \( \mathbb{R}^n \) to be a finite set \( R \) of unit vectors, such that \( \rho_r[R] = R \) for all \( r \in R \). Given such an \( R \), let \( G_R \) be the reflection group generated by \( \{ \rho_r : r \in R \} \). If \( g \in G_R \) fixes \( R \) pointwise it fixes both \( \text{Span}(R) \) and \( \text{Span}(R)^\perp \); thus, \( G_R \) acts faithfully on \( R \), so is a finite reflection group.

One readily verifies that the product of two reflections is a rotation in the plane of the two normals. Indeed, choosing the signs of the normals \( r \) and \( s \) so that the angle between them is \( \theta \leq \pi/2 \), then the angle of the rotation is \( 2\theta \). In a finite reflection group, \( \theta = \pi/m \) for some integer \( m \geq 2 \).

The group with generators \( \rho \) and \( \sigma \), and relations \( \rho^2 = 1 \), \( \sigma^m = 1 \), and \( \rho \sigma = \sigma^{-1} \rho \) is the dihedral group (see section 4.3) for \( m \geq 3 \), and the four group \( \mathbb{Z}_2 \times \mathbb{Z}_2 \) when \( m = 2 \). Indeed, there are \( 2m \) elements \( \rho^i \sigma^j \) where \( i, j \in \mathbb{Z}_m \), and \( \rho^i \sigma^j \rho^{i+2} \sigma^{j+2} = \rho^{i+1} \sigma^{j+1} \rho^{i-1} \sigma^j \), where the minus sign is taken if \( i_2 = 1 \). If the vertices of the \( m \)-gon for \( m \geq 3 \) are numbered \( 1, \ldots, m \) counterclockwise, \( \rho \) may be taken as \( (2, n) \ldots (i, j) \), where \( i = [n/2] \) and \( j = i + 1 \) if \( n \) is odd and \( j = i + 2 \) if \( n \) is even; and \( \sigma \) as \( (1, \ldots, n) \).

The subgroup of a finite reflection group generated by two reflections \( \rho_r \) and \( \rho_s \) is of the above form, with \( \sigma = \rho_s \rho_r \). It is readily verified that when \( m \) is even it acts transitively on the orbits of \( r \) and \( s \); and when \( m \) is odd it acts transitively on a single orbit. In either case, it acts on \( 2m \) unit vectors lying in the plane and spaced at angles of \( \pi/m \).

If \( G \) is a finite reflection group let \( R \) be the set of unit vectors such that \( \rho_r \in G \). Given \( r, s \in R \), it follows from the above description of the subgroup generated by \( \rho_r \) and \( \rho_s \) that \( \rho_r \rho_s \in R \). Thus, \( R \) is a unit root system, called the root system of the group. By hypothesis \( G_R = G \).

An order on a real vector space is one on the additive group (see section 6.5), where in addition if \( v \) is positive and \( \alpha > 0 \) then \( \alpha v \) is positive. Call a linear combination positive (negative) if it is nontrivial and its coefficients are all positive (negative). A positive linear combination of positive elements is clearly positive.

Suppose \( R \) is a unit root system. A subset \( P \subseteq R \) is called a positive system if there is an order on \( \mathbb{R}^n \) such that \( P \) lies in the positive subset. Suppose \( P \) and \( Q \) are positive systems with \( P \subseteq Q \). If \( r \notin P \) then \(-r \in P \), so \(-r \in Q \), so \( r \notin Q \); thus, \( Q = P \).

In exercise 9 it is shown that a positive set may alternatively be defined as those root vectors lying strictly on one side of some hyperplane. In particular \( \text{Cone}(P) \) is a pointed polyhedral cone.

A subset \( S \subseteq R \) is called a simple (or fundamental) system if it is linearly independent, and every \( r \in R \) is either a positive linear combination of elements of \( S \), or a negative one. Given a simple system \( S \), let \( P \) be those \( r \in R \) which are a positive linear combination. Then \( P \) is a positive system containing \( S \) (extend \( S \) to a basis and consider the order as in exercise 9). If \( Q \) is any other positive system containing \( S \) then \( P \subseteq Q \), whence \( Q = P \); that is, \( S \) is contained in a unique positive system.

For the following, given a a positive system \( P \), let \( \text{Plc}(S) \) denote those \( r \in P \) which are a positive linear combination of elements of \( S \).

**Lemma 19.** Suppose \( P \) is a positive system, and \( S \) is a minimal subset such that \( P = \text{Plc}(S) \). Then
a. $S$ is the unique simple system contained in $P$;
b. if $r, s \in S$ then $r \cdot s \leq 0$; and
c. if $r \in P$ then there is an $s$ in $S$ with $r \cdot s > 0$.

**Proof:** If $r \in P - S$ then clearly $P = \text{Plc}(P - \{r\})$, so $r \in \text{Plc}(P - \{r\})$. Suppose $r \in S$. Suppose also that $r \in \text{Plc}(P - \{r\})$. Write $S = \{r_1, \ldots, r_s\}$ where $r = r_1, r_2 = \alpha_1 p_1 + \cdots + \alpha_t p_t$ where $\alpha_i > 0$ and $p_i \in P - \{r\}$, and $p_i = \beta_1 r_1 + \cdots + \beta_t r_t$ where $\beta_1 \geq 0$. Then $(1 - \gamma_1) r_1 = \gamma_2 r_2 + \cdots + \gamma_s r_s$ where $\gamma_i \geq 0$, whence $\gamma_i \leq 1$.

If $\gamma_1 < 1$ then $r_1 \in \text{Plc}(P_2, \ldots, r_s)$, which contradicts the minimality of $S$. Thus, $\gamma_1 = 1$, and so $\gamma_j = 0$ for $j > 1$. From this, $\beta_1 = 0$ for $j > 1$ whence $p_i = \beta_1 r_1$ for all $i$, whence $p_i = r$ for all $i$, a contradiction. We have thus shown that for $r \in P, r \in S$ iff $r \notin \text{Plc}(P - \{r\})$. Next, suppose for $r, s \in S$ that $r \cdot s > 0$. Let $t = \rho_r(S)$; then $t = s - \lambda r$ where $r > 0$. If $t \in P$ then since $s = t + \lambda r, s \in \text{Plc}(P - \{s\})$. If $-t \notin P$ then since $r = (1/\lambda) t + (1/\lambda)(-t), r \in \text{Plc}(P - \{r\})$. This is a contradiction in either case, so $r \cdot s \leq 0$, proving part b.

If $S$ is not linearly independent then $\sum_{s \in S_1} \alpha_s s = \sum_{s \in S_2} \alpha_s s$ where $S_1, S_2$ are disjoint nonempty subsets of $S$ and $\alpha_s > 0$ for all $s \in S_1 \cup S_2$. Letting $v$ denote the common value, $v \cdot v \leq 0$, whence $v = 0$. But this is impossible since $v$ is positive in the order on $\mathbb{R}^n$. This proves that $S$ is linearly independent, and so simple.

Suppose $S' \subseteq P$ is also a simple system. If $r \in P - S'$ then since $r \in \text{Plc}(S'), r \in \text{Plc}(P - \{r\})$. It follows that $S \subseteq S'$; and since $S$ and $S'$ both generate $\text{Span}(R), S = S'$. This proves part a. Suppose $r \in P$; write $r = \sum_{s \in S} \alpha_s s$ where $\alpha_s \geq 0$. Since $1 = r \cdot r = \sum_s \alpha_s (r \cdot s), r \cdot s > 0$ for some $s \in S$.

**Lemma 20.** Suppose $S$ is a simple system, $P$ is the positive system containing it, and $r \in S$. Then $\rho_r(P - \{r\}) = P - \{r\}$.

**Proof:** If $s \in P - \{r\}$ then $s = \sum_i \xi_i \alpha_i t$ where $\alpha_i > 0$ for some $t \neq r$, say $t_1$. Also, $\rho_r(s) = s - \lambda r$, and $\rho_r(t) = \sum_i \xi_i \beta_i t$. Since $\beta_1 = \alpha_1, \rho_r(s) \in P$. Finally, $\rho_r(s)$ cannot equal $r$, since $-r \notin P$.

**Theorem 21.** If $R$ is a unit root system and $S \subseteq R$ is a simple system then $R = G_R[S], \{\rho_r : r \in S\}$ generates $G_R$.

**Proof:** Let $P$ be the positive system containing $S$. Let $G_S$ be the group generated by $\{\rho_s \mid s \in S\}$. For $r \in P$ let $w(r) = \sum_{s \in S} c_s$ where $r = \sum_{s \in S} c_s s$. Let $t$ be an element in $G_S[r] \cap P$ of minimum $w$; we claim that (*) $t \in S$. It follows that $P \subseteq G_S[S]$. Since $-S \subseteq G_S[S], -P \subseteq G_S[S]$ also, which proves that $R = G_R[S]$. The second claim of the theorem follows, since by facts about reflections noted above $\rho_r \in G_S$ for any $r \in R$. To prove (*), let $u \in S$ be such that $t \cdot u > 0$, and let $v = s_u(t)$. Then $v = t - \alpha u$ where $\alpha > 0$; and $w(t) = w(v) + \alpha$. If $t \notin S$ then $v \in P$ by lemma 20, a contradiction.

**Theorem 22.** If $R$ is a unit root system then $G_R$ acts transitively on the positive systems.

**Proof:** We prove this by induction on $|P \cap (-Q)|$, for positive systems $P$ and $Q$. If this is 0 then $Q = P$. Otherwise, there must be an element $r \in S \cap (-Q)$ where $S$ is the simple system of $P$. Let $P' = \rho_r[P]$ and $Q' = \rho_r[Q]$. Then $|P \cap (-Q')| = |P \cap (Q) - 1$, the last equality following by lemma 20.

In fact, the action is regular; but we omit this. [BourL] is an encyclopedic reference for topics covered in this section; a proof of regularity for finite reflection groups may be found in [Vara]
such a graph is called a Coxeter graph, or Coxeter diagram. In a common variation, the label $m_{rs}$ is omitted if it is 3.

A Coxeter group has a standard realization. Letting $M$ denote the Coxeter matrix, define a symmetric bilinear form with matrix $Q$ on $\mathbb{R}^n$, where $Q_{rs}$ equals $-\cos(\pi/m_{rs})$. $Q$ is a symmetric matrix with 1’s on the diagonal, and negative values off the diagonal. Let $\rho_r$ for $1 \leq r \leq n$ denote the linear transformation $x \mapsto x - 2(\varepsilon_r^t Q x)e_r$, where $e_r$ is as usual the $r$th standard unit vector.

If $w$ and $u$ are vectors in $\mathbb{R}^n$ with $w \cdot u \neq 0$, the transformation with matrix $I - 2(uu^t)/(u^t u)$ is readily verified to map $u$ to $-u$, to fix vectors $x$ with $u^t x = 0$, and to have determinant $-1$. Such a transformation might be called a “shear reflection”. The map $\rho_r$ defined above is of this form, with $u = e_r$ and $w = Q e_r$.

The $\rho_r$ no longer lie in $O(n)$. Let $O_Q(n)$ denote the group of linear transformations which preserve $Q$. For $u$ with $u^t Q u = 1$ let $\rho_u$ be the map taking $x$ to $x - 2(u^t Q x) u$; then $\rho_u \in O_Q(n)$. To see this, write $x$ as $x_H + au$, where $u^t Q x_H = 0$, so that $\rho(x) = x_H - au$. Similarly write $y$, and expand $x^t Q y$ and $\rho(x)^t Q \rho(y)$ by linearity. By an argument similar to one given above, it follows that if $g \in O_Q(n)$ is a transformation with $g(u) = v$, then $\rho_v = g \rho_u g^{-1}$.

Given two such maps $\rho_u$ and $\rho_v$, in the basis $u,v$,

$$\rho_u = \begin{bmatrix} -1 & -2u^t Q v \\ 0 & 1 \end{bmatrix}.$$ 

From this, if $u^t Q v = -\cos \theta$ where $\theta = \pi/m$ for $m$ a positive integer and $i$ an integer,

$$(\rho_u \rho_v)^i = \frac{1}{\sin \theta} \begin{bmatrix} \sin(2i+1)\theta & -\sin 2i\theta \\ \sin 2i\theta & -\sin(2i-1)\theta \end{bmatrix}$$

(for the induction step use the identity $\sin(\phi_1 + \theta) \sin(\phi_2 + \theta) - \sin \phi_1 \sin \phi_2 = \sin \theta \sin(\phi_1 + \phi_2 + \theta)$). Since $\rho_u$ and $\rho_v$ act on the subspace orthogonal to $u$ and $v$, it follows that the order of $\rho_u \rho_v$ is $m$. Thus, the group generated by $\rho_u$ and $\rho_v$ is the dihedral group.

Let $G$ denote the Coxeter group with generators $\rho_r$ and relations $(\rho_r \rho_s)^{m_{rs}} = 1$. Letting $W$ denote the set of words in the letters $\rho_r$, and $W_K$ the words $\rho_r \rho_s \cdots \rho_r \rho_s$ where there are $m_{rs}$ pairs, An element of $G$ is an equivalence class of words $w$, where $w_1 \equiv w_2$ iff there is a sequence of insertions and deletions of words from $W_K$. Letting $\bar{\rho}_r$ denote the reflection as above, by the above observations there is a homomorphism $w \mapsto \bar{w}$ from $G$ to the subgroup $\bar{G}$ of $O_Q(n)$ generated by the $\bar{\rho}_r$.

For $w \in W$ let $l(w)$ denote the minimum length of a word $w'$ such that $w \equiv w'$. Clearly, $l(w_1 w_2) \leq l(w_1) + l(w_2)$. In particular, $l(w \rho_r) \leq l(w) + 1$; also $l(w) \leq l(w \rho_r) + 1$ because $w = w \rho_r \rho_r$. Further, $l(w)$ mod 2 is a homomorphism on $W$, because the words in $W_K$ all have even length. It follows that $l((w \rho_r) \bar{w}(w) \rho_r) = l(w) \pm 1$.

In the dihedral group with generators $\rho_r, \rho_s$ and relations $\rho_s^2 = 1, \rho_r^2 = 1$, and $(\rho_r \rho_s)^m = 1$, every alternating product $\rho_r \rho_s \cdots$ and $\rho_s \rho_r \cdots$, of length less than $m$, is a distinct group element. The two products of length $m$ are the same. This can readily be seen using the characterization $\rho^i \sigma^j$ of the elements given above, where $\rho = \rho_r$, $\sigma = \rho_r \rho_s$, using the relation $\rho_r \rho_s \cdots = \rho_s \rho_r \cdots$, which is often called the braid relation.

For the following, let Plc denote those vectors in $\mathbb{R}^n$ which are positive linear combinations of the basis vectors $\{e_r\}$. Let $v$ be a word as in the preceding paragraph, of length $k < m$, and ending in $\rho_s$. We claim that $\bar{v}(e_r) \in$ Plc. This follows using the matrix $M$ for $\rho_r \rho_s$ given above, noting that $\sin i \pi/m \geq 0$ if $i \leq m$. If $k$ is even and $i = k/2$ then $2i + 1 \leq m$, so the first column of $M^i$ has nonnegative entries. If $k$ is odd and $i = (k - 1)/2$ then $2i + 2 \leq m$; also $(\bar{\rho}_s \rho_r)^i \bar{\rho}_s(e_r) = (\bar{\rho}_s \rho_r)^i \bar{\rho}_s(-\rho_r(e_r)) = -(\bar{\rho}_s \rho_r)^{i+1}(e_r)$, and the second column of $-M^{i+1}$ has nonnegative entries.

Lemma 23. For a Coxeter group $G$, and $w \in G$, if $l(w \rho_r) = l(w) + 1$ then $\bar{w}(e_r) \in$ Plc.
Proof: Suppose \( w, r \) is a counterexample with \( l(w) \) least. Since \( l(e_r) = 1, l(w) = 1 \). For some \( w_1 \) and \( s, w = w_1 \rho_s \) and \( l(w_1) = l(w) - 1 \). Inductively, if \( l(w_i \rho_k t) = l(w_i) - 1 \) where \( t = r \) if \( i \) is odd, or \( t = s \) if \( i \) is even, let \( w_{i+1} \) be such that \( w_i = w_{i+1} \rho_t \) and \( l(w_{i+1}) = l(w_i) - 1 \). Let \( k \) be the least \( i \) least such that \( l(w_i \rho_k t) = l(w_i) + 1 \) for both \( t = r \) and \( t = s \). Then \( w = w_k v \) where \( v \) is an alternating product of \( k \) factors \( \rho_r, \rho_s \) ending in \( \rho_s \). Further \( l(v) = k \) since \( l(w) = l(w_k) + k \); from this, \( 1 \leq k \leq m \). Further, since \( l(w_i \rho_k t) = l(w) + 1 \) by hypothesis, \( l(w_k v \rho_r t) = l(w_k) + k + 1 \), whence no minimal expression for \( v \) can end in \( \rho_r \), whence \( k < m \). By remarks above, \( \lambda v = \lambda \rho_r + \mu \rho_s \), where \( \lambda, \mu \geq 0 \). By the induction hypothesis, \( \bar{w}_k(e_r) \in \text{Plc} \) and \( \bar{w}_k(e_s) \in \text{Plc} \). Thus, \( \bar{w}(e_r) = \bar{w}_k \bar{v}(e_r) = \lambda \bar{w}_k(e_r) + \mu \bar{w}_k(e_s) \), and so \( (w \rho_r e_r) \in \text{Plc} \), a contradiction.

Theorem 24. For a Coxeter group \( G \), the homomorphism \( w \mapsto \bar{w} \) from \( G \) to \( \bar{G} \) is an isomorphism.

Proof: It is surjective by definition. If \( \bar{w} = 1 \) and \( w \neq 1 \) then for some \( r, w = w' \rho_r \) and \( l(w) = l(w') + 1 \).

By lemma 23, \( \bar{w}'(e_r) \in \text{Plc} \). But \( -\bar{w}'(e_r) = \bar{w}' \bar{\rho}_r(e_r) = \bar{w}(e_r) = e_r \), so \( -\bar{w}'(e_r) \in \text{Plc} \), a contradiction.

Figure 1 gives the complete list of Coxeter graphs which are those of an irreducible finite Coxeter group, and all are finite reflection groups.
Except for the last graph, the subscript of the name is the number of nodes. The restrictions on the subscript ensure that there is no repetition. The names are given as used in the classification of simple Lie algebras over \( \mathbb{C} \). The last three graphs do not occur in this context. There are two families of such which yield the same Coxeter graph, \( B_n \) and \( C_n \).

We now define a root system in \( \mathbb{R}^n \) to be a finite set \( R \) of arbitrary vectors, such that \( \rho_r|\mathbb{R}| = R \) for all \( r \in R \). A root system is said to be reduced in if for every \( r \in R \), \( \lambda r \in R \) for a real number \( \lambda \) if \( \lambda \neq \pm 1 \). Henceforth only reduced root system will be considered.

If \( R \) is a root system then the unit vectors \( r/|r| \) for \( r \in R \) form a unit root system, which we denote \( R_u \). A positive system in \( R \) is defined as for \( R_u \); the positive systems in \( R \) are in obvious 1-1 correspondence with those in \( R_u \). Similar remarks apply to simple systems.

\( R \) is said to be integral, or crystallographic, if \( 2(r \cdot s)/(r \cdot r) \) is an integer for all \( r, s \in R \). This restriction is of interest in many applications of root systems. For all \( r, s \in R \) in an integral root system, \( \mu_{ij} = 4(r \cdot s)^2/((r \cdot r)(s \cdot s)) \), or \( 4 \cos \theta_{rs} \), is an integer, and clearly this integer is between 0 and 4. The value 4 occurs when \( s = \pm r \), and need not be considered further. In the other cases, the angle \( \theta \), the ratio \( \lambda \) of the larger of \(|r|, |s|\) to the smaller, and the order \( m_{rs} \) of \( \rho_r \rho_s \), are as follows.

\[
\begin{array}{cccc}
\mu_{ij} & \theta & \lambda & m_{rs} \\
0 & \pi/2 & & 2 \\
1 & \pi/3, 2\pi/3 & 1 & 3 \\
2 & \pi/4, 3\pi/4 & \sqrt{2} & 4 \\
3 & \pi/6, 5\pi/6 & \sqrt{3} & 6 \\
\end{array}
\]

The second angle is the one whose cosine is negative, i.e., the angle when \( r, s \) are elements of a simple system. The relative lengths are obtained by noting that there is (up to order) only one possible factorization \((2(r \cdot s)/(r \cdot r))/((r \cdot r)(s \cdot s))\) of \( \mu_{ij} \); there is no relative length information when \( \mu_{ij} = 0 \).

The finite reflection group of an integral root system is called a Weyl group. In a variation of the Coxeter diagram, called the Dynkin diagram, the labels 4 and 6 are replaced by double and triple edges; and an arrow is added for these pointing toward the shorter root. In \( B_n \) to the left, and in \( C_n \) to the right.

We will determine the allowed connected Dynkin diagrams. Define a scheme to be a linearly independent subset \( \Sigma \subseteq \mathbb{R}^n \), where \( 2(r \cdot s)/(r \cdot r) \) is a nonpositive integer for all \( r, s \in \Sigma \). A scheme has a graph, with the vectors as vertices, and an edge labeled \( \mu_{rs} \), defined exactly as above, between vertices \( r \) and \( s \) when \( 1 \leq \mu_{rs} \leq 3 \). Note that a subset \( \Sigma' \subseteq \Sigma \) of a scheme is a scheme, and the graph of \( \Sigma' \) is a full subgraph of the graph of \( \Sigma \).

**Lemma 25.** If \( \Sigma \) is a scheme with \( n \) elements then the graph contains at most \( n - 1 \) edges.

**Proof:** Let \( v = \sum_{r \in \Sigma} r/|r| \). Since \( \Sigma \) is linearly independent \( v \neq 0 \), whence \( 0 < |v|^2 = n + 2 \sum_{E} (r \cdot s)/(|r||s|) \) where \( E \) is the set of edges \( \{r, s\} \). Each term of the sum is less than or equal to \( 1 \), so the size of \( E \) can be at most \( n - 1 \).

It follows that the components of the graph of a scheme are trees.

**Lemma 26.** For any vertex in a scheme, the sum of the labels on the incident edges cannot exceed 3.

**Proof:** Suppose \( s_1, \ldots, s_l \) are the vertices adjacent to \( r \). Then \( s_i \) and \( s_j \) are perpendicular for \( i \neq j \). Let \( v = \sum_i ((r \cdot s_i)/(s_i \cdot s_i))s_i \). Then \( |v| < |r| \), and \( 4v \cdot v = \sum_i \mu_{rs} r \cdot s \), so \( \sum_i \mu_{rs} < 4 \).

If \( \Sigma' \) is a subscheme of \( \Sigma \), the contraction of the graph of \( \Sigma \) by \( \Sigma' \) is obtained by replacing \( \Sigma' \) by a single node \( t \), and for \( r \notin \Sigma \), adding an edge \( \{r, t\} \) whenever there was an edge \( \{r, s\} \) with \( s \in \Sigma \). By lemma 25, given \( r \) there can be at most one \( \{r, s\} \), and \( \{r, t\} \) is given the same label.
LEMMa 27. If Σ is a scheme, and Σ′ is a path in Σ, all of whose edges have label 1, then the contraction of Σ by Σ′ is a scheme.

PROOF: Let r_1, ..., r_n be the vertices of the path, and let v = r_1 + ... + r_n. By remarks above, r_i · r_j = r_i · r_j, 2r_i · r_{i+1} = -r_i · r_i, and r_i · r_j = 0 otherwise. It follows that v · v = r_1 · r_1. Further, if \{r_i, s\} is any edge of Σ with s \notin Σ′ then v · s = r_1 · s. The lemma follows.

THEOREM 28. If the graph of a scheme is connected it must be one of A_n, B_n, C_n, D_n, G_2, F_4, E_6, E_7, or E_8, in the list given in figure 1.

PROOF: Let Γ be the graph. If Γ has an edge labeled 3 it must be G_2. Γ cannot contain two edges labeled 2, else the path joining them could be contracted, resulting in a graph contradicting lemma 26. Similarly if Γ contains an edge labeled 2 it cannot contain a vertex of degree 3 (there would be one closest to one of the vertices of the edge labeled 2); thus, it must be a simple path. Suppose r_1, ..., r_p are the vertices away from one end of the edge labeled 2, and s_1, ..., s_q are those away from the other. Let v = \sum_i (p + 1 - i)r_i, and w = \sum_i (q + 1 - i)s_i. Then

v · v = \frac{p(p + 1)}{2}r_1 · r_1, \quad w · w = \frac{q(q + 1)}{2}s_1 · s_1, \quad v · w = pq r_1 · s_1.

We may assume w.l.o.g. that s_1, s_1 = 2r_1, and r_1 · s_1 = -r_1 · r_1. Since the r_i and s_j together are linearly independent, w is not a multiple of v, and so by the Cauchy-Schwarz inequality (v · w)^2 < (v · v)(w · w). From this, pq < \frac{1}{2}((p - 1)(q - 1), from which it follows that either p = 1, q = 1, or p = q = 2, giving the graphs B_n, C_n, and F_4 respectively. As usual, if Γ contains a degree 3 vertex it can only contain one. Let r_{ij} for 1 \leq i \leq 3 and 1 \leq j \leq p_j be the vertices of the i-th path from the degree 3 node r, with r_{ii} = r_{ii} = r. Suppose p_1 \geq p_2 \geq p_3 \geq 2. Let v_i = \sum_{j \geq 2} (p_i + 1 - j)r_{ij}. Then

v_i · v_i = \frac{p_i(p_i - 1)}{2} c, \quad v_i · v_i = 0, i \neq j; \quad r · v_i = -\frac{1}{2} c.

Let M be the Gram matrix, with entries as above. Since M is full rank (by the linear independence of the scheme vectors), det(M) is positive. It is readily seen that det(M) = \frac{c^4}{16}(p_1 - 1)(p_2 - 1)(p_3 - 1)(p_1p_2 + p_1p_3 + p_2p_3 - p_1p_2p_3). It follows that 1/p_1 + 1/p_2 + 1/p_3 > 1, whence p_3 = 2, whence 1/p_1 + 1/p_2 > 1/2. From this, either p_2 = 2 and p_1 \geq 2 (D_n), or p_2 = 3 and 3 \leq p_1 \leq 5 (E_6, E_7, E_8). Finally, if none of the above hold then Γ is A_n for some n.

Vectors for the irreducible root system are as follows; we leave it to the reader to determine simple systems and verify the correspondence with the Dynkin diagrams.

A_n. The vectors in \mathbb{R}^{n+1} with a one +1 and one -1.
B_n. The vectors in \mathbb{R}^n with one ±1, or two ±1.
C_n. The vectors in \mathbb{R}^n with two ±1, or one ±2.
D_n. The vectors in \mathbb{R}^n with two ±1.
E_8. The vectors in \mathbb{R}^8 with two ±1, or all entries ±(1/2), with an even number of -(1/2).
E_7. The vectors of E_8 in the hyperplane normal to a root.
E_6. The vectors of E_8 in the intersection of two hyperplanes normal to a root, where the roots are not parallel or orthogonal.
F_4. The vectors in \mathbb{R}^4 with two ±1, one ±1, or all entries ±(1/2).
G_2. Let a = (1, 0), b = (-3/2, \sqrt{3}/2) (these are a fundamental system). The vectors are a, 3a + b, 2a + b, 3a + 2b, a + b, b, and their negatives.

A root lattice may be defined as one generated by any of the above root systems. It suffices to consider \mathbb{Z}^n, A_n, D_n, and E_n. Note that the version of the D_4 lattice given here differs from that mentioned in chapter 20. See [ConSl] for further discussion.

310
7. The wallpaper groups. An isometry of Euclidean space $\mathbb{R}^n$ is a function $\phi : \mathbb{R}^n \to \mathbb{R}^n$, which preserves distance, i.e., such that $|\phi(x) - \phi(y)| = |x - y|$. The isometries of $\mathbb{R}^n$ form a group under composition, called the Euclidean group and denoted $E(n)$. Although we are principally interested in $n = 2$ in this section, some facts will be proved more generally.

The map $x \mapsto x + a$ where $a$ is a fixed element of $\mathbb{R}^n$ is clearly an isometry, called a translation; we will use $\tau_a$ to denote it. If $\phi$ is any isometry, suppose $\phi(0) = a$, and consider the map $\gamma = \tau_{-a}\phi$; then $\gamma$ is an isometry fixing 0, and $\phi = \tau_a\gamma$. Note that the decomposition depends on the coordinatization of Euclidean space.

An isometry $\gamma$ fixing 0 may be seen to be a linear transformation (in particular it is not necessary to assume the latter in addition). Indeed, consider the points 0, $x, cx$ where $c \in \mathbb{R}$; these lie on a line. They map to 0, $\gamma(x), \gamma(cx)$; it is clear that these also lie on a line, and $\gamma(cx) = c\gamma(x)$. If $x$ and $y$ do not lie on a line, then $z = x + y$ iff $0, x, y, z$ are the vertices of a parallelogram; and by further geometric arguments it follows that $\gamma(z) = \gamma(x) + \gamma(y)$.

An isometry preserves $x \cdot x$. Since $x \cdot y = ((x + y) \cdot (x + y) - x \cdot x - y \cdot y)/2$, if $\gamma$ is an isometry fixing 0 then $\gamma \in O(n)$. Thus, if $\phi \in E(n)$ then $\phi = \tau\gamma$ where $\gamma \in O(n)$ and $\tau$ is a translation. Clearly $\tau$ is unique, namely $\tau_a$ where $a = \phi(0)$. Also, $\phi$ is determined by its action in $n + 1$ points, which do not lie in an affine hyperplane (for example the origin and the $n$ standard unit vectors).

Writing $E$ for $E(n)$, $O$ for $O(n)$, and letting $\gamma, \tau$ denote elements of $O$ and translations respectively, the following are readily verified.

- $O$ is the subgroup of $E$ of isometries which fix 0.
- The translations form a commutative subgroup $T \subseteq E$, indeed $\tau_a \tau_b = \tau_{a+b}$.
- $E = TO$, and $T \cap O = \{1\}$.
- Since $\gamma(x + a) = \gamma(x) + \gamma(a)$, $\gamma\tau_a = \tau_{\gamma(a)}\gamma$.
- Writing $\tau\gamma$ for $\gamma\gamma^{-1}$, $\tau_a^2 = \tau_{\gamma(a)}$. Indeed $\tau_a^2 = \tau_{\phi(a)}$ for $\phi \in E$; in particular $T$ is a normal subgroup of $E$.
- $\tau_2\tau_1\gamma_1 = \tau_2\tau_1^2\gamma_1\gamma_1$.
- The map $\tau\gamma$ is a homomorphism.
- The inclusion of $O$ in $E$ demonstrates that the extension $1 \mapsto T \mapsto E \mapsto G \mapsto 1$ splits.
- $E$ is the semidirect product of $T$ by $O$ (see section 19.9); $\gamma \mapsto \tau\gamma$ is the map $O \mapsto \text{Aut}(T)$.

As in the previous section let $\rho_v \in O(n)$ denote the map $x \mapsto x - 2\frac{v \cdot x}{v^2}v$, the reflection in the hyperplane normal to $v$.

**Theorem 29.** Any $\gamma \in O(n)$ is a product of at most $n$ reflections.

**Proof:** We map suppose $\gamma$ is not the identity, say $\gamma(u) \neq u$. Let $v = f(u) - u$, so that $\rho_v$ is the reflection in the “perpendicular bisector” of the line segment from $u$ to $f(u)$. Then

$$2f(u) \cdot v f(u) = 2\frac{f(u) \cdot (f(u) - u)}{(f(u) - u) \cdot (f(u) - u)}v = v = f(u) - u,$$

whence $\rho_v(f(u)) = u$. Letting $\gamma' = \rho_v\gamma$, $\gamma'$ fixes $u$. Also, if $v \cdot x = 0$ then $\gamma'(v) \cdot \gamma'(x) = 0 = v \cdot \gamma(x)$; that is, $\gamma'$ acts on $v^\perp$. Inductively, $\gamma'$ is a product of at most $n - 1$ reflections.

This theorem holds more generally for the group preserving a symmetric bilinear form on a finite dimensional vector space over a field of characteristic other than 2; see [Jacobson] for example. In $E(n)$, a map $\tau_a\rho_v\tau_{-a}$ is also called a reflection. It maps $x$ to $2\frac{v \cdot x}{v^2}v$, and is a reflection in the affine hyperplane $\{x : u \cdot (x - a) = 0\}$. It is readily verified that whether a map is of this form does not depend on the coordinatization. Further, by the theorem any $\phi \in E(n)$ is a product of at most $n + 1$ such maps, since if $\phi(0) = a$, $a$ may be mapped to 0 by a reflection.
The map $\tau_0 \mapsto \det(\gamma)$ is a homomorphism from $E(n)$ to $\{+1, -1\}$. Isometries mapping to $+1$ are called proper, or direct; and those mapping to $-1$ improper, or inverse.

The elements of $E(2)$ may be classified as follows.
- Reflections. These are improper, and leave a line fixed.
- Translations. These are a product of two reflections, with parallel reflection lines. They are proper, and except for the identity have no fixed points.
- Rotations (about an arbitrary center). These are a product of two reflections, with the center of rotation being the intersection of the reflection lines (see the previous section). They are proper, and except for the identity have one fixed point.
- Glide reflections. The only remaining case is an improper isometry $\phi$ having no fixed points. Let $u$ be a point, let $v = \phi(u)$, and let $w = (u + v)/2$. Let $\psi$ be the rotation around $w$ through the angle $\pi$. Then $\psi\phi$ is improper and fixes $u$, hence is a reflection $\rho$. Then $\phi = \psi\rho$; from this, it follows that $\phi = \rho'\tau_0 = \tau_a\rho'$ where $\rho'$ is the reflection in the line through $w$ and $\rho(w)$ (consider the midpoint $o$ and a third point, on the reflection line of $\rho$ and distinct from $o$). Such an isometry (a reflection followed by a translation along the reflection line) is called a glide reflection. In the special case that $w$ lies on the reflection line of $\rho$, $\psi$ is the reflection in the perpendicular line through $w$.

A group $G$ acting on a topological space $X$ is said to act discontinuously if the orbits are discrete subsets of $X$. In the case of a discrete subgroup $D \subseteq E(n)$, with notation as above, $T \cap D$ is a commutative normal subgroup of $D$. Letting $L$ be the orbit of $0$ under $T \cap D$, $L$ is a lattice. For $a \in L$ and $\phi \in D$, $\tau_a \in T \cap D$. Since $\tau_{\phi(a)} = \phi\tau_a\phi^{-1}$, $\tau_{\phi(a)} \in T \cap D$ also, whence $\phi(a) \in L$. We have shown that $D$ acts on $L$.

A plane crystallographic group, or "wallpaper" group, is defined to be a discontinuous subgroup of $E(2)$, which contains translations $\tau_a$ and $\tau_b$ where $a$ and $b$ are linearly independent. Such a group may be realized as the symmetry group of a wallpaper pattern, although this topic will not be considered.

**Theorem 30.** If a discontinuous subgroup of $E(2)$ contains a translation and a rotation, then the angle of the rotation is $2\pi/n$ where $n$ equals $2, 3, 4, \text{ or } 6$.

**Proof:** Suppose that the rotation is about $x$, and let $y \neq x$ be such that $|y - x|$ is as small as possible. The angle of a rotation $\psi_x$ around $x$ may be taken as $2\pi/n$ for some $n$, where $n$ is the order of the cyclic group generated by $\psi_x$, which must be finite by the discreteness hypothesis. If $n > 6$ then $|y - \psi_x(y)| < |y - x|$, and if $n = 5$ then $|\psi_y(x) - \psi_x(y)| < |y - x|$, proving the theorem.

Suppose from hereon that $E = E(2)$, and $D \subseteq E$ is a wallpaper group. If $L$ is a lattice on which $D$ acts, it will be assumed to have $a, b$ as a basis, where $a$ is as short as possible, and the angle between $a$ and $b$ is as large as possible, not exceeding $\pi/2$. If $D$ contains a rotation, we assume that $0$ is a center of rotation. We let $T$ denote the translations of $E$ and $P \subseteq D$ the stabilizer of 0.

By general facts about permutation groups, if $L$ is the orbit of $0$ under $T \cap D$, $\gamma \in P$, and $a \in L$, then $\tau_a \phi \tau_{\gamma^{-1}}$ stabilizes $a$, and is the "same" orthogonal transformation as $\phi$, translated to $a$. Similarly, if $l$ is any line of $L$, a reflection in $l$ is conjugate of a reflection in a line through $0$; and similarly for glide reflections. On the other hand, any rotation center, reflection line, or glide reflection line may be translated so that it intersects the fundamental parallelogram.

The full symmetry group of a two-dimensional lattice will certainly be a wallpaper group. A subgroup $D$ of the full symmetry group of $L$ has $a$ and $b$ in the orbit of $0$ iff it contains $\tau_a \gamma_a$ for some $\gamma_a \in O$, and $\tau_b \gamma_b$ for some $\gamma_b \in O$. The orbit of $0$ under $D \cap T$, which will be denoted $L_T$, might be a proper subset of the orbit under $D$, which will be denoted $L_0$. As will be seen, in all cases $L_0$ is a lattice, with $L_T$ a sublattice.

To classify the wallpaper groups, we consider the symmetry possessed by a lattice. Various details will be left to the reader. Let $\psi_n$ denote the rotation through angle $2\pi/n$ around the origin. Let $\rho_A \ (\rho_B)$ denote the reflection in the line through $\{0, a\} \ (\{0, b\})$. 

312
If \(|b| > |a|\) and the angle between \(a\) and \(b\) is less than \(\pi/2\), the lattice is said to be oblique. The wallpaper groups acting on an oblique lattice, and their generators, are

\[
\begin{align*}
p2 & \quad \psi_2, \tau_a, \tau_b \\
p1 & \quad \tau_a, \tau_b \\
\end{align*}
\]

The full symmetry group of an oblique lattice \(L\) is \(p2\), as may be seen by considering each of the four types of isometries. The only subgroup with \(L_0 = L\) is \(p1\). These are the only wallpaper groups acting on an oblique lattice, and they act on any lattice.

If \(|b| > |a|\) and the angle between \(a\) and \(b\) is equal to \(\pi/2\), the lattice is said to be rectangular. Additional wallpaper groups in this case are

\[
\begin{align*}
pmm & \quad \rho_{A}, \rho_{B}, \tau_a, \tau_b \\
pmg & \quad \rho_{A}, \tau_a, \tau_b \rho_{B} \\
pgg & \quad \tau_a \rho_{A}, \tau_b \rho_{B} \\
pg & \quad \tau_a, \tau_b \rho_{B} \\
\end{align*}
\]

The full symmetry group of a rectangular lattice \(L\) is \(pmm\). To see this, \(L\) yields a tiling of the plane with rectangles. The symmetries which fix 0 form a copy of the four group \(\mathbb{Z}_2 \times \mathbb{Z}_2\); this is also true for the symmetries which fix the center of a rectangle edge, or a rectangle center.

To see that all these are in \(pmm\), it suffices to note that the reflection in the line parallel the \(x\)-axis, through the center of the rectangle, is \(\tau_b \rho_{A}\).

To see that the subgroups with \(L_0 = L\) are all listed, note that if \(\tau_a \notin D\) then \(\tau_a \rho_{A}\) must be in \(D\). By further such reasoning, any wallpaper group with \(L_T\) rectangular must be one of the above (including \(p2\), \(p1\), with \(L_0\) containing \(L_T\) as a sublattice of index 2 or 4.

If \(|b| = |a|\) and the angle between \(a\) and \(b\) is neither \(\pi/2\) nor \(\pi/3\), the lattice is said to be rhomboidal. By the restrictions on \(a\) and \(b\), the angle is between \(\pi/3\) and \(\pi/2\). Let \(\rho_C\) (\(\rho_D\)) denote the reflection in the line through 0 and \(a + b\) (\(a - b\)). Additional wallpaper groups in this case are

\[
\begin{align*}
cmm & \quad \rho_C, \rho_D, \tau_a, \tau_b \\
cm & \quad \rho_C, \tau_a, \tau_b \\
\end{align*}
\]

The full symmetry group of a rombohedral lattice \(L\) is \(cmm\). To see this, \(L\) yields a tiling of the plane with rhombuses. The symmetries which fix 0 form a copy of the four group \(\mathbb{Z}_2 \times \mathbb{Z}_2\); this is also true for the symmetries which fix a rhombus center.

The proof that these are all in \(cmm\) is similar to the rectangular lattice case. The list of subgroups is clearly complete. These groups act on the rectangular lattice; the orbit of 0 is a rhomboidal sublattice.

None of the groups listed so far are isomorphic. The actions of \(\psi_2, \rho_{A}\), and \(\rho_C\) by conjugation on the \(\mathcal{Z}\)-module with basis \(\{\tau_a, \tau_b\}\) yield matrices with different eigenvalues, and this implies that \(p2\), \(pm\), and \(cm\) are non-isomorphic. Other possibilities are left to the reader.

If \(|b| = |a|\) and the angle between \(a\) and \(b\) is \(\pi/2\), the lattice is said to be square. Note that this is the case if \(D\) contains \(\psi_4\). Additional wallpaper groups in this case are

\[
\begin{align*}
p4m & \quad \psi_4, \rho_{A}, \tau_a, \tau_b \\
p4g & \quad \psi_4, \tau_a, \tau_b \rho_{B} \\
p4 & \quad \psi_4, \tau_a, \tau_b \\
\end{align*}
\]

The full symmetry group of a square lattice \(L\) is \(p4m\). To see this, \(L\) yields a tiling of the plane with squares. The symmetries which fix 0 form a copy of the dihedral group \(D_4\). The symmetries which fix the center of a square edge form a copy of \(\mathbb{Z}_2 \times \mathbb{Z}_2\). The symmetries which fix the center of a square form a copy of \(D_4\).

Finally there are glide reflections, with line through the midpoints of two adjacent sides of a square.

To see that all these are in \(p4m\), some symmetries follow as for a rectangular lattice. For the remaining cases, it suffices to observe that \(\tau_a \psi_4\) maps \(0 \mapsto a \mapsto a + b \mapsto b \mapsto 0\), and \(\tau_a \rho_{A} \psi_4^2\) maps \(0 \mapsto a \mapsto a + b \mapsto a + 2b \mapsto b \mapsto 0\), which yields a tiling of the plane with squares.
Also, \( \tau \) clockwise from \( \tau \) that approximations to \( r \) of a triangle (other glide reflections are the product of a reflection and a translation).

The symmetries which fix 0 form a copy of the dihedral group \( D \). The symmetries which fix the center of a triangle edge form a copy of \( \mathbb{Z} \). The symmetries which fix the center of a triangle form a copy of \( D_3 \). Finally there are glide reflections, with line through the midpoints of two sides of a triangle (other glide reflections are the product of a reflection and a translation).

To see that all these are in \( p6m \), number the points of a hexagon 0 for the center, and 1 to 6 counterclockwise from \( a \). Then using permutation cycle notation, on these points \( \psi_6 = (123456) \) and \( \rho_A = (26)(35) \). Also, \( \tau_a \) maps 0,3,4,5 to 1,2,0,6 respectively, and \( \tau_b \) maps 0,1,2 to 1,0,2; and \( \tau_b \rho_A \psi_6^3 \) maps 0,1,2 to 1,0,2; and \( \tau_b \rho_A \) maps 4 \( \mapsto \) 3 \( \mapsto \) 0 \( \mapsto \) 2 \( \mapsto \) 1 \( \cdots \).

Some of the rectangular lattice groups act on the hexagonal lattice, for example \( p6m \). This is not considered a distinct wallpaper group; two groups are considered the same if there is an affine transformation which induces a group isomorphism.

[Kozo] was a useful reference for this section. The crystallographic groups in three dimensions were classified in 1891. These are also called the space groups. There are 230 of them; see [Zach], [CHFT]. Crystallographic groups in higher dimension are a topic of research. Some fundamental theorems were proved by Bieberbach in 1910, including that there are only finitely many in any dimension; see [Charl].

8. Continued Fractions. Continued fractions provide a representation of a real number with useful properties. In particular there is a theory of Diophantine approximation to a single real based on continued fractions; we give a treatment of this. The section concludes with a useful fact about Euclid’s algorithm for the gcd.

Any real number \( r = r_0 \) other than an integer can be written uniquely as \( r_0 = [r_0] + 1/r_1 \) where \( r_1 > 1 \). This step may be repeated, letting \( r_n = [r_n] + 1/r_{n+1} \), or \( r_{n+1} = 1/(r_n - [r_n]) \), as long as \( r_n \) is not an integer. Let \( a_n = [r_n] \), and call the sequence of \( a_n \), which may be finite or infinite, the continued fraction expansion of \( r \). The values \( a_n \) are integers, positive if \( n > 0 \); further if the sequence is finite, with last member \( a_N \), then \( a_N \neq 1 \), unless \( r_0 = 1 \) and \( N = 0 \). Call such a sequence regular.

Given any finite sequence \( a_0, \ldots, a_n \) of real numbers with \( a_n \neq 0 \) if \( n > 0 \), let \( [a_0, \ldots, a_n] \) denote \( a_0 \) if \( n = 0 \), or inductively \( a_0 + 1/[a_1, \ldots, a_n] \) if \( n > 0 \). That is, \( [a_0, \ldots, a_n] \) denotes

\[
a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}.
\]

A simple induction shows that

\[
[a_0, \ldots, a_n] = [a_0, \ldots, a_i, [a_{i+1}, \ldots, a_n]], \quad 0 \leq i < n.
\]

Also, if \( a_n \) is the continued fraction expansion of \( r \) then \( r = [a_0, \ldots, a_{n-1}, r_n] \) for \( n \geq 0 \); the quantities \( [a_0, \ldots, a_n] \) are called the convergents to \( r \). It should be clear that these form a sequence of rational approximations to \( r \); in fact, as we shall see, they are in a sense the best rational approximations to \( r \).
Lemma 31. Given a (finite or infinite) sequence \(a_n\) of real numbers, let
- \(p_0 = a_0, q_0 = 1,\)
- \(p_1 = a_1a_0 + 1, q_1 = a_1,\)
- \(p_n = a_np_{n-1} + p_{n-2}, q_n = a_nq_{n-1} + q_{n-2}\) for \(n \geq 2.\)

Then the following hold.

a. \([a_0, \ldots, a_n] = p_n/q_n.\)

b. \(p_nq_{n-1} - p_{n-1}q_n = (-1)^{n-1}, n \geq 1.\)

c. \(p_nq_{n-2} - p_{n-2}q_n = (-1)^n a_n, n \geq 2.\)

d. If \(a_n > 0\) for \(n > 0\) then \(q_n\) is positive, and increasing for \(n > 0.\)

e. If \(a_n > 0\) for \(n > 0\) then the even convergents are increasing, the odd are decreasing, and every odd is greater than every even.

f. If \(a_n\) is an integer for all \(n\) then \(p_n\) and \(q_n\) are integers, and \(p_n/q_n\) is in lowest terms. For \(n \geq 2.\)

Proof: Part a is proved by induction on \(n; n = 0\) and \(n = 1\) are immediate (since \([a_0, a_1] = (a_1a_0 + 1)/a_1).\)

For \(n \geq 2,\) write 
\[
[a_0, \ldots, a_n, a_{n+1}] = [a_0, \ldots, a_n + 1/a_{n+1}];
\]
by induction this equals 
\[
(a_n + 1/a_{n+1})p_{n-1} + p_{n-2} \over (a_n + 1/a_{n+1})q_{n-1} + q_{n-2},
\]
and the result follows by algebra and the equations for \(p_n, q_n\) and \(p_{n+1}, q_{n+1}.\) For part b, the claim is immediate if \(n = 1,\) and for \(n > 1,\)
\[
p_nq_{n-1} - p_{n-1}q_n = (a_n p_{n-1} + p_{n-2})q_{n-1} - (a_n q_{n-1} + q_{n-2})p_{n-1}
\]
\[
= -(q_{n-2} - p_{n-2}q_{n-1});
\]
the claim follows inductively. For part c,
\[
p_nq_{n-2} - p_{n-2}q_n = (a_n p_{n-1} + p_{n-2})q_{n-1} - (a_n q_{n-1} + q_{n-2})p_{n-1}
\]
\[
= -(p_{n-1}q_{n-2} + p_{n-2}q_{n-1}),
\]
and the claim follows using part b. Part d follows by induction. Part e follows using parts c and b. For part f, \(p_n, q_n\) are integers by induction; that they are relatively prime follows by part b.

If \(r\) is rational so is each \(r_i;\) write \(r_i = h_i/k_i\) in lowest terms. Then
\[
{h_{i+1}} \over k_{i+1} = {k_i} \over h_i - k_i a_i;
\]
the right side is in lowest terms, so \(h_{i+1} = k_1\) and \(k_{i+1} = h_i - k_i a_i.\) These are exactly the recursion equations for computing \(\text{gcd}(h_0, k_0)\) by the Euclidean algorithm, where \(a_i = [h_i/k_i].\) The algorithm terminates when \(k_{i+1} = 0;\) it always terminates since the \(k_i\) form a decreasing sequence of nonnegative integers. When it terminates, \(k_i = h_{i+1} = 1\) (or \(\text{gcd}(h_0, k_0)\) in general). Thus, the continued fraction expansion of a rational number is finite; the converse is clearly true.

If \(a_n\) is a regular infinite sequence it follows using lemma 31 that the limit of \([a_0, \ldots, a_n]\) exists (by part e, part b, and the fact that the \(q_n\) are increasing). We let \([a_0, \ldots,]\) denote the limit. It is easily seen that \([a_0, \ldots, a_n, [a_{n+1}, \ldots]] = [a_0, \ldots];\) indeed \(r_n = [a_n, \ldots].\) The continued fraction expansion of \(r\) is easily seen to be the unique regular sequence whose value is \(r,\) so there is a bijection between the real numbers and the regular sequences, with the rational numbers corresponding to the finite sequences.

If \(r\) is rational, with continued fraction \([a_0, \ldots, a_n],\) the sequence \([a_0, \ldots, a_n - 1, 1]\) also has value \(r.\) This is the only other sequence for \(r\) with integer \(a_i\) and \(a_i > 0\) for \(i > 0,\) since \([a_i, \ldots, a_n] \geq 1\) for \(i > 0,\) with equality iff \(n = i\) and \(a_i = 1.\)

We use \(r = [a_0, \ldots,]\) to denote that the finite or infinite sequence of \(a_n\) is the continued fraction expansion of \(r.\) If the sequence is finite, \(N\) denotes the highest index value. Given \(a_n, p_n\) and \(q_n\) are as in lemma 31.
LEMMA 32. Suppose $r = [a_0, \ldots]$; then $|q_n r - p_n|$ is decreasing, and for $n \leq N - 1$ $q_n r - p_n = \delta_n (-1)^n / q_{n+1}$, where $0 < \delta_n < 1$ if $n \leq N - 2$ and $\delta_{N-1} = 1$.

PROOF: Using lemma 31.a

$$r = [a_0, \ldots, a_n, r_{n+1}] = \frac{r_{n+1} p_n + p_{n-1}}{r_{n+1} q_n + q_{n-1}}, \quad 1 \leq n \leq N - 1,$$

and so

$$q_n r - p_n = \frac{q_n P_{n-1} - p_n q_{n-1}}{r_{n+1} q_n + q_{n-1}} = \frac{(-1)^n}{s_{n+1}}, 1 \leq n \leq N - 1$$

where $s_n = r_n q_{n-1} + q_{n-2}$ for $2 \leq n \leq N$. Letting $s_1 = r_1$, the claim is true for $n = 0$ also. Now, $a_n < r_n < a_n + 1$ for $n \leq N - 1$, whence $q_n a_{n+1} + q_{n-1} < q_n r_{n+1} + q_{n-1}$, or $q_{n+1} < s_{n+1}$, for $1 \leq n \leq N - 2$; this is also true for $n = 0$. Also, $q_n r_{n+1} + q_{n-1} < q_n a_{n+1} + q_{n} + q_{n-1}$, or $s_{n+1} < q_{n+1} + q_n$, for $1 \leq n \leq N - 2$; but $q_{n+1} + q_n \geq a_{n+2} q_{n+1} + q_n = q_{n+2}$, so $s_{n+1} < q_{n+2}$ for $1 \leq n \leq N - 2$, and this also is true for $n = 0$.

Thus, $1/q_{n+2} < |q_n r - p_n| < 1/q_{n+1}$ for $0 \leq n \leq N - 2$; $q_{n+2} r_n - p_{n+1} = (-1)^{N-1}/q_N$; and $q_n r - p_n = 0$. The $q_n r - p_n$ are thus decreasing in absolute value. The value $\delta_n$ equals $q_{n+1}/s_{n+1}$ for $0 \leq n \leq N - 1$, which is $< 1$.

A quadratic surd is a real number of the form $a + b\sqrt{c}$, where $a, b, c$ are rational. Equivalently, it is a real root of a quadratic equation with integer coefficients. A continued fraction is repeating if there is an $m$ and a $j > 0$ such that $a_{n+j} = a_n$ for $n \geq m$; equivalently, such that $r_{m+j} = r_m$. It is easy to see that if $r$ has a repeating continued fraction then $r$ is a quadratic surd. Indeed, $r_m = [a_m, \ldots, a_{m+j-1}, r_m]$, and the right side is of the form $(a r_m + b)/(c r_m + d)$ for integer $a, b, c, d$, yielding a quadratic equation for $r_m$, $r$ is of the same form, and since the quadratic surds are closed under multiplication and multiplicative inverse, $r$ is a quadratic surd. Conversely, a quadratic surd has a repeating continued fraction; see [HardWr] for a proof.

In particular, the real represented by the continued fraction where $a_n = 1$ for all $n$ is a quadratic surd. Letting $\phi$ denote its value, $\phi = 1 + 1/\phi$, whence $\phi = (\sqrt{5} + 1)/2$, or $1.618034$ to 7 significant figures. This number is called the golden ratio; it was so named by the Greek geometers. One verifies that $q_1 = q_0 = 1$, $q_n = q_{n-1} + q_{n-2}$ for $n \geq 2$, and $p_n = q_{n+1}$ (for readers familiar with the definition, note that the $q_n$ are the well-known Fibonacci numbers).

LEMMA 33. If $A > \sqrt{5}$ then $|\phi - p_n/q_n| < 1/(A q_n^2)$ has only finitely many solutions.

PROOF: Write $\phi = p_n/q_n + \delta_n/q_n^2$. Then $q_n \phi - q_n = \phi - q_n$, whence $\delta_n/q_n - \sqrt{5} q_n/2 = q_n/2 - p_n$, whence $\delta_n^2/q_n - \sqrt{5} \delta_n = -q_n^2 - p_n q_n + p_n^2$. If the absolute value of the left side ever decreases below 1 then for some positive integers $p$ and $q$, $-q^2 - p^2 + p^2 = 0$ a contradiction.

THEOREM 34.

a. $|r - p_n/q_n|$ is decreasing with $n$.

b. $|r - p_n/q_n| < 1/q_n^2$ for $n \leq N - 2$.

c. For $r$ irrational, $|r - p_n/q_n| < 1/\sqrt{5} q_n^2$ for infinitely many $n$.

d. $\sqrt{5}$ is the largest constant for which part c is true.

e. For $r$ irrational, if $|r - p/q| < 1/2q^2$ then $p/q$ is a convergent to $r$.

f. If $n > 1$, $0 < q \leq q_n$, and $p/q \neq p_n/q_n$ then $|r_n - p_n/q_n| < |r_q - p|$, and a fortiori $|r - p_n/q_n| < |r - p/q|$.

PROOF: Using lemma 32 and the fact that the $q_n$ are increasing, the values $|r - p_n/q_n|$ are decreasing; and $|r - p_n/q_n| < 1/q_n q_{n+1} < 1/q_n^2$. For part c, $|r - p_n/q_n| = 1/(q_n s_{n+1}) = 1/(A_n q_n^2)$ where $s_n$ is in the proof of lemma 32, and $A_n = r_{n+1} + q_{n-1}/q_n$. We derive a contradiction from supposing that $A_i \leq \sqrt{5}$ for
Then the steps in the continued fraction expansion of $u$ are relatively prime. $\mu, \nu$ are such that $|\mu q - \nu p| \leq 1$ and $\mu(rq - p) > |r|$. Let $\mu, \nu$ be such that $\begin{bmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} = \begin{bmatrix} p \\ q \end{bmatrix}$, whence $\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \pm \begin{bmatrix} q_{n-1} - p_{n-1} \\ -q_n \\ p_n \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix}$.

Then $\mu, \nu$ are integers, both nonzero, and since $q = \mu q_n + \nu q_{n-1} < q_n$ they have opposite signs. Using Lemma 31, $rq_n - p_n$ and $rq_{n-1} - p_{n-1}$ have opposite signs, whence $\mu(rq_n - p_n)$ and $\nu(rq_{n-1} - p_{n-1})$ have the same sign. Since their sum is $rq - p$, $|rq - p| > |rq_n - p_n|$ follows.

In the case $n = 1$ of part f, the case $r = k + 1/2$ for integral $k$ is exceptional; see [HardWr].

By the recursion for $[a_0, \ldots, a_n]$, this value as a function of $a_1, \ldots, a_n$ may be written as a quotient of polynomials $P_n(a_0, \ldots, a_n)/Q(a_0, \ldots, a_n)$, where $P_0(a_0) = a_0$, $Q_0(a_0) = 1$, and for $n > 0$

$$P_n(a_0, \ldots, a_n) = a_0 P_{n-1}(a_1, \ldots, a_n) + Q_{n-1}(a_1, \ldots, a_n)$$
$$Q_n(a_0, \ldots, a_n) = P_{n-1}(a_1, \ldots, a_n).$$

Writing $P_n$ for $P_n(a_0, \ldots, a_n)$, etc., one verifies by induction that $P_n Q_{n+1} = Q_n P_{n+1} = (-1)^{n+1}$. Indeed, write $P_n^+$ for $P_n(a_1, \ldots, a_{n+1})$, etc., so that $P_n = a_0 P_{n-1}^+ + Q_{n-1}^+$ and $Q_n = P_{n-1}^+$, expand all four factors to apply the induction hypothesis. In particular, $P_n$ and $Q_n$ are relatively prime.

Suppose $u_0 > u_1 > 0$ are integers; write the steps of Euclid’s algorithm as $u_n = a_n u_{n+1} + u_{n+2}$. Then the steps in the continued fraction expansion of $u_0/u_1$ are $u_n/u_{n+1} = a_n + u_{n+2}/u_{n+1}$. It follows that $u_0/u_1 = P_n/Q_n$, where $[a_0, \ldots, a_n]$ is the continued fraction expansion of $u_0/u_1$. In fact, since $P_n$ and $Q_n$ are relatively prime, $u = dP_n(a_0, \ldots, a_n)$ and $v = dQ_n(a_0, \ldots, a_n)$ where $a_i$ are the integers and $d = \gcd(u, v)$.

The coefficients of $P_n$ are easily seen to be nonnegative. It follows that the smallest value $u_0$ may have in an execution of Euclid’s algorithm with a given value of $n$ is $P_n(1, \ldots, 1, 2)$. This in turn may readily be verified to be $F_{n+2}$, where $F_n$, the $n$th Fibonacci number, is defined by the recursion $F_0 = 0, F_1 = 1$, and for $n \geq 2$, $F_n = F_{n-1} + F_{n-2}$.

Let $\phi = (1 - \sqrt{5})/2$; $\phi$ and $\hat{\phi}$ are the roots of $x^2 - x - 1$. It is readily verified by induction that $F_n = (\phi^n - \hat{\phi}^n)/\sqrt{5}$. From this, $n \leq \log_2(u_0)/\log_2(\phi)$.

Exercises.

1. Show that the measurable sets form a ring. Hint: Suppose $E$ and $F$ are measurable and $A$ is arbitrary. Then

$$\mu^*(A) = \mu^*(A \cap E \cap F) + \mu^*(A \cap E^c \cap F) + \mu^*(A \cap E \cap F^c) + \mu^*(A \cap E^c \cap F^c).$$

Considering this with $A = B \cap (E \cup F)$, the last term drops out. Considering it with $A = B$ and using the latter fact the measurability of $E \cup F$ follows. For $E \cap F$ proceed similarly with $A = B \cap (E - F)^c$ and then $A = B$. 

317
2. Show that $\mu^*(E \cup F) = \mu^*(E) + \mu^*(F)$ for disjoint measurable sets $E$ and $F$. Hint: $(E_1 \cup E_2) \cap E_1 = E_1$ and $(E_1 \cup E_2) \cap E_2^c = E_2$.

3. Show that the measurable sets form a $\sigma$-ring. Hint: Immediately from the definition, if $S \subseteq T$ then $\mu^*(S) \leq \mu^*(T)$. Let $E = \cup E_n$ be a union of a disjoint family of measurable sets. Let $A$ be any set, and let $F_n = \cup_{i=0}^n E_i$. Then

$$\mu^*(A) = \mu^*(A \cap F_n) + \mu^*(A \cap F_n^c) \geq \mu^*(A \cap F_n) + \mu^*(A \cap E^c) = \sum_{i=0}^n \mu^*(A \cap E_i) + \mu^*(A \cap E^c).$$

Since this holds for all $n$ we have

$$\mu^*(A) \geq \sum_{i=0}^\infty \mu^*(A \cap E_i) + \mu^*(A \cap E^c) \geq \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

by countable subadditivity. The reverse inequality follows by countable subadditivity.

4. Show that if $\{E_i\}$ is a disjoint collection of measurable sets and $E = \cup E_i$ then $\mu^*(E) = \sum \mu^*(E_i)$. Hint: Define $F_n$ as in the previous problem; then $\mu^*(E) \geq \mu^*(F_n) = \sum_{i=0}^n \mu^*(E_i)$. It follows that $\mu^*(E) \geq \sum_{i=0}^\infty \mu^*(E_i)$. The reverse inequality follows by countable subadditivity.

5. Show that the cells are measurable. Hint: If $A$ is an arbitrary set and $\epsilon > 0$ there is a disjoint collection $\{C_i\}$ of cells with $\sum \nu(C_i) \leq \mu^*(A) + \epsilon$. If $C$ is a cell then $\sum \nu(C_i) = \nu(C_i \cap C) + \nu(C_i - C)$. It follows that $\mu^*(A) \geq \mu^*(A \cap C) + \mu^*(A \cap C^c)$. The reverse inequality follows by countable subadditivity.

6. Show that a set of outer measure 0 is measurable. Hint: If $A$ is any set and $E$ has outer measure 0 then $\mu^*(A) \geq \mu^*(A \cap E) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$.

7. Prove without using theorem 9 that for real numbers $r_i$, $1 \leq i \leq n$, there are integers $p_i$ and $q \geq 1$ with $|r_i - p_i/q| < 1/2^{1+1/n}$. Hint: Divide the cell $0 \leq x_i < 1$ into subcells of edge $1/N$ where $N$ is an integer. Consider the $N^n + 1$ points $w_i$ where $w_i$ is the fractional part of $lr_i$; two of these must lie in the same subcell. It follows that there are $p_i$ and $q$ with $q \leq N^n$ and $|qr_i - p_i| < 1/N$.

8. Prove the formula for the volume of a unit ball given in section 3. Hint: Let $I_k = \int_{-1}^{+1} (1 - x^2)^k \, dx$, where $k$ is an integer or half-integer. Prove by integration by parts that $I_k = (2k) / (2k + 1) I_{k-1}$.

9. Show that a subset of a root system is positive iff it lies strictly on one side of some hyperplane. Hint: If $S$ lies in a positive set then $Cone(S)$ contains no line, so the origin is a face. For the converse, take a hyperplane intersecting $Cone(S)$ in 0. Take an orthonormal basis $\{v_i\}$ containing the normal toward $Cone(S)$. Say that $x$ is positive iff the first nonzero coefficient in $x = \sum c_i v_i$ is positive.

1. Topological structures. The interplay between algebra and topology underwent a series of dramatic advances beginning in the 1960's, resulting in perspectives and methods which have found uses throughout mathematics and other sciences, and contributing to major discoveries. This interplay can be seen in algebraic structures on topological spaces, families of algebraic structures defined on a topological space (either its points or its open sets), and functors from Top to algebraic categories. The latter is the subject matter of algebraic topology, which had its beginnings circa 1900, and has been modernized under the impact of category theory. This chapter includes some algebraic topology in section 14, but covers other topics as well, hence its title.

To begin with some categories are defined, which provide a convenient setting for general definitions. These have been a more specialized topic, with the generality they provide usually being achieved by less concrete methods; but the simplicity of the concrete setting makes it attractive. If \( L \) is a first order language, the category \( \text{TopS}_L \) has as objects the topological spaces which are also objects of \( \text{Struct}_L \), and where the functions are all continuous (the structure and topology are said to be compatible). The morphisms are the \( L \)-morphisms.

\( \text{TopS}_L \) is the full subcategory of \( \text{TopS}_L \times \text{Struct}_L \) where the objects are the models of \( T \).

\( \text{TopS}_L \) must be generalized slightly. The action of a structure \( A \) on a structure \( X \) can be described as a multisorted structure, but for fixed \( A \) acting on the structures of a category this results in annoying complications which can be avoided. The action may be specified as a function symbol for each \( a \in A \); in many important cases axioms for the behavior of the action can be given as equations. When topologies are present, \( A \) can be a topological structure, and the action is required to be continuous from \( A \times X \) to \( X \). Let \( \text{TopS}_{L,A} \) to be the resulting category. Names are still used for the elements of \( A \), and in addition the topology is given. The category \( \text{TopM}_{T,A} \) is defined similarly.

The most important examples of categories of topological structures are topological groups, and topological vector spaces over some field equipped with a topology. For the former, the language is assumed to contain a symbol for inverse. The latter is a category of type \( \text{TopM}_{T,A} \) where \( A \) is the field. Also, the paradigm is stretched for a field with a topology, in that the multiplicative inverse is a partial function rather than a function; it must be continuous. Further categories of interest include some subcategories of the topological vector spaces.

In theorem 3 below parts a-d show how the forgetful functor from \( \text{TopS}_L \) to \( \text{Struct}_L \) determines various constructions. This is not the case for the coproduct. The underlying structure is the coproduct structure, but it must be given a topology specific to the new category.

If \( t \) is a term over \( L \) the function it determines in any structure \( X \) of \( \text{TopS}_L \) is continuous, a fact which will be used below. To see this, suppose that the distinct variables from left to right are \( x_1, \ldots, x_r \). If \( t \) is \( f(t_1, \ldots, t_n) \), inductively each \( t_i \) may be considered as determining a continuous function from \( X^r \) to \( X \). A continuous function from \( X^r \) to \( X^n \) is induced, and \( t \) determines the composition of \( f \) with this.

Suppose again that \( t \) is a term with variables \( x_1, \ldots, x_r \). Given objects \( X_i \) in \( \text{TopS}_L \), let \( H \) denote the set of closed terms \( t \) of the form \( t(c_1, \ldots, c_r) \) where \( c_i \in X_i \), considered as a structure in the usual way. Suppose \( U_i \) is an open set of \( X_i \), and let \( U_{i_1, U_{i_2}, \ldots, U_{i_r}} \) denote the closed terms where \( c_i \in U_i \). These sets may be taken as comprising the subbasic open sets of a topology on \( H \). For the remainder of the section \( H \) is supposed to be equipped with this topology.

**Lemma 1.** If \( f \) is an \( n \)-ary function of \( L \) then \( f \) is a continuous function on \( H \).

**Proof:** A closed term \( f(t_1, \ldots, t_n) \) is in a subbasic open set iff the term \( t \) has the closed term as an instance, and the open sets contain the constants. An (subbasic) open set for \( t_k \) is obtained by taking \( t_k \) as the term, and the open sets for its constants as for \( t \).

319
Lemma 2. The map \( X_i \mapsto H \) mapping \( c \) to itself is continuous.

Proof: The inverse image of \( U_{c,W} \) is \( W \).

Theorem 3.

a. Given \( X \) in \( \text{TopS}_L \) and a substructure \( W \subseteq X \), the subspace topology is compatible with the substructure.

b. Given objects \( X_i \) in \( \text{TopS}_L \), the product topology is compatible with the product structure. The product topology is that induced by the projections. The product structure equipped with the product topology is the product in \( \text{TopS}_L \).

c. The limit of a diagram in \( \text{TopS}_L \) is the limit in \( \text{Struct}_L \), equipped with the topology induced by the limit cone.

d. Given \( X \) in \( \text{TopS}_L \) and a congruence relation \( \equiv \) on \( X \), the quotient topology is compatible with the quotient structure.

e. With \( H \) as above, let \( \equiv \) be the equivalence relation where two closed terms are equal iff all their constants are from the same structure and they are equal in the structure. This is a congruence relation, and \( H/\equiv \), equipped with the quotient topology is the coproduct in \( \text{TopS}_L \).

f. Colimits exist in \( \text{TopS}_L \).

g. If there is a topological space of actions the action on the limit or colimit is continuous.

h. The preceding facts hold for \( \text{TopM}_T \) for \( T \) a set of equations, provided in part e \( \equiv \) is modified to include the equivalences arising from \( T \).

Proof: For part a, suppose \( V \subseteq W \) is open, and \( n \) is an \( n \)-ary function; let \( f_r \) denote the restriction to \( W^n \). Then \( w \in f_r^{-1}[V] \) iff \( w \in W^n \) and \( w \in f^{-1}[V] \). That is, \( f_r^{-1}[V] = W^n \cap f^{-1}[V] \), so \( f_r^{-1}[V] \) is open in \( W^n \). For part b, let \( U \) be the subbasic open set which is \( U_j \) in component \( j \) and \( X_i \) for \( i \neq j \); suppose \( f \) is \( n \)-ary. Then \( f(\langle x_{i1}, \ldots, x_{in} \rangle) \in U \) iff \( f(x_{ij}, \ldots, x_{jn}) \in U_j \). So \( f^{-1}[U] \) is in fact the subbasic open set which is \( f^{-1}[U_j] \) in component \( j \) and \( X_i^t \) for \( i \neq j \), and in particular is open. The product topology is that induced by the projections since this is true of the topological spaces. Thus, given a cone from the \( X_i \) to any \( Y \in \text{TopS}_L \), the induced map from \( Y \) to \( \times_i X_i \) (the product in \( \text{Struct}_L \)) is continuous. Part c follows from parts a and b by usual arguments, noting that the equalizer in \( \text{Set} \) is the equalizer in \( \text{TopS}_L \). For part d, for \( x \in X \) let \( \bar{x} \) denote the image under the canonical epimorphism, and extend the notation to subsets of \( X \), to the functions of the structure. Suppose \( \bar{f}(\bar{x}_1, \ldots, \bar{x}_n) \in \bar{V} \) where \( \bar{V} \) is open in \( \bar{X} \). Then \( f(x_1, \ldots, x_n) \in V \) (where \( V = \cup \bar{V} \) is the saturated open set whose image is \( \bar{V} \)). Thus there are open sets \( U_1, \ldots, U_n \) such that if \( x_k \in U_k \) for \( 1 \leq k \leq n \) then \( f(x_1, \ldots, x_n) \in V \). Since \( V \) is saturated and \( \equiv \) is a congruence relation, \( U_k \) may be replaced by its saturation. Then if \( \bar{x}_k \in \bar{U}_k \) for \( 1 \leq k \leq n \) then \( \bar{f}(\bar{x}_1, \ldots, \bar{x}_n) \in \bar{V} \). For part e, that \( H/\equiv \) is the coproduct in \( \text{Struct}_L \) follows by arguments as in chapter 13. Given \( K \in \text{TopS}_L \) and \( \text{TopS}_L \) morphisms \( \nu_i : X_i \mapsto K \), there is a unique \( \text{Struct}_L \) morphism \( \phi : H \mapsto K \) such that \( \phi(c) = \nu_i(c) \) for each \( c \in X_i \). Suppose \( V \subseteq K \) is open and \( \phi(t) \in V \). Write \( t \) as \( t(c_1, \ldots, c_r) \) where \( c_i \in X_{t_i} \). Let \( t_K \) be the function on \( K \) interpreting \( t \). Since \( t_K \) is continuous there are open sets \( U_{t_i} \subseteq X_{t_i} \) such that for \( \langle c'_1, \ldots, c'_r \rangle \in U_{t_1} \times \cdots \times U_r \) \( t_K(\nu_i(c'_1), \ldots, \nu_i(c'_r)) \in V \). Thus, \( \phi(U_{t_{U_{t_1} \cdots U_r}}) \subseteq V \), and \( \phi \) has been shown to be continuous. If \( d = f(c_1, \ldots, c_n) \) in \( X \), then \( \phi(d) = \phi(f(c_1, \ldots, c_n)) \), so \( \phi \) respects \( \equiv \). Thus, \( \phi \) factors through the canonical epimorphism in \( \text{Top} \), and part e is proved. For part f, the colimit of a diagram in \( \text{TopS}_L \) is obtained from the coproduct and the coequalizer in the usual way. For part g, for limits let \( \alpha_i(\alpha) \) denote the action of \( A \) on the \( X_i \) (the limit \( Y \)). The map \( \alpha_i \circ (i_A \times f) \) is continuous, and since \( f_i \) preserves \( \alpha \) equals \( f_i \alpha \). By part c and theorem 17.1 \( \alpha \) is continuous. For colimits, one verifies directly that \( A \) acts continuously on \( H \), and the rest of the claim follows by general arguments. Part h holds for limits because substructures and products of models are models; it follows for colimits by arguments as in chapter 13.
There is another approach to defining topological groups. Let $C$ be a category with finite products (in particular a terminal object $T$). A group in $C$ is defined to be an object $G$ and arrows $m : G \times G \mapsto G$, $e : T \mapsto G$, and $i : G \mapsto G$, such that the following diagrams are commutative.

![Diagram]

In the first diagram canonical equivalences between $G \times G \times G$, and $G \times (G \times G)$ and $(G \times G) \times G$, are ignored. In the third diagram $(\iota, \iota)$ is the map induced by $\iota$ and $i$. It is worth noting that in the second diagram, $\pi_1$ and $\pi_2$ are isomorphisms (exercise 1).

It is easy to see that a group in $\text{Set}$ is exactly a group, and a group in $\text{Top}$ is exactly a topological group, according to the “concrete” definitions. Theorem 1 shows the advantage of concrete definitions; on the other hand the “abstract” category-theoretic definition has advantages. For example, a group in the category of manifolds is a “Lie group”; these are of great importance in modern algebra and an introduction is given in chapter 27. Lie groups have a standard concrete definition. There is a category $\text{TopH}_*$, defined below, where a concrete definition of a group object (H-group) is more specialized.

The abstract definition has another advantage, that it can be dualized; the product is replaced by the coproduct, the terminal object replaced by an initial object, and the arrows reversed. $G$ is called a cogroup, $m$ the comultiplication, etc. There is an important example of cogroups in $\text{TopH}_*$, which will be discussed below.

2. Topological groups. In a topological group, a left or right translation is a continuous, and since its inverse is also a left or right translation, it is a homeomorphism. Similarly $x \mapsto x^{-1}$ is a homeomorphism.

If $\{U\}$ is a neighborhood base at the identity $e$ then its translation to $x$ is a neighborhood base at $x$. Since the union of the neighborhood bases over all $x$ is a base for the topology, the topology is determined by a neighborhood base at $e$. A function between topological groups is continuous iff it is continuous at $e$; to see this, note that if $f(x) = y$ then $f^{-1}[y^{-1}V] = x^{-1}f^{-1}[V]$.

A group equipped with a topology is readily verified to be a topological group if $(x, y) \mapsto xy^{-1}$ is continuous. This is often given as the definition. In particular for any open neighborhood $U$ of the identity there is an open neighborhood $V$ of the identity with $VV^{-1} \subseteq U$. In fact, if $V_1V_2 \subseteq U$ let $W = V_1 \cap V_2 \cap V_1^{-1} \cap V_2^{-1}$; then $W^2 \subseteq V$ and $W = W^{-1}$.

The requirement that the inverse be continuous is necessary; see [Robert], exercise 1.20.

If $G$ is a group, equipping it with either the discrete or the indiscrete topology yields a topological group. In particular a topological group need be neither connected nor $T_0$. If $G$ is a $T_0$ topological group

321
then it is $T_2$. Indeed, suppose $x \in U$ and $y \notin U$. Choose $V$ with $VV^{-1} \subseteq x^{-1}U$. If $xV \cap yV$ is nonempty, say $xv_1 = yv_2$, then $v_1v_2^{-1} = x^{-1}y$, which is a contradiction since it implies $x^{-1}y \in x^{-1}U$, or $y \in U$.

If $H \subseteq G$ is a subgroup then by facts in section 1 it is a topological group when given the subspace topology. $H^{cl}$ is also a subgroup: for if $xy \in U$ where $U \cap H = \emptyset$, choose $V_x$ and $V_y$ with $x \in V_x$, $y \in V_y$, and $V_xV_y \subseteq U$. If both $V_x$ and $V_y$ intersect $H$ then $U$ does, so one of $x, y$ is not in $H^{cl}$.

If $H$ contains an open set $U$ then $gH = \cup_{h \in H}ghU$, and so the cosets are all clopen. If $H$ contains a set $U$ which is open in $H^{cl}$ then $H$ is closed in $H^{cl}$, hence $H = H^{cl}$.

If $N$ is a normal subgroup then $N^{cl}$ is also a normal subgroup. Indeed, suppose $x^{-1}yx \in U$ where $U \cap N = \emptyset$; then $y \in xUx^{-1}$ and $xUx^{-1}$ is open. Further

$$xUx^{-1} \cap N = xUx^{-1} \cap xNx^{-1} = x(U \cap N)x^{-1} = \emptyset.$$

If $U$ is a component of $G$ then any translate of $U$ is a component. If $U_e$ is the component of the identity and $x \in U_e$ then $xU_e = U_e$, because $x \in U_e \cap xU_e$ and distinct components are disjoint. Thus, $U_e$ is a subgroup, in fact by similar arguments a normal subgroup.

If $H$ is a subgroup of a group $G$, $G/H$ is commonly used to denote the set of left cosets of $H$. We use this notation for the next few paragraphs, but generally reserve it to denote the quotient group when $H$ is normal. Recall from chapter 5 that $G$ acts transitively on $G/H$ by left multiplication (right cosets and right multiplication could equally well be used). Further any transitive action of $G$ on a set $X$ is isomorphic to the action of $G$ on $G/G_x$ where $x \in X$ and $G_x$ is the stabilizer of $x$.

If $G$ is a topological group and $H$ is a subgroup then $G/H$ is presumed to be equipped with the quotient topology. The projection map $\eta$ maps $g$ to $gH$ and is continuous. It is also open, because the saturation $\cup\{gH : g \in U\}$ of $U$ is $UH$, a union of translates of $U$.

The action of $G$ on $G/H$ is continuous. To see this, note that $G \times (G/H)$ has the topology coinduced by $\iota_G \times \eta$, and the map $(g_1, g_2) \mapsto g_1g_2H$, being the composition of multiplication and $\eta$, is continuous. The action of any fixed element $g \in G$ on $G/H$ is a homeomorphism (the action of $g^{-1}$ being the inverse map).

If $N$ is a normal subgroup then $G/N$ is a topological group and $\eta$ is a morphism of topological groups.

Given an action of $G$ on a topological space $X$, and $x \in X$, the map $gG_x \mapsto gx$ is a continuous bijection, because $G/G_x$ has the coinduced topology and $g \mapsto gx$ is continuous. For the map to be a homeomorphism, $X$ must have the coinduced topology coinduced by $g \mapsto gx$; this is easily seen to be the case iff the map is open. Under these circumstances $X$ is called a homogeneous space for $G$.

If $H$ is a closed subgroup of $G$ then the homogeneous space $G/H$ is Hausdorff. It suffices to show that it is $T_0$. Given two cosets $g_1H$ and $g_2H$, the open set $\cup_{g \neq g_1, g_2}gH$ contains the second but not the first. Conversely if $G/H$ is Hausdorff then $\{H\}$ is a closed singleton of $G/H$, so its inverse image under $\eta$, namely $H$, is a closed subgroup.

A group homomorphism $f : G \rightarrow H$ is a morphism of topological groups iff $f^{-1}[U]$ is open for any open neighborhood $U$ of the identity. If $f : G \rightarrow H$ is a morphism of topological groups the usual facts hold, by group theory and remarks in section 1 (that is, the usual functions are continuous). Note, however, that the canonical isomorphism from $G/\text{Ker}(f)$ to $f[G]$ is a continuous bijection, and a homeomorphism iff $f[G]$ has the coinduced topology, iff $f$ is open.

3. **Topological vector spaces.** Topological vector spaces may be defined over any field with a topology. As additional restrictions are imposed more facts may be proved. Frequently the field is the real or complex numbers, although a field whose topology is given by an absolute value will often suffice. The vector space might be required to be Hausdorff (in which case the topology on the field must be; see exercise 2).
We let $F$-TVS denote the category of topological vector spaces over a field $F$. Theorem 1 and the remarks following it pertain to $F$-TVS. A finite product is a biproduct; this is easily seen by verifying directly that the product with the usual injections is a coproduct.

A translation is a homeomorphism, because this is true of the additive group. Multiplication by a nonzero scalar is also a homeomorphism, indeed an isomorphism in the category, again because it is continuous by definition and its inverse map, being a scalar multiplication, is also.

If $X$ is Hausdorff, $Y$ is a subspace, and $U$ is an open subset of $Y$ then $Y = X$. We may assume by translation that $0 \in U$. If $U = \{0\}$ then $X = \{0\}$, so suppose $x \neq 0$ and $x \in U$. Since $x \mapsto \alpha x$ is continuous there is an open set $W \subseteq F$ such that if $\alpha \in W$ then $\alpha x \in U$. In particular (exercise 2) there is some $\alpha \neq 0$ such that $\alpha x \in U$. Since $Y$ is a subspace, $x \in Y$.

The closure of a subspace is again a subspace. It is a subgroup of the additive group by section 2, and a similar argument shows that it is closed under scalar multiplication.

By section 2 a linear transformation between topological vector spaces is continuous iff it is continuous at 0.

Suppose $X$ is, as a vector space, the internal direct sum of subspaces $Y_1$ and $Y_2$. The map $(y_1,y_2) \mapsto y_1 + y_2$ from $Y_1 \times Y_2$ to $X$ is a continuous vector space isomorphism. The inverse map $y_1 + y_2 \mapsto (y_1, y_2)$ is not necessarily continuous (for a counterexample, see [Dieudonné], problem 2 of section 6.5). By the universality of the product it is continuous iff the maps $y_1 + y_2 \mapsto y_1$ and $y_1 + y_2 \mapsto y_2$ are both continuous. If one of these maps is continuous then both are, since for example $y_2 = (y_1 + y_2) - y_1$.

In the vector spaces, $\text{Hom}(X,Y)$ is a vector space when equipped with the pointwise operations. This remains true in $F$-TVS; that $f + g$ and $af$ are continuous follows since $+$ and scalar multiplication are. As for Top, $\text{Hom}(X,Y)$ can have various topologies defined on it as well, for example the pointwise topology, which has as a subbase the sets $\{f : f(x) \in V\}$ where $x \in X$ and $V$ is an open subset of $Y$. One readily verifies that this is the topology induced by the maps $f \mapsto f(x)$. As will be seen below, in the case of normed linear spaces there is a natural topology defined on $\text{Hom}(X,Y)$; we omit any further discussion of the general case.

For the rest of the section suppose that the topology on $F$ is that of an absolute value $|x|$; we follow [Wilde]. A subset $S \subseteq F$ is said to be bounded if $|x| < r$ for some real number $r$. Say that a nonempty subset $B$ of a topological vector space $X$ over $F$ is balanced if $aB \subseteq B$ whenever $|a| \leq 1$. A balanced subset of $F$ itself is either all of $F$, or is bounded; for if $y \notin B$ and $x \in B$ then $|x| < |y|$.

**Lemma 4.** If $V$ is an open neighborhood of 0 then there is a balanced open neighborhood $U$ of 0 with $U \subseteq V$.

**Proof:** By the continuity of scalar multiplication there is a real number $\delta$, and an open neighborhood $W$ of 0, such that $aW \subseteq V$ for $|a| < \delta$. Let $U = \cup_{|a| < \delta} aW$. If $x \in U$, say $x \in aW$, and $|b| \leq 1$, then $bx \in baW$.

**Lemma 5.** Suppose $f : X \to F$ is a linear functional whose kernel is not all of $X$; then the following are equivalent.

a. $f$ is continuous.

b. $\text{Ker}(f)$ is closed.

c. For some open neighborhood $U$ of 0, $f[U]$ is bounded.

**Proof:** Since $\{0\}$ is closed, if $f$ is continuous then $\text{Ker}(f)$ is closed. If $\text{Ker}(f)$ is closed, let $x$ be some element not in it. Since $\text{Ker}(f)$ is closed, there is an open neighborhood $U$ of 0 such that $x + U$ is disjoint from $\text{Ker}(f)$; hence there is a balanced such $U$. Then $f[U]$ is a balanced. Further, $-f(x) \notin f[W]$, else $0 \in f[x + W]$. Thus, $f[W]$ is bounded, and c follows from b. Suppose $|f(u)| < r$ for $u \in U$, and let $\epsilon > 0$ be a real number. Then $|f(u)| < \epsilon$ for $u \in (\epsilon/r)U$, proving that $f$ is continuous.

323
The next theorem requires the definition of a norm on a linear space, so it is given here rather than section 5. Suppose $F$ is a field equipped with an absolute value. By a norm on a vector space $X$ over $F$ is meant a function $|x|$ from $X$ to $\mathbb{R}^+$ such that $|x| = 0$ iff $x = 0$, $|ax| = |a||x|$, and $|x + y| \leq |x| + |y|$. The function $|x - y|$ is readily verified to be a metric on $X$, and the topology of this metric is called the topology of the norm.

**Theorem 6.** Suppose $F$ is a field equipped with an absolute value, which is locally compact in the metric topology. Then an $n$-dimensional Hausdorff topological space $X$ over $F$ is isomorphic to $F^n$ (equipped, say, with the norm $\max(|a_1|, \ldots, |a_n|))$.

**Proof:** Suppose $x_1, \ldots, x_n$ is a basis for $X$. The map $\phi : F^n \mapsto X$ mapping $\langle a_1, \ldots, a_n \rangle$ to $a_1 x_1 + \cdots + a_n x_n$ is clearly a continuous isomorphism of vector spaces. Let $S = \{ a \in F^n : |a| = 1 \}$. By local compactness of $F$ $S$ is compact, so $\phi[S]$ is compact, so by the assumption that $X$ is Hausdorff $\phi[S]$ is closed. Since $0 \notin \phi[S]$, $\phi[S]^c$ is an open neighborhood of 0, so contains a balanced open neighborhood $V$. $\phi^{-1}[V]$ is therefore a balanced open neighborhood of 0 in $F^n$, and is disjoint from $S$. For $a \in \phi^{-1}[V]$, if $|a| \geq 1$ then $a/|a|$ is in $S$, and also in $\phi^{-1}[V]$ since $\phi^{-1}[V]$ is balanced, a contradiction. It follows that for each projection map $\pi_i$ on $F^n$, $\pi_i \phi^{-1}$ is bounded. By lemma 5 each $\pi_i \phi^{-1}$ is continuous. Hence $\phi^{-1}$ is continuous.

The requirement that $F$ be locally compact is satisfied by $\mathcal{R}$ and $C$, and also the $p$-adic number fields $\mathbb{Q}_p$ defined in section 20.12. We will not prove the latter; see the references given in section 20.12.

The norm used in the theorem is called the “max” or “infinity” norm. It will be seen below that the choice is immaterial; other suitable norms include the “taxicab” or “$1$” norm $|a_1| + \cdots + |a_n|$, and the “Euclidean” or “$2$” norm $\sqrt{|a_1|^2, \ldots, |a_n|^2}$. The proof of the theorem is unchanged with either of these.

4. Additional facts about metric spaces. We introduce the notation $B^r_{xy}$ for the closed ball $\{ y : d(x, y) \leq r \}$. This need not equal $B^1_{xy}$ (consider the discrete metric $d(x, y) = 1$ if $y \neq x$). The diameter of a subset $S \subseteq X$ of a metric space is defined to be $\sup\{ d(p, q) : p, q \in S \}$; this may be $\infty$, and the set is bounded iff it is finite. The diameter of $S^d$ equals the diameter of $S$. The diameter $d$ of a ball $B_{xy}$ or $B^-_{xy}$ satisfies $d \leq 2r$. We use $\text{diam}(S)$ to denote the diameter.

**Lemma 7.** Suppose $X$ is a metric space; the following are equivalent.

a. $X$ is complete.

b. $S_n$ is a descending chain of closed sets with $\text{diam}(S_n)$ converging to 0 then $\bigcap_n S_n$ is nonempty.

c. $S_n$ is a descending chain of closed balls with $\text{diam}(S_n)$ converging to 0 then $\bigcap_n S_n$ is nonempty.

**Proof:** Suppose $X$ is complete and $S_n$ is as in b; choose $x_n \in S_n$. Then $x_n$ is a Cauchy sequence, so has a limit $x$. If $x \notin S_k$ then $d(x, S_k) = r$ for some $r > 0$; but then $d(x, x_n) \geq r$ for $n \geq k$, contradicting the fact that $x$ is the limit of $x_n$. Thus, $x \in S_k$. This shows that $a$ implies $b$; $b$ implies $c$ trivially. Suppose $c$ holds, and let $x_n$ be a Cauchy sequence. Let $n_0$ and $r_0$ be such that for $n \geq n_0$, $x_n \in B_{r_0, r_0/2}$. Given $n_k$, let $n_{k+1}$ be such that if $k, l \geq n_{k+1}$ then $d(x_k, x_l) < r_0/2^{k+2}$. One verifies that if $n \geq n_k$ then $x_n \in B_{x_{n_k}, r_0/2^{k+1}}$, and $B^-_{x_{n_k}, r_0/2^{k+1}} \subseteq B^-_{x_{n_k}, r_0/2^k}$. The intersection of the $B^-_{x_{n_k}, r_0/2^k}$ is readily verified to consist of a single point, which is the limit of the $x_n$.

A metric space $X$ is said to be ultrametric if the ultrametric inequality $d(x, z) \leq \max(d(x, y), d(y, z))$ holds. By the ultrametric inequality, for any $r$ the relation $d(x, y) \leq r$ is an equivalence relation, and also $d(x, y) < r$. The following facts are readily verified.

- If $y \in B_{xy}$ then $B_{xy} = B_{yr}$, and similarly for $B^-_{xy}$.
- The open or closed balls of radius $r$ form a partition of $X$.
- Given two open or closed balls, if they intersect then one is a subset of the other.
- Open balls are closed and $X$ has a base of clopen sets.

324
It is also true that closed balls are open; for if \( y \in B_{x,\epsilon}^- \) then \( B_{y,\epsilon} \subseteq B_{y,\epsilon}^- \subseteq B_{x,\epsilon}^- \).

Continuing to suppose \( X \) ultrametric, suppose \( B \) is a ball of diameter \( d \), and suppose \( x \in B \). If \( y \in B \) then \( d(x,y) \leq d \), so \( y \in B_{x,d}^- \); that is, \( B \subseteq B_{x,d}^- \). If \( y \in B_{x,d} \) then \( d(x,y) = e \) for some \( e < d \), so \( B_{x,e} \subseteq B \) since \( B \subseteq B_{x,e} \) is impossible; that is, \( B_{x,d} \subseteq B \).

A metric space is said to be spherically complete if a descending chain of closed balls has nonempty intersection (the definition may be given in general, but it is mostly used in ultrametric spaces). By lemma 7 a spherically complete metric space is complete.

A locally compact normed linear space over a field with a nontrivial absolute value is spherically complete. Let \( S_n \) be a descending chain of closed balls, and choose \( x_n \in S_n \). Take a ball around \( x_0 \) which has compact closure, and multiply it by a scalar to enclose the entire sequence of \( x_n \). By theorem 17.15 the sequence has a limit point. As in the proof of b from a in lemma 7, this is in every \( S_k \).

Suppose \( X \) is ultrametric and spherically complete, and \( S \) is a set of closed balls such that for \( B_1, B_2 \in S \), \( B_1 \cap B_2 \neq \emptyset \). By the above, \( S \) is a chain. Let \( d = \inf \{ \text{diam}(B) : B \in S \} \). Choose a countable descending chain \( B_n \) from \( S \), with \( \text{diam}(B_n) \) converging to \( d \). This set is nonempty; let \( x \) be an element. Then \( B_{x,d} \subseteq B \) for every \( B \in S \), and \( \cap S \) is nonempty.

**Theorem 8**. Suppose \( X \) is a complete metric space. Suppose \( D_n \) for \( n \in \mathcal{N} \) is a dense open set; then \( \cap_n D_n \) is dense.

**Proof**: Let \( U \) be an open set. Define a sequence \( U_n \) of nonempty open sets inductively; let \( U_0 = U \). \( U_n \) and \( D_n \) are open and \( U_n \cap D_n \) is nonempty. Since \( X \) is regular, there is a nonempty open set \( V \) such that \( V^\text{cl} \subseteq U_n \cap D_n \). Let \( U_{n+1} = \text{a nonempty open subset of } V \text{ of diameter at most } 1/(n+1) \). By construction \( \cap U_{n+1}^\text{cl} \subseteq U \cap \cap_n D_n \). By lemma 7 \( \cap U_{n+1}^\text{cl} \) is nonempty, proving the theorem.

A topological space with the property of the theorem is called a Baire space. Note that a subset is dense iff its complement has empty interior. A subset is called nowhere dense if its closure has empty interior. Suppose \( E_n \) for \( n \in \mathcal{N} \) is a closed set with empty interior; then \( \cup_n E_n \) has empty interior. A meager set has empty interior.

The theorem shows that a topological space which can be equipped with a metric under which it is complete is a Baire space. A locally compact space is also a Baire space; the proof of the theorem may be modified as follows. \( U_1^\text{cl} \) may be assumed to be compact, and \( \cap U_{n+1}^\text{cl} \) is nonempty since \( \{ U_{n+1}^\text{cl} \} \) has the finite intersection property.

If \( f : X \rightarrow Y \) is a function between metric spaces, and \( x_0 \in X \), \( y_0 \in Y \) is said to be the limit of \( f \) as \( x \) approaches \( x_0 \) (written \( \lim_{x \to x_0} f(x) = y_0 \)) if for all \( \epsilon > 0 \) there is a \( \delta > 0 \) such that if \( d(x,x_0) < \delta \) and \( x \neq x_0 \) then \( d(f(x),y_0) < \epsilon \). We say that \( f \) approaches \( y_0 \) as \( x \) approaches \( x_0 \). No matter how \( x_0 \) is approached (e.g., for any sequence \( x_n, n \geq 1 \), converging to \( x_0 \) with no \( x_n \) equaling \( x_0 \) \( f(x) \) approaches \( y_0 \) (e.g., \( f(x_n) \) does).

Observe that \( f \) is continuous at \( x_0 \) iff the limit is \( f(x_0) \). Also observe that \( f \) need only be defined on \( U \setminus \{ x_0 \} \), where \( U \) is an open set containing \( x_0 \). A frequent use of the notion of a limit is to define \( f \) at \( x_0 \), when the expression defining it at values near \( x_0 \) is undefined at \( x_0 \).

5. **Normed linear spaces.** The notion of a norm \( x \) on a vector space \( X \) over a field \( F \) equipped with an absolute value was defined above. As for absolute values on fields, the notation \( |x| \) conflicts with one used in chapter 10, where it denoted \( \sqrt{x \ast x} \) for an inner product \( \ast \), which is a map from \( F \times F \) to \( F_\mathbb{F}^\times \). From
hereon a norm will be assumed to be a map to \( \mathbb{R} \). Of course, in the cases \( F = \mathbb{R} \) or \( F = \mathbb{C} \) \( \sqrt{x^*x} \) is a standard norm.

The requirement that addition, scalar multiplication, and the norm be continuous is redundant and follows from the definition of a norm. The proof is similar to that of theorem 20.9. It also follows that the norm is continuous.

Let \( F\text{-NLS} \) denote the category of normed linear spaces over \( F \). Morphisms must preserve the linear space structure and be continuous. In general they need not preserve the norm; morphisms which do are called isometries, and are necessarily isomorphisms. Forgetting the norm yields a full subcategory of \( F\text{-TVS} \). An object in the latter category is called normable if it arises this way. We also require the absolute value on \( F \) to be nontrivial; the case of \( F \) trivial is of little interest, and also some basic theorems fail to hold.

**Lemma 9.** Let \( f : X \rightarrow Y \) be a vector space homomorphism spaces in \( F\text{-NLS} \). Then \( f \) is continuous iff there is a real number \( r > 0 \) such that \( |f(x)| \leq r|x| \).

**Proof:** It was observed in section 3 that \( f \) is continuous iff it is continuous at 0. If \( r \) exists then \( f \) is continuous at 0; indeed, given \( \epsilon \) if \( |x| < \epsilon/r \) then \( |f(x)| < \epsilon \) (nontriviality is not required for this direction). For the converse, let \( a \in F \) be such that \( |a| > 1 \). For any \( x \neq 0 \) let \( b_x \) be \( a^n \) where \( |a^n| \leq |x| < |a^{n+1}| \); then \( |b_x| \leq |x| < K|b_x| \) where \( K = |a| \). If \( f \) is continuous then there is a real number \( \delta > 0 \) such that \( |x| < \delta \) implies \( |f(x)| < 1 \). Let \( c \in F \) be such that \( |c| < \delta/K \). Let \( y = (cx)/b_x \); then \( |y| < \delta \), so \( |f(y)| < 1 \), so \( |f(x)| < |b_x|/|c| \leq (1/|d|)/|x| \).

The lemma is false for the trivial absolute value. Let \( X \) be the finite sequences over \( F \) equipped with the norm \( |x| = 1 \) if \( x \neq 0 \); let \( Y \) be the same set where the norm is the index (from 1) of the last nonzero element; and let \( f \) be the identity.

**Corollary 10.** Two norms \( |x|_1 \) and \( |x|_2 \) in a vector space \( X \) induce the same topology iff there are real numbers \( r_1 > 0 \) and \( r_2 > 0 \) such that \( |x|_1 < r_1|x|_2 \) and \( |x|_2 < r_2|x|_1 \).

**Proof:** The identity function is continuous from norm 1 to norm 2 iff \( r_1 \) exists, and from norm 2 to norm 1 iff \( r_2 \) exists.

Under the circumstances of the corollary the norms are said to be equivalent. Another consequence of lemma 9 is the uniform continuity of a continuous function between normed linear spaces.

Biproducts exist in \( F\text{-NLS} \). It is readily verified that \( |x_1| + \cdots + |x_n| \) (where the norm of \( x_i \) is taken in \( X_i \)) is a norm on the product. As in lemma 17.28 it induces the product topology. Various other norms could be equally well used, for example \( \max(|x_1|,\ldots,|x_n|) \) or \( \sqrt{|x_1|^2 + \cdots + |x_n|^2} \) (exercise 3). The last of these has properties which make it preferable in some contexts.

By essentially the same construction arbitrary coproducts (direct sums) exist in \( F\text{-NLS} \). An element of the direct sum is an element of the direct product which is nonzero in only finitely many components, and the norms of the nonzero components may be added.

A subspace of a normed linear space is a normed linear space when equipped with the restriction of the norm. In particular the equalizer in Set is an equalizer, so finite limits exist in \( F\text{-NLS} \). The coinage-image factorization in Set is a coinage-image factorization. If \( Y \subseteq X \) is a subspace then the quotient \( X/Y \) is a topological vector space.

**Lemma 11.** Suppose \( Y \subseteq X \) is a subspace. Let \( |x + Y| = \inf \{|x + y| : y \in Y\} \).

a. \( |x + Y| \) is a pseudo-norm on \( X/Y \).

Suppose \( X/Y \) is equipped with this pseudo-norm. Let \( \eta \) be the canonical epimorphism onto the quotient vector space.

b. \( \eta \) is continuous.
c. $\eta$ is open, and the pseudo-norm topology on $X/Y$ is the quotient topology.

**Proof:** Letting $y$ vary over $Y$,

$$|ax + Y| = \inf \{|ax + y|\} = \inf \{|ax + ay|\} = |a| \inf \{|x + y|\} = |a||x + Y|.$$  

Also

$$|x_1 + x_2 + Y| = \inf \{|x_1 + x_2 + y|\} = \inf \{|(x_1 + y_1) + (x_2 + y_2)|\} \leq \inf \{|(x_1 + y_1)|\} + \inf \{|(x_2 + y_2)|\} = |x_1 + Y| + |x_2 + Y|.$$  

This proves part a. Clearly $|x + Y| \leq |x|$, and part b follows. For part c, we claim that $\eta[B_{0,\epsilon}] = B_{0,\epsilon}$; we have already seen that $\eta[B_{0,\epsilon}] \subseteq B_{0,\epsilon}$. Conversely if $|x + Y| < \epsilon$ then there is an $x'$ such that $|x'| < \epsilon$ and $x - x' \in Y$. For part d, if $Y$ is closed and $|x + Y| = 0$ then there is a sequence $y_i$ such that $|x + y_i|$ converges to 0, and since $Y$ is closed, $x \in Y$; that is, $x + Y$ is the zero element of the quotient.

In particular if $Y$ is closed than $X/Y$ is Hausdorff, which was shown in section 2 more generally for topological groups. If $Y$ is not closed, then since $\eta$ is continuous and $Y = \text{Ker}(\eta)$, $\{0\}$ is not closed in $X/Y$, whence $X/Y$ is not $T_1$, and the pseudo-norm is not a norm. For an example, let $X$ be the continuous functions from $[0,1]$ to $\mathbb{R}$, and let $Y$ be $\{f: f(0) = 0\}$. With the norm $\sup_x |f(x)|$ $Y$ is closed and $X/Y$ is isomorphic to $\mathbb{R}$. With the norm $\int_0^1 |f(x)| dx$ $X/Y$ is the indiscrete space; sequences in $Y$ converging to the function $f(x) = 1$ are readily constructed.

In $F$-NLS if $f \in \text{Hom}(X,Y)$ then by lemma 9 $\inf \{r \in B^2 : |f(x)| \leq r|x| \text{ for all } x \in X\}$ exists; denote this value as $|f|$. It may alternatively be characterized as $\sup \{|f(x)/|x| : x \neq 0\}$.

**Lemma 12.** $|f|$ as defined above is a norm on $\text{Hom}(X,Y)$.

**Proof:** This is proved by the following three observations.

1. $|f| = 0$ iff $f(x) = 0$ for all $x$ iff $f = 0$.
2. $|af(x)| \leq |a|r|x|$ iff $|f(x)| \leq r|x|$, whence $|af| = |a||f|$.
3. $|(f + g)(x)| \leq (|f| + |g|)|x|$. whence $|f + g| \leq |f| + |g|$.

This norm is a natural one for many uses; one name it goes by is the strong norm. It may readily be generalized to the vector space $L(X_1, \cdots, X_n; Y)$ of multilinear maps (defined in chapter 8). Recall the vector space isomorphism $f \leftrightarrow \bar{f}$ between $L(X,Y; Z)$ and $L(X; L(Y; Z))$ defined in section 18.3, where $f(x,y) = \bar{f}(x)(y)$, and $\bar{f}(x) = f_x$ where $f_x(y) = f(x,y)$. Then $|\bar{f}| \leq r$ iff for all $x$ $|\bar{f}(x)| \leq r|x|$ iff for all $x$ $|f_x| \leq r|x|$ iff for all $x, y$ $|f_x(y)| \leq r|x||y|$ iff for all $x, y$ $|f(x,y)| \leq r|x||y|$. Thus, defining $|f| = \inf \{|r : |f(x,y)| \leq r|x||y| \text{ for all } x, y\}$ yields a norm on $L(X,Y; Z)$ which makes the map $f \mapsto \bar{f}$ an isometry. Similarly, defining $|f| = \inf \{|r : |f(x_1, \cdots, x_n)| \leq r|x_1| \cdots |x_n| \text{ for all } x_1, \cdots, x_n\}$ for $f \in L(X_1, \cdots, X_n; Y)$ yields a norm under which various vector space isomorphisms are isometries.

Defining a norm on the tensor product ("tensornorm" or "crossnorm") is a complex subject (see [Ryan]). However, as already noted in chapter 18, in many cases of interest the required space can be regarded as a space of multilinear maps, and given the norm defined in the preceding paragraph.

The strong norm satisfies $|g \circ f| \leq |g||f|$, for $f : X \mapsto Y$ and $g : Y \mapsto Z$, as may be seen by noting that $|fg(x)| \leq |f||g||x|$. Further, composition is readily verified to be continuous, using $|fg - f'g'| \leq |f||g - g'| + |f - f'||g'|$. In particular, composition is a continuous operation on $\text{Hom}(X,X)$ with the strong norm.

Recalling the map $x \mapsto \phi_x$ from $V$ to $V^{**}$ of section 10.2, $|\phi_x| \leq r$ iff $|\phi_x(f)| \leq r|f|$ for all $f$ iff $|f(x)| \leq r|f|$ for all $f$; but $|f(x)| \leq |x||f|$ for all $f$, so $|\phi_x| \leq |x|$. In particular $x \mapsto \phi_x$ is continuous.

The "Hahn-Banach" theorem states that a continuous linear functional $f$ on a subspace $Y \subseteq X$ can be extended to all of $X$ without increasing the norm. If this holds in $F$-NLS, then $x \mapsto \phi_x$ is a homeomorphic
embedding; and with an additional condition an isometry. If \( F = \mathcal{R} \) or \( \mathcal{C} \) then both the Hahn-Banach theorem and the additional condition hold. The Hahn-Banach theorem also holds if the absolute value on \( F \) is non-Archimedean, \( F \) is spherically complete, and \( X \) is ultrametric. These facts are proved in the exercises. (By definition a normed linear space is ultrametric iff \( |x - z| \leq \max(|x - y|, |y - z|) \); this is readily seen to be equivalent to the requirement \( |x + y| \leq \max(|x|, |y|) \). This was the version used in section 20.4.)

The strong norm of a linear transformation has been seen to have a number of desirable properties. One might ask how it might be computed for an \( m \times n \) matrix \( M \) when \( \mathcal{R}^m \) and \( \mathcal{R}^n \) are equipped with the Euclidean norm. It turns out that it is the square root of the largest eigenvalue of \( M^T M \), which is the root of an algebraic equation in the matrix elements. There are various inequalities in terms of more easily computed matrix norms; see [GvL] for this and also the computation of eigenvalues, a modern achievement of numerical analysis.

**Lemma 13.** Suppose \( X \) is a normed linear space and \( f : X \rightarrow F \) is a linear functional whose kernel is not all of \( X \). Then \( f \) is continuous iff \( \text{Ker}(f) \) is closed. Further in this case \( X \) is homeomorphic to the product of \( \text{Ker}(f) \) and \( F \).

**Proof:** The first claim follows by lemma 5. In the proof of of b implies c, open balls may be used rather than balanced sets, noting that \( f[B_{0,r}] \) is bounded if it is not all of \( F \); also lemma 9 may be appealed to. Now, if \( w \not\in \text{Ker}(f) \) then \( X \) is the internal direct sum of \( \text{Ker}(f) \) and \( Fw \), that is, each \( x \in X \) can be written uniquely as \( y_1 + y_2 \) where \( y_1 \in \text{Ker}(f) \) and \( y_2 \in Fw \); this is a fact about vector spaces and can be readily verified directly. The kernel of \( x \mapsto y_2 \) is equal to \( \text{Ker}(f) \), so the map is continuous. By remarks in section 3 the second claim follows.

As remarked in chapter 22, the kernel of a linear functional which is not identically 0 is called a hyperplane. In \( F\text{-NLS} \), a hyperplane is closed iff the linear functional is continuous. An example of a non-continuous linear functional will be given in the next section.

**Theorem 14.** If \( F \) is a complete field and \( X \) an \( n \)-dimensional normed linear space over \( F \) then \( X \) is isomorphic to \( F^n \).

**Proof:** Let \( x_1, \ldots, x_n \) and \( \phi \) be as in the proof of theorem 6. We show by induction on \( n \) that \( \phi^{-1} \) is continuous. If \( n = 1 \) the map \( ax_1 \mapsto a \) is continuous because \( |ax_1| < |x_1| \epsilon \) implies \( |a| < \epsilon \). For \( n > 1 \) let \( Y_1 \) be the subspace generated by \( x_1 \) and let \( Y_2 \) that generated by \( x_2, \ldots, x_n \). By lemma 13 \( X \) is homeomorphic to \( Y_1 \oplus Y_2 \). Inductively \( Y_2 \) is homeomorphic to \( F^{n-1} \), and the theorem follows.

As observed in section 4, a locally compact field is complete. Complete fields need not be locally compact, although we omit an example (by Ostrowski’s theorem the absolute value must be non-Archimedean). Theorem 14 thus applies in some cases where theorem 6 does not. Completeness is necessary for the theorem; the real numbers 1 and \( \sqrt{2} \) generate a 2-dimensional subspace of \( \mathcal{R} \) over \( \mathcal{Q} \), and equipped with the inherited norm this space is not isomorphic to \( \mathcal{Q}^2 \).

If \( F \) is complete (locally compact) and \( X \) is a finite dimensional normed linear space then \( X \) is complete (locally compact), since the complete metric (locally compact) spaces are closed under finite product. If \( F \) is complete then a finite dimensional subspace of any normed linear space \( X \), being complete, is a closed subspace of \( X \).

It is also true that a locally compact normed linear space \( X \) over a locally compact field is finite dimensional; a proof is as follows. Suppose w.l.g. that \( B_{0,1} \subseteq \cup_{i \in S} B_{c_i,1/2} \) where \( S \) is finite. Let \( Y \) be the subspace generated by \( \{c_i\} \). If there is some \( x \in X - Y \), then \( d(x, Y) \) is positive since \( Y \) is closed. Suppose \( y \in Y \) and let \( w = (x - y)/|x - y| \); then since \( w \in B_{0,1} \) there is some \( i \in S \) such that \( |w - c_i| < 1/2 \). Then \( x = y + |x - y|c_i + |x - y|(w - c_i) \), whence \( d(x, Y) \leq |x - y||w - c_i| \), whence \( |x - y| > 2d(x, Y) \), a contradiction since \( y \) was arbitrary.
Recall from section 10.11 the correspondence between projection operators and direct sum decompositions in a vector space. In F-NLS, a continuous projection operator \( p \) yields a direct sum decomposition into closed subspaces. This follows because the subspaces are \( \text{Ker}(p) \) and \( \text{Ker}(\iota - p) \). In the next section the converse will be shown for complete normed linear spaces.

The following strengthening of Theorem 14 will be used in section 8.

**Theorem 15.** If \( F \) is a complete field, \( X \) is a space in F-NLS, \( S \) is a closed subspace of \( X \), and \( x \notin S \) then \( T = S + Fx \) is a closed subspace.

**Proof:** The map \( s + rx \mapsto r \) is a linear functional on \( T \) whose kernel is \( S \). Let \( g \) denote it; by lemma 13 it is continuous. If \( t_n \) is a sequence in \( T \) converging to \( t \in X \) then \( f(t_n) \) is a Cauchy sequence in \( F \), so has a limit \( r \). The sequence \( t_n - f(t_n)x \) is a sequence in \( S \) converging to \( t - rx \), which is in \( S \) since \( S \) is closed. Thus, \( t \in T \) and \( T \) is closed.

6. Complete normed linear spaces. In this section suppose \( F \) is a field complete under a given absolute value. Let F-CNLS denote the full subcategory category F-NLS, where the vector space is complete with respect to the norm. If \( F = R \) (\( F = C \)), a complete normed linear space is called a real (complex) iBanach space; some authors use the term over any \( F \). A space in F-TVS is called Banachable if it arises from forgetting the norm of a Banach space.

Banach spaces are the setting of functional analysis ([Dieudonné], [Simmons], [Rudin2]). Other complete fields are a more specialized topic, but are considered, in particular the \( p \)-adic fields. Various facts may be proved in general.

Suppose \( X \) is any set and \( Y \) is in F-CNLS (in the most usual case \( Y = F = R \)); the functions \( f : X \mapsto Y \) comprise a vector space with the pointwise operations. Say that such an \( f \) is bounded if there is an \( r \) such that \( |f(x)| < r \) for all \( x \in X \). Let \( B \) denote the bounded functions from \( X \) to \( Y \); these comprise a subspace of the vector space of all the functions. Indeed if \( r \) is a bound for \( f \) and \( s \) is a bound for \( g \) then \( |a|r \) is a bound for \( af \) and \( r + s \) is a bound for \( f + g \).

For \( f \in B \) define \( |f| \) to be \( \sup\{|f(x)| : x \in X\} \). This is readily verified to be a norm. Indeed \( |af| = \sup\{|af(x)|\} = \sup\{|a||f(x)|\} = |a|\sup\{|f(x)|\} = |a||f| \); and \( |f + g| = \sup\{|f(x) + g(x)|\} \leq \sup\{|f(x)|\} + \sup\{|g(x)|\} = |f| + |g| \). When \( Y \) is \( R \) or \( C \) this norm is known as the uniform norm; it is also called the “sup” norm.

Suppose \( f_n \) is a Cauchy sequence in \( B \). Then at each \( x \) \( f_n(x) \) is a Cauchy sequence in \( Y \) (because \( |f(x)| \leq |f| \)), and hence, since \( Y \) is complete, converges to some value, denoted \( f(x) \). For any \( \epsilon \) \( n_0 \) can be found such that if \( n \geq n_0 \) then \( |f_n - f_n(0)| < \epsilon \). It follows that \( |f - f_{n_0}| \leq \epsilon \), because this is true pointwise. Thus, \( B \) is complete.

The bounded functions from a set \( X \) to a space \( Y \) in F-CNLS thus form a space in F-CNLS. If \( X \) is a topological space the functions \( f : X \mapsto Y \) which are bounded and continuous form a subspace of \( B \), which we denote \( C \). This follows by general facts; for example \( f_1 + f_2 \) is the composition of + with an induced map, and so is continuous if \( f_1 \) and \( f_2 \) are.

Suppose \( S \subseteq C \) and \( f \in S^\dagger \). Given \( \epsilon \) and \( x \in X \), there is a \( g \in S \) such that \( |f - g| < \epsilon/3 \), and an open \( U \subseteq X \) such that \( g[U] \subseteq B_{g(x),\epsilon/3} \). If \( y \in U \) then

\[
|f(y) - f(x)| \leq |f(y) - g(y)| + |g(y) - g(x)| + |g(x) - f(x)| < \epsilon.
\]

That is, \( f[U] \subseteq B_{f(x),\epsilon} \); this shows that \( f \) is continuous, and \( C \) is a closed subset of \( B \), and finally that \( C \) is complete.

The continuous bounded functions from a topological space \( X \) to to a space \( Y \) in F-CNLS thus also form a space in F-CNLS. These functions may also be considered in the pointwise topology. In this case a
pointwise convergent sequence may not converge to a continuous function. When \( Y = \mathcal{R} \) or \( \mathcal{C} \) a sequence which converges in the uniform norm is called uniformly convergent, and in this case, as just shown, the limit is continuous (see also exercise 17.7). Note also that in the continuous functions on, say, \([0,1]\) there are examples of non-continuous linear functionals, namely \( f \mapsto f(x) \) for any \( x \in [0,1] \).

**Theorem 16.**

a. \( F \)-CNLS is closed under finite products.

b. The quotient of a space in \( F \)-CNLS by a closed subspace is in \( F \)-CNLS.

c. For \( X, Y \) in \( F \)-CNLS, \( \text{Hom}(X,Y) \) equipped with the strong topology is in \( F \)-CNLS (in fact, only \( Y \) need be complete).

**Proof:** Part a has already been observed. For part b, suppose \( x + Y \) is a Cauchy sequence in \( X/Y \). Refine it to a subsequence \( w_n + Y \) where \( |w_n - w_{n+1} + Y| < 1/2^n \). Choose \( v_0 \in w_0 + Y \), and inductively choose \( v_{n+1} \in w_{n+1} + Y \) with \( |v_n - v_{n+1}| < 1/2^n \). Then for \( m > n \) \( |v_n - v_m| < 1/2^n + \cdots + 1/2^{m-1} \leq 1/2^m \). So \( v_n \) is a Cauchy sequence in \( X \), and so has a limit \( v \). One readily verifies that \( v + Y \) is the limit of \( x_n + Y \). For part c, suppose \( f_n \) is a Cauchy sequence in \( \text{Hom}(X,Y) \). Given \( x \) and \( \epsilon \), there is an \( n_0 \) such that for \( n_1, n_2 \geq n_0 \) \( |f_n(x) - f_n(x)| < \epsilon/|x| \), whence \( |f_n(x) - f_n(x)| < \epsilon \). That is, \( f_n(x) \) is a Cauchy sequence, so has a limit, for which we write \( f(x) \). By an argument as above, \( f_n \) converges to \( f \) in the strong norm. By continuity one verifies that \( f \) is linear. Finally, by arguments already given, \( |f(x) - f_n(x)| \leq \epsilon/|x| \) for \( n \) sufficiently large, whence if \( |f_n(x)| \leq r|x| \) then \( |f(x)| \leq (r + \epsilon)|x| \), showing that \( f \) is continuous.

Part b in fact holds in the category of metrizable topological groups; see [BourT]. By part c and facts already proved, the vector space \( L(X_1, \ldots, X_n; Y) \) of multilinear maps is complete.

The special nature of linear maps may be seen in the following theorem, called the open mapping theorem, which does not hold for arbitrary continuous functions.

**Theorem 17.** In \( F \)-CNLS, if \( f \in \text{Hom}(X,Y) \) is surjective then \( f \) is open.

**Proof:** It suffices to show that for \( B_{0r} \subseteq X \), \( f[B_{0r}] \subseteq Y \) contains some \( B_{0s} \subseteq Y \). Write \( B_r \) for \( B_{0r} \subseteq X \) and \( B'_s \) for \( B'_{0s} \subseteq Y \); it suffices to show that for some \( s, B'_s \subseteq f[B_1] \). Since \( f \) is surjective, \( Y = \cup_n f[B_n] \), so by theorem 8 \( f[B_n]^{cl} \) has nonempty interior for some \( n \). It follows that \( f[B_n] \) has an interior point \( y \); \( f[B_n] - y \subseteq f[B_2n] \), so \( f[B_n]^{cl} - y \subseteq f[B_2n]^{cl} \), so \( 0 \) is an interior point of \( f[B_2n]^{cl} \), so \( 0 \) is an interior point of \( f[B_1]^{cl} \), so for some \( s, B'_s \subseteq f[B_1]^{cl} \). For \( y \in B'_s \) we will find an \( x \in B_2^{cl} \) such that \( f(x) = y \), from which \( f'_s \subseteq f[B_1] \). There is an \( x_1 \in B_1 \) such that \( |y - f(x_1)| < s/2 \). Inductively, there is an \( x_n \in B_{2^{n-1}} \) such that \( |y - f(\sum_{i \leq n} x_i)| < s/2^n \), since \( B'_s/2^n \subseteq f[B_1/2^n]^{cl} \). The partial sums \( \sum_{i \leq n} x_i \) form a Cauchy sequence, which converges to a value \( x \) with \( |x| \leq 2 \) and \( f(x) = y \).

In particular, a continuous vector space isomorphism is a homeomorphism. Even though the completion functor from \( F \)-NLS to \( F \)-CNLS is very well behaved, the completeness assumption cannot be dropped. Let \( X \) be the finite sequences of reals with the max norm, let \( Y = X \), and let \( f(x_1, \ldots, x_n) = \langle y_1, \ldots, y_n \rangle \) where \( y_i = x_i/n \); \( f \) is a continuous surjection which is not open. Its completion is not surjective.

If \( X \) and \( Y \) are topological spaces with \( Y \) Hausdorff, and \( f : X \rightarrow Y \) is continuous, then the graph of \( f \) is a closed subset of \( X \times Y \) (exercise 6). The closed graph theorem states that for linear transformations in \( F \)-CNLS the converse holds (exercise 7). Other conditions under which the converse holds may be found in [Kelley].

Another consequence of the open mapping theorem is the converse promised at the end of section 6. Suppose \( X = Y \oplus Z \) where \( Y \) and \( Z \) as closed; we show that the corresponding projection operator is continuous. Let \( |x| = |y| + |z| \), where the norms on \( Y \) and \( Z \) are the restrictions. The projection onto \( Y \) (or \( Z \)) is clearly continuous when \( X \) is equipped with this norm. Letting \( X_1 \) denote \( X \) equipped with \( |x| \), the identity map is continuous from \( X_1 \) to \( X \), and so is a homeomorphism by the open mapping theorem.
If $X,Y$ are spaces in $F$-CNLS, $U$ is an open subset of $X$, $f : U \mapsto Y$ is any function, and $x_0 \in U$, a function $\phi \in \text{Hom}(X,Y)$ is said to be the derivative of $f$ at $x_0$ if $|f(x) - f(x_0) - \phi(x - x_0)|/|x - x_0|$ approaches 0 as $x$ approaches $x_0$. Thus, $f(x_0) + \phi(x - x_0)$ “approximates”, or is “tangent to” $f$, where $g$ is said to be tangent to $f$ at $x_0$ if $\lim_{x \to x_0} |f(x) - g(x)|/|x - x_0| = 0$. This definition agrees with that of ordinary calculus, where $Y = X = \mathbb{R}$, in that the linear function is specified by giving a real number $r$ such that $\phi(x) = rx$.

In this section we will use $E(x)$ to denote $f(x) - f(x_0) - \phi(x - x_0)$. Then $\phi$ is the derivative iff $|E(x)|/|x - x_0|$ approaches 0 as $x$ approaches $x_0$, which is to say that $E(x)$ approaches 0 “faster than” $x - x_0$. In particular $E(x)$ approaches 0.

It is readily verified that if there is any linear function $\phi$ satisfying the requirement, then there is only one; further if $\phi$ exists then $\phi$ is continuous at $x_0$. Indeed, if $\phi_1$ and $\phi_2$ are linear functions satisfying the requirement then their difference $\psi = \phi_1 - \phi_2$ satisfies $\lim_{x \to 0} |\psi(x)|/|x| = 0$, whence $|\psi(x)| < \epsilon|x|$ for any $\epsilon$, whence $|\psi| < \epsilon$ for any $\epsilon$, whence $|\psi| = 0$ and so $\psi$ is identically 0. Since $|f(x) - f(x_0)| \leq |\phi(x - x_0)| + |E(x)|$, if $\phi$ is continuous both terms approach 0, whence $f$ is continuous at $x_0$. Similarly since $|\phi(x - x_0)| \leq |f(x) - f(x_0)| + |E(x)|$, if $f$ is continuous at $x_0$ then $\phi$ is continuous.

If the derivative $\phi$ exists $f$ is said to be differentiable at $x_0$. Since we have required the derivative to be continuous (some authors do not), $f$ is continuous at $x_0$. If $f$ is differentiable at every point of $U$ then there is a function $f'$ mapping $U$ to $\text{Hom}(X,Y)$ whose value at $x$ is the derivative at $x$. Often $f'$ is written indifferently for $\phi = f'(x_0)$, but to avoid confusion we have used different symbols.

It is well known from ordinary calculus that even if $f'$ exists throughout $U$ it may not be continuous; $f$ is said to be $C^1$ in $U$ if $f'$ exists and is continuous. More generally, if $f^{(k)} = f'^{k-1}$ exists $f$ is said to be $k$ times differentiable, and if $f^{(k)}$ is continuous $f$ is said to be $C^k$. By definition the $k$th derivative is in $\text{Hom}(X,\ldots,\text{Hom}(X,Y)\ldots)$; as seen in section 5 it may equally well be considered to be a multilinear map from $X^k$ to $Y$.

Suppose $f$ has derivative $\phi$ at $x_0$, and $g$ has derivative $\psi$, and $c$ is a scalar. We claim that at $x_0$

- $f + g$ has derivative $\phi + \psi$,
- $cf$ has derivative $c\phi$, and
- $fg$ has derivative $f(x_0)\psi + g(x_0)\phi$.

Indeed, the first fact follows since $E_{f+g} = E_f + E_g$, where we have subscripted the “error term” $E$ with the function. Similarly, the second follows since $E_{cf} = cE_f$. Finally, $E_{fg}$ equals $\phi(x - x_0)\psi(x - x_0) + f(x_0)E_g + g(x_0)E_f$. Clearly the last two terms approach 0 faster than $x - x_0$; the first does also, because its norm is bounded by a constant times $|x - x_0|^2$.

Suppose $f : U \mapsto Y$ where $U \subseteq X$, $g : V \mapsto Z$ where $V \subseteq Y$, $f[U] \subseteq V$, $x_0 \in U$, $y_0 = f(x_0)$, $f$ has derivative $\phi$ at $x_0$, and $g$ has derivative $\psi$ at $y_0$. Noting that $g \circ f : U \mapsto Z$, we claim that at $x_0$

- $g \circ f$ has derivative $\psi \circ \phi$.

Indeed, $E_{g \circ f}(x) = E_g(f(x)) + \psi(E_f(x))$, and the reader may verify that both terms approach 0 faster than $x - x_0$. This fact is called the chain rule.

By the uniqueness of the derivative it follows that if $f$ is a linear function on $X$ then the derivative of $f$ exists throughout $X$, and equals $f$. Similarly if $f$ is a constant function then $f' = 0$.

As has been seen, many basic facts of ordinary calculus continue to hold in $F$-CNLS, where $F$ may be non-Archimedean, the spaces infinite dimensional, or both. On the other hand, there are facts which fail to hold in general. For example, for non-Archimedean fields it does not follow that if $f' = 0$ then $f$ is constant; we omit further discussion (see [Robert] for example).

Many facts of ordinary calculus continue to hold for arbitrary spaces when $F = \mathbb{R}$ or $F = \mathbb{C}$ (the Archimedean case). A useful observation in this case is that $X$ (indeed any convex subset) is path connected.
(see below for convex sets over $C$). Given $y, z \in X$, for the next lemma let $S_{yz}$ be the line segment between $y$ and $z$; it is (the range of) the path from $y$ to $z$ with with parameterization $t \mapsto y + t(z - y)$ for $t \in [0,1]$.

**Lemma 18.** Suppose $F = R$ or $F = C$.

a. Suppose $f : [l, u] \mapsto X$, and for $t \in [l, u] f'(t)$ exists and $|f'(t)| \leq M$. Then $|f(u) - f(l)| \leq M(u - l)$.

b. Suppose $y, z \in X$, and for $x \in S_{yz} f'(x)$ exists and $|f'(x)| \leq M$. Then $|f(z) - f(y)| \leq M|z - y|$.

**Proof:** Suppose $\epsilon > 0$, $l \leq v < u$, and $|f(t) - f(l)| \leq (M + \epsilon)(t - l)$ for $t \in [l, v]$. There is a $\delta > 0$ such that $v + \delta \leq u$ and for $s \leq \delta$, $|f(v + s) - f(v) - f'(v)(s)| < \epsilon s$, whence $|f(v + s) - f(v)| \leq (M + \epsilon)s$, whence $|f(t) - f(l)| \leq (M + \epsilon)(t - l)$ for $t \in [l, v + \delta]$; it follow that $|f(t) - f(l)| \leq (M + \epsilon)(t - l)$ for $t \in [l, u]$, and since $\epsilon$ was arbitrary part a is proved. For part b, let $x(t) = y + t(z - y)$ and $g(t) = f(x(t))$. Then $g'(t) = f'(x)(z - y)$, so $|g'(t)| \leq M|z - y|$, so $|g(1) - g(0)| \leq M|z - y|$ and part b follows.

In [Dieudonné] for example, many facts are proved using this lemma, including the implicit function theorem, and the following. Suppose $f : X \mapsto Y$ is differentiable in an open subset $U$ of $X$, and the second derivative $\psi$ exists at a point $x \in U$. Then $\psi$ is a symmetric bilinear form. In the non-Archimedean case, theorems such as this fail. A notion of “strict differentiability” is defined to remedy the fact. Again, see the references for more information. See also section 12 for some additional facts about differentiation of functions $f : R^n \mapsto R^n$.

Analytic functions may be considered in general. Here we give a general definition; further discussion of a particular case may be found in chapter 27. Recall that a monomial in $n$ variables is an expression of the form $x_1^{\nu_1} \cdots x_n^{\nu_n}$, where $\nu_j \in \mathbb{N}$ is the power to which the $j$th variable $x_j$ is taken; the notation $x^\nu$ is used for a monomial, where $\nu \in \mathbb{N}^n$. Suppose $X$ is a space in $F$-CNLS. A formal power series in $n$ variables with coefficients in $X$ is a map $\nu \mapsto c_\nu$ from the monomials to $X$, $c_\nu$ being the coefficient of the monomial $\nu$; such may be denoted as $\sum_\nu c_\nu x^\nu$. A monomial $\nu$ determines a value in $F$ at each value in $F^n$ in the usual way; multiplying this by a coefficient yields a value in $X$.

A formal power series yields a series in $X$ at each point of $F^n$. As is readily verified, exercise 17.9 holds for series in $X$. Thus at each point in $F^n$ where the series is absolutely convergent a value in $X$ is determined. A basic open set (“polydisc”) in $F^n$ is of the form $B_1 \times \cdots \times B_n \subseteq U$, where $B_i = B_{w_i,r_i}$. Say that a formal power series $\sum_\nu c_\nu x^\nu$ is normally convergent in a polydisc if $\sum_\nu c_\nu s_1^{\nu_1} \cdots s_n^{\nu_n}$ converges absolutely whenever $s_i < r_i$ for $1 \leq i \leq n$. This implies that the series is absolutely convergent at $x_i - w_i$ for $\langle x_1, \ldots, x_n \rangle$ in the polydisc. The reason for this technical restriction will be seen in chapter 26; clearly it is unnecessary when $F = R$ or $F = C$.

A function $f : F^n \mapsto X$ is said to be analytic in an open set $U \subseteq F^n$ if for each $\langle w_1, \ldots, w_n \rangle \in U$ there is a polydisc $B_1 \times \cdots \times B_n \subseteq U$, where $B_i = B_{w_i,r_i}$, and a formal power series, such that the series converges normally in the polydisc, and for $\langle x_1, \ldots, x_n \rangle$ in the polydisc, the sum of the series at $x_1 - w_1, \ldots, x_n - w_n$ equals $f(x_1, \ldots, x_n)$.

**7. Inner product spaces.** If $\ast$ is a bilinear map on vector space over $F$, in order that $\sqrt{x \ast x}$ be a norm, we must suppose that $F = R$ or $F = C$. Thus, we only define topological inner product vector spaces for $F = R$ or $F = C$. As in chapter 10, $\ast$ is required to be symmetric bilinear or Hermitian sesquilinear respectively; and positive definite. Also as in chapter 10, letting $|x| = \sqrt{x \ast x}$, the Cauchy-Schwarz inequality $|x \ast y| \leq |x||y|$ holds, and $|x|$ is a norm, called the norm of the inner product.

An object of the category $F$-IPS is defined to be a vector space over $F$, equipped with an inner product. It is considered an object of $F$-NLS by equipping it with the norm of the inner product. Using the Cauchy-Schwarz inequality, it is easy to see that the inner product is continuous. $F$-IPS is the full subcategory of $F$-NLS with these objects; that is, the morphisms are the continuous linear transformations.

As mentioned in chapter 10 the parallelogram law may be written as $|x + y|^2 + |x - y|^2 = 2(|x|^2 + |y|^2)$ (in fact in an inner product space over any field; but here we will only be concerned with $R$ and $C$). When
\( F = \mathcal{R} \) the polarization identity may be written as \( x \ast y = (1/4)(|x + y|^2 - |x - y|^2) \). When \( F = \mathbb{C} \) the polarization identity becomes \(|x + y|^2 - |x - y|^2 = 2(x \ast y + (x \ast y)^*)\); replacing \( y \) by \( iy \), and multiplying by \( i \) and adding, yields \( x \ast y = (1/4)(|x + iy|^2 - |x - iy|^2) \).

By the foregoing, an inner product is determined by its norm. In fact, if the parallelogram law holds then the expression above for \( x \ast y \) (in either case) yields an inner product with the given norm; the proof is left to exercises 8 and 9. On \( \mathbb{R}^n \), for any real number \( p \geq 1 \) the function \((\sum |x_i|^p)^{1/p}\) is a norm (see any real analysis book for a proof). The “max” norm may be viewed as the case \( p = \infty \). We state without proof that among these, the value \( p = 2 \) yields the only norm arising from an inner product (the norms are all equivalent, though).

If \( \phi_w \) is the linear functional with \( \phi_w(x) = x \ast w \), by the Cauchy-Schwarz inequality \(|\phi_w(x)| \leq |w||x|\). Thus, \( \phi_w \) is continuous, and \( |\phi_w| \leq |w| \). Since \(|w * (w/|w|)| = |w||w/|w||\), in fact \(|\phi_w| = |w|\), whence the map \( w \mapsto \phi_w \) is an isometry. It is linear in the case \( F = \mathbb{R} \), and antilinear if \( F = \mathbb{C} \).

For \( X \) in \( F \)-IPS, recall that for \( S \subseteq X \), \( S^\perp = \{ x \in X : y \ast x = 0 \text{ for all } y \in S \} \). In section 10.8 it was observed that the maps \( \perp \) from the subsets of \( X \) ordered by inclusion to \( / \) from the subsets ordered by opposite inclusion, form a Galois adjunction (indeed this is so under more general hypotheses on \( \ast \)). In section 10.11 some further properties were given (for general inner products). In \( F \)-IPS, \( S^\perp \) is a closed subspace of \( X \); indeed it equals \( \cap \{ \text{Ker} (\phi_w) : w \in S \} \).

8. Hilbert spaces. Suppose \( F = \mathbb{R} \) or \( F = \mathbb{C} \). Let \( F \)-CIPS denote the full subcategory of \( F \)-IPS, where the vector space is complete with respect to the norm. Such a space is called a real (complex) Hilbert space. In this section let \( X \) denote a (real or complex) Hilbert space.

Various spaces of real and complex analysis are Hilbert spaces; a simple and useful example is as follows. Let \( \ell_2 \) denote those infinite sequences \( x_i \) of real numbers, such that \( \sum x_i^2 \) is finite. Exercise 10 outlines a proof that this is a real Hilbert space, with inner product \( \sum_i x_i y_i \). There is a similar complex Hilbert space, where \( x_i \) is complex, \( \sum |x_i|^2 \) must be finite, and \( \sum_i x_i y_i^* \) is the inner product.

A vector space over \( \mathbb{C} \) may be considered a vector space over \( \mathbb{R} \) by only considering multiplication by reals (if \( X \) is \( n \)-dimensional over \( \mathbb{C} \) it is \( 2n \)-dimensional over \( \mathbb{R} \)). This fact is useful in that definitions given for real vector spaces may be applied in the complex case. For example a subset \( S \) of a Hilbert space \( X \) is convex if \( rx + (1-r)y \in S \) whenever \( x \), \( y \in S \) and \( r \) is a real with \( 0 \leq r \leq 1 \). Lemma 22.1 holds; the proof is unchanged, using only the parallelogram law. Lemma 22.2 holds as is for a real Hilbert space. For a complex Hilbert space, the real part of \((x - z) \ast (y - z)\) must be \( \leq 0 \); the proof requires only a minor modification to the algebra.

Lemma 19. Suppose \( S \) is a subspace of a Hilbert space \( X \); the following are equivalent.

a. \( S \oplus S^\perp = X \).

b. \( S = S^{\perp \perp} \).

c. \( S \) is closed.

Proof: That a implies b is proved in section 10.11. As noted above, \( S^\perp \) is closed for any \( S \), so b implies c. Suppose \( S \) is a closed subspace, and \( x \notin S \). Let \( z \in S \) be the closest point. By lemma 22.2, for \( y \in S \) \((x - z) \ast (y - z) \leq 0 \), and since \( -y \in S \) \((x - z) \ast (y - z) \geq 0 \) also, whence \((x - z) \ast (y - z) = 0 \). In the complex case, the real part of \((x - z) \ast (y - z)\) must be taken. Thus, \( x - z \in S^\perp \), which shows that \( x \in S \oplus S^\perp \) since \( x = z + (x - z) \).

Thus, the orthogonal decompositions are in bijective correspondence with the closed subspaces. It follows by section 10.12 that a projection operator is continuous. The proof of the lemma shows that for a closed subspace \( S \), the map \( x \mapsto z \) is the orthogonal projection corresponding to \( S \). If \( p \) is an orthogonal projection write \( x = x_1 + x_2 \) where \( x_1 = p(x) \); then \(|x|^2 = |x_1|^2 + |x_2|^2 \), a fact known as the Pythagorean theorem (one sees also from this that \( p \) is continuous).
If $S$ is a proper subspace and $x \notin S$ then $v = x - z$ where $z$ is as above, is a vector orthogonal to $S$. If $S$ is a hyperplane then $Fw$ is the orthogonal complement to $S$. If $f$ is a linear functional not identically 0, let $S$ be the kernel, and let $v$ be as just defined. Let $w = (f(v^*)/|v|^2)v$. Then $f(x) = x * w$; this is true for $x \in S$ since both sides are 0, and is true for $v$, and the claim follows. Thus, each linear functional on a Hilbert space is of the form $\phi_w$ as defined above ($w = 0$ for the identically 0 functional).

Given an orthonormal set $\{e_i\}$ of vectors in a Hilbert space, define the coefficient of $x$ with respect to $\{e_i\}$ to be $x * e_i$, and denote it as $c_i$. For a finite orthonormal set, let $r = x - \sum_i c_i e_i$; then $r$ is orthogonal to every $e_i$, and by the Pythagorean theorem $|r|^2 = |x|^2 - \sum_i |c_i|^2$. In particular $\sum_i |c_i|^2 \leq |x|^2$. From this, if $\{e_i\}$ is any orthonormal set, $|x * e_i| \geq |x|/n$ for at most $n$ $e_i$, whence $x * e_i \neq 0$ for at most countably many $e_i$.

Given any orthonormal set, and an element $x$, $\sum_i c_i e_i$ may be taken as a countable series. This series is absolutely convergent since $\sum_{i \leq n} |c_i|^2 < |x|^2$ for any $n$. Writing $r = x - \sum_i c_i e_i$, one verifies that $r$ is orthogonal to every $e_i$. The orthonormal sets are inductively ordered, so there are maximal orthonormal sets. If $\{e_i\}$ is such, then $r = 0$. Thus, every $x$ has a series expansion in terms of a maximal orthonormal set, which is clearly unique. Finally, note that $\sum_i |c_i|^2 \leq |x|^2$, a fact known as Bessel’s inequality.

Much as in the case of a basis for a vector space, the cardinality of a maximal orthonormal set in a Hilbert space is determined (indeed, given two infinite such, the cardinality of one is at most $\aleph_0$ times that of the other). Up to isomorphism, the Hilbert space is determined by this cardinality. For example a Hilbert space with a countable maximal orthonormal set is isomorphic to $\ell^2$.

Exercise 12 shows that a Hilbert space is separable iff it has a countable maximal orthonormal set. The Hilbert spaces of real and complex analysis are separable, and they have a variety of important countable maximal orthonormal sets.

The subspaces of a Hilbert space, like those of any vector space, form a modular lattice. The closed subspaces, being closed under arbitrary intersection, also form a (complete) lattice. In a finite dimensional Hilbert space every subspace is closed; but in an infinite dimensional space the sum of two closed subspaces need not be closed.

A standard example in $\ell_2$ is as follows. Let $S_1$ be the sequences where $x_{2n} = 0$. Let $S_2$ be the sequences where $x_{2n+1} = (n + 1)x_{2n}$. The vector space of all sequences is the direct sum of $S_1$ and $S_2$. For $i = 1, 2$ $T_i = S_i \cap \ell_2$ is a closed subspace of $\ell_2$. Since the finite sequences are in $T_1 + T_2$, the latter is dense. However, it is not all of $\ell_2$; indeed the sequence $y$ with $y_{2n} = 1/(n + 1)$ and $y_{2n+1} = 0$ is not in $T_1 + T_2$, as can be seen by looking at its decomposition in $S_1 + S_2$.

By lemma 19 and some other facts already proved, the lattice $L$ of closed subspaces is an ortholattice, with $\perp$ as complementation. Using lemma 22.2, if $S$ is a proper subspace of $T$ then there is a vector in $T \cap S^\perp$; thus, $L$ is an orthomodular lattice.

Let $T_1$, $T_2$, and $y$ be as above, and let $U = T_2 + Fy$; by theorem 15 $U$ is closed. Let $\sqcup$ denote the join in $L$. Then $T_2 \sqcup U$, $y \in (T_2 \sqcup T_1) \cap U$, and $y \notin T_2 \sqcup (T_1 \cap T_2)$, showing that $L$ is not modular.

9. Bundles. Let $C$ be a full subcategory of some $\text{TopS}_L$. Define a family in $C$ to be a 4-tuple $\langle X, B, p, \{X_b : b \in B\} \rangle$, where
1. $X$ is a topological space (the total space),
2. $B$ is a topological space (the base space),
3. $p : X \to B$ is a continuous function, and
4. $X_b$ is a structure in $C$ whose space is $p^{-1}(b)$ with the induced topology. If $C$ is $\text{Top}$ the family $\{X_b\}$ is absent. The notation may be abused by using $p^{-1}(b)$, which may be called the fiber at $b$, indifferently to denote $X_b$. The structure operations act “fiberwise”, i.e., individually on each fiber. $C$ may include the action of a topological group, which then acts on a family according to the general
Let $E$ be a component as transition functions. Thus, $p$ the first component and $C$ the subcategory of families in $\phi$. Lemma 20. Given a topological space $X_1$ a morphism from family 1 to family 2 is a pair $\langle \phi, \psi \rangle$ of continuous functions such that

1. $\phi : X_1 \rightarrow X_2$,
2. $\psi : B_1 \rightarrow B_2$,
3. $\phi p_1 = p_2 \psi$, and
4. the induced map $\phi_b : p_1^{-1}(b) \rightarrow p_2^{-1}(\psi(b))$ is a morphism in $C$.

That is, $\phi$ “acts on the fibers”, and does so preserving the algebraic structure.

The families in $C$ are readily verified to form a category. If $B$ is fixed, and $\psi$ required to be the identity, the subcategory of families in $C$ over $B$ is obtained. Henceforth only such categories will be considered.

If $F \in C$, the space $B \times F$ is a $C$-family over $B$ in an obvious manner, namely $p$ is the projection on the first component and $X_b$ is $F$. This family is called the trivial bundle over $B$, with fiber $F$. Note that the structure operations act continuously on the total space, when paired with the identity on $B$.

A bundle over $B$ with fiber (or “typical fiber”) $F$ is defined to be a family over $B$ which is locally isomorphic to the trivial bundle with fiber $F$; that is, together with an open cover $\{U_\alpha\}$ of $B$, and isomorphisms of families $\Phi_\alpha : p^{-1}(U_\alpha) \rightarrow U_\alpha \times F$. The isomorphisms are called “local trivializations”.

If $F$ is merely a topological space the bundle is called a fiber bundle. If $F$ is a topological vector space the bundle is called a vector bundle; if $F$ is a one (n) dimensional real or complex vector space, a bundle is called a line (n-plane) bundle. Vector bundles are the most common type, but occasionally the general definition is of use. For example $F$ may be required to have additional structure; another example will be given below. The basic definition of bundles may be given in various ways; the foregoing method follows the definition of vector bundle in [Atiyah].

Given a bundle, for $b \in U_\alpha$ let $\phi_{ab} : p^{-1}(b) \rightarrow F$ be the map $(\pi_2 \circ \Phi_\alpha) | p^{-1}(b)$. For $b \in U_\alpha \cap U_\beta$ let $\phi_{\beta ab}$ denote $\phi_{\beta \beta} \circ \phi_{\alpha \beta}^{-1}$; this map is an automorphism of $F$. If $U_{\alpha} \cap U_{\beta} \neq \emptyset$ let $\Phi_{\beta \alpha}$ denote the map taking $\langle b, w \rangle$ to $\langle b, \phi_{\beta \alpha}(w) \rangle$; these maps are called the transition functions. The $\phi_{\beta \alpha}$ satisfy

$(\ast) \phi_{\alpha \beta} = \phi_{\alpha \beta} \circ \phi_{\beta \alpha}^{-1}$, and $\phi_{\alpha \beta} \circ \phi_{\alpha \beta} = \phi_{\alpha \beta}$. The restrictions are redundant, in that the others follow from the last. Other maps are often defined, such as $b \mapsto \phi_{\beta \alpha}$ from $B$ to $\text{Aut}(F)$.

The basic definition has various specializations. For example Top may be replaced by a subcategory, in particular the smooth manifolds of some type; or the spaces having some property such as compactness. For another example, the Hom sets of $C$ might be given a topology, and the map $b \mapsto \phi_{\beta \alpha}$ from $B$ to $\text{Aut}(F)$ required to be continuous.

The notion of a section in a bundle is frequently encountered, and will be used below. A section in a bundle is a continuous function $s : B \rightarrow X$ such that $ps = t_B$.

**Lemma 20.** Given a topological space $B$, an open cover $\{U_\alpha\}$ of $B$, and family automorphisms $\Phi_{\beta \alpha}$ on $(U_\alpha \cup U_\beta) \times F$ such that $(\ast)$ above is satisfied, there is a bundle, unique up to isomorphism, having the $\Phi_{\beta \alpha}$ as transition functions.

**Proof:** Let $D$ be the disjoint union of the $U_\alpha \times F$, and let $\equiv$ be the binary relation on $D$ where $\langle \alpha, b, w_1 \rangle \equiv \langle \beta, b, w_2 \rangle$ when $b \in U_\alpha \cap U_\beta$, and $w_2 = \phi_{\beta \alpha}(w_1)$. Using $(\ast)$ is readily verified to be an equivalence relation. Let $E = D/\equiv$ be the quotient space, with $\eta$ the canonical epimorphism. Let $p_0 : D \rightarrow B$ map $\langle \alpha, b, w \rangle$ to $b$; since $p_0$ respects $\equiv$ a map $p : E \rightarrow B$ is induced. If $U \subseteq B$ is open then $p^{-1}[U]$ is open in each component $U_\alpha \times F$, so is open. It is also readily verified to be saturated, from which it follows that $p^{-1}[U]$ is open. Thus, $p$ is continuous. To define $E_\alpha$ given $w_1, \ldots, w_k \in F$ and a function $f$ of the structure, if $b \in U_\alpha$ let $f([\langle \alpha_1, b, w_1 \rangle], \ldots, [\langle \alpha_1, b, w_1 \rangle]) = [\langle \alpha_1, b, f(w_1, \cdot, w_k) \rangle]$. It is readily verified that this specifies a well-defined function. To define $\Phi_\alpha$, let $\Phi_\alpha^{-1} : \langle b, w \rangle \rightarrow [\langle \alpha, b, w \rangle]$. It is readily verified that $\Phi_\alpha^{-1}$ is a
fiberwise isomorphism of structures (in particular there is no "collapsing" in the equivalence classes). Φ⁻¹ is a homeomorphism because it maps a basic open rectangle to the image of a saturated open set, and these images form a basis for the topology on E. To prove uniqueness, suppose X₁ for i = 1, 2 are the total spaces of two bundles. Suppose w.l.g that the open covers of the base space are the same, and let Φ₁,α be the local trivializations. Finally suppose the transition functions are the same. Define f : X₁ → X₂ on p⁻¹₁[Uₐ] to be Φ₂αΦ⁻¹₁. By hypothesis f is well-defined; also it is a bundle isomorphism.

There is a similar construction of bundle morphisms. Given a bundle morphism f : X → X', suppose the open cover of B is the same for both bundles; let Φ₁(Φₙ) be the trivializations for X (X'), and let F,F' be the fibers. A system of maps f₁ : U₁ × F → U₁ × F' is determined, where f₁ = Φ₁fΦ⁻¹₁. On the other hand, given a system f₁, in order for it to determine a bundle morphism it is necessary and sufficient that Φ⁻¹₁f₁Φ₁ = Φ⁻¹₁f₁Φ₁ on Uₐ ∩ U₁.

It is not difficult to see that limits exist in the category of families in C over B, indeed are obtained “fiberwise”. The resulting limit may not be a limit in the bundles, though. For example in the real line H₀ be the fibers. A system of maps a functor by category theory; the resulting morphisms will be the same as those of the above outline. Hom(C, X) comes equipped with a norm, and a trivial bundle can be considered to have this norm on every fiber. For an arbitrary bundle X, each fiber of p⁻¹[Uₐ] may be given the image of this norm under Φ⁻¹. However, these norms need not be the same for different α, but only equivalent; and there is no natural norm on the fibers of X.

Indeed, if H is an n-ary functor from C to C (allowing C or C^{op} in each argument position), and F₁, . . . , Fₙ are objects of C, let G = H(F₁, . . . , Fₙ). If U is an open subset of B, U × G is a trivial bundle. Given bundles X₁, with fiber F₁, we may assume the cover of B is the same for each X₁. To determine a bundle Y whose fiber is G, it is only necessary to construct the fiber automorphism φ₁αb from the φ₁αb, and show that the resulting Φ₁α is continuous. We will only consider the case that φ₁αb is H(φ₃αb, . . . , φₙαb).

In this case the conditions (⋆) are satisfied, because they are satisfied for each X₁ and H is a functor.

Suppose bundle morphisms f₁ : X₁ → X₁' are given, and let Y and Y' be the bundles constructed in this manner. The maps f₁ : U₁ × F₁ → U₁ × F₁' determine a map g₁ : U₁ × G → U₁ × G' fiberwise by applying H. Again, in the case being considered the compatibility requirement for the g₁ is satisfied, because the required identity follows from that for the f₁ by applying H. Thus, once the continuity of the Φ₁α and g₁ is established, the functor on C has been shown to induce a functor on the bundles.

**Theorem 21.** Suppose C is TopM_{T} for T a set of equations, possibly with a space acting continuously. In the bundles in C over a base space B, the product of two bundles exists.

**Proof:** First, note that the product of U × F₁ and U × F₂ is U × (F₁ × F₂); by theorem 17.1 it suffices to verify that the topology on U × (F₁ × F₂) is induced by the maps ⟨u, x₁, x₂⟩ → ⟨u, x₁⟩ for i = 1, 2. In the notation of the outline given above, the continuity of Φ₁α follows again using theorem 17.1, constructing the appropriate diagram of trivial bundles with U = Uₐ ∩ U₁. Let Y denote the resulting bundle; given bundle morphisms f₁ : Y' → X₁ for i = 1, 2, a unique map g : Y' → Y is determined by trivialization, and is a bundle morphism, showing that Y is the product. In more detail, for i = 1, 2 let f₁α : U₁ × (F₁ × F₂) → U₁ × F₁ be as above, and let g₁α : U₁ × (F₁ × F₂) → U × (F₁ × F₂) be the induced map. By theorem 17.1 g₁α is continuous; the compatibility requirement follows also, using uniqueness of the induced map. Note that the product is a functor by category theory; the resulting morphisms will be the same as those of the above outline.

Further constructions may depend on C, and possibly other specializations. For example in the vector bundles the product is a biproduct.

For another example, when C is F-NLS for a field F, a natural topology can be constructed on Hom(X₁, X₂). An object V comes equipped with a norm, and a trivial bundle can be considered to have this norm on every fiber. For an arbitrary bundle X, each fiber of p⁻¹[Uₐ] may be given the image of this norm under Φ⁻¹. However, these norms need not be the same for different α, but only equivalent; and there is no natural norm on the fibers of X.
For \( i = 1, 2 \) suppose \( X_i \) has fiber \( V_i \), and let \( W \) denote \( \text{Hom}(V_1, V_2) \); then \( U \times W \) may be equipped with the strong norm. Using the continuity of \( \Phi_{\alpha \beta}^{V_1} \) and \( \Phi_{\alpha \beta}^{V_2} \), and composition it follows that \( \langle b, g \rangle \mapsto \langle b, \Phi_{\alpha \beta}^{V_2} g \Phi_{\alpha \beta}^{V_1} \rangle \), that is, \( \Phi_{\alpha \beta}^W \), is continuous. A similar argument shows that \( g_a \) is continuous.

Thus, there is a bundle, which we denote as \( \text{Hom}(X_1, X_2) \), which fiberwise is \( \text{Hom}(V_1, V_2) \). Further, the morphisms from \( X_1 \) to \( X_2 \) in are bijective correspondence with the sections of this bundle, as may be seen by localizing.

Even though there is no natural system of norms on the fibers of a bundle, an obvious question is whether there is any such. This is the case provided \( B \) is paracompact Hausdorff. As noted above, for each \( \alpha \) each fiber of \( p^{-1}[U_\alpha] \) may be given a norm. Let \( \{\alpha \beta\} \) be a partition of unity subordinate to a locally finite refinement of \( \{U_\alpha\} \) (that such exists is shown in section 17.4). For each \( \beta \) choose some \( U_\alpha \) containing the support of \( \alpha \beta \); call the norm obtained using \( U_\alpha \) the norm corresponding to \( \alpha \beta \). On the fiber at some \( b \in B \), only a finite set \( \{a_i\} \) of the functions of the partition of unity are nonzero at \( b \), and the sum of their values is 1. Let \( |x|_i \) be the norm corresponding to \( a_i \). Let \( |x| = \sum_i a_i |x|_i \). It is easy to check that \( |x| \) is a norm on each fiber. Continuity follows using local finiteness.

The least strict definition of a subbundle is a subfamily which is a bundle. In some contexts this is too general, and authors give a more strict definition; in other contexts, for example finite dimensional real vector bundles, the least strict definition is adequate. The treatment of quotient bundles, exact sequences, etc., similarly varies with the context. Further discussion of this topic is omitted.

Suppose there is a topological group \( G \) acting on the fibers. A bundle is said to have \( G \) as a structure group if each \( \phi_{\alpha \beta} \) is in \( G \). A bundle always has the automorphism group (in \( \text{TopS}_H \) of the fiber \( F \) as a structure group. For example, a real \( n \)-plane bundle \( X \) has the “general linear group” \( GL(n) \) of linear transformations of nonzero determinant as a structure group. \( X \) is said to be oriented if it has \( GL_+(n) \) as a structure group, where \( GL_+(n) \) is the group of linear transformations of positive determinant.

Suppose \( C \) is \( R \)-IPS, and let \( B \) be an object with inner product \( \cdot \). The orthogonal group \( O(F) \) is those automorphisms preserving \( \cdot \). We claim that a bundle has \( O(F) \) as a structure group if it has a “global inner product”, i.e., a fiberwise inner product \( * \) which is carried to \( \cdot \) by each \( \phi_{\alpha \beta} \).

For \( x, y \in p^{-1}(b) \) define \( x * y = \phi_{\alpha \beta}(x) \cdot \phi_{\alpha \beta}(y) \). This is well-defined since for \( b \in U_\alpha \cap U_\beta \), \( \phi_{\alpha \beta}(x) \cdot \phi_{\alpha \beta}(y) = \phi_{\beta \alpha}(x) \cdot \phi_{\beta \alpha}(y) \) by the hypothesis; and is a global inner product. Supposing \( * \) exists, reversing the argument shows that \( O(F) \) is a structure group.

The argument showing the existence of a global norm when \( B \) is paracompact Hausdorff works equally well for a global inner product. That \( \sum_i a_i * x \) is symmetric, bilinear, and positive definite is readily verified. The foregoing arguments are readily adapted to \( C \)-IPS. Recall that in this case the group preserving the inner product is called the unitary group; it is denoted \( U(F) \) or something similar.

An \( n \)-sphere bundle is a fiber bundle whose fiber is \( S^n \). In an \((n + 1)\)-plane bundle with a global norm, the fiberwise \( S^n \) is an \( n \)-sphere subbundle.

A bundle where the fiber is a topological space with a topological group \( G \) acting regularly on the left is called a principal bundle. A typical example of a principal bundle is the bundle of unit tangent vectors on a 2-sphere, with the action of the group of rotations in the plane.

A fiber \( p^{-1}(b) \) of a principal bundle may be “identified” with \( G \) by choosing a “base point” \( s_b \). This induces an action preserving homeomorphism \( g \mapsto g s_b \) from \( G \) to \( p^{-1}(b) \), as in theorem 5.5. The map is bijective by the hypothesis that the bundle is principal, and continuous since the action of \( g \) is. The inverse is \( g^{-1} \) followed by a right translation of \( G \), where \( x \in U_\alpha \), so it also is continuous. A regular right action of \( G \) which commutes with the left action is a structure group for the bundle; indeed choosing a base point \( w_0 \) for the fiber, \( g \phi_{\beta \alpha}(w_0) = g \phi_{\beta \alpha}(w_0) = g(s_b h) \) where \( h \in G \).

The transition functions of a bundle are fiberwise maps in \( \text{Aut}(F) \). These may also be considered as acting on \( \text{Aut}(F) \) by right multiplication, and using lemma 20 a principal bundle with \( \text{Aut}(F) \) acting on itself

337
by left multiplication may be constructed. Reversing the process, a bundle with fiber $F$ can be constructed from a principal bundle with fiber $\text{Aut}(F)$. These two bundles are said to be associated.

Another example of a principal bundle occurs when $G$ is a topological group and $H$ a subgroup; the canonical epimorphism makes $G$ a principal bundle with fiber $H$, over the base space $G/H$.

In a principal bundle, if $s$ is a section then the map $(b,g) \mapsto g \cdot s(b)$ is readily verified to be a bundle isomorphism. It is easy to see from this that a principal bundle has a section iff it is isomorphic to the trivial bundle on $B$. A vector bundle, on the other hand, always has at least one section, namely the 0 element of each fiber.

A frame in a finite dimensional vector bundle is a set $\{ s_i \}$ of sections such that at each $b \in B$, $\{ s_i(b) \}$ is a basis for $p^{-1}(b)$. If $\{ e_i \}$ is a basis for the fiber $F$, then the maps $b \mapsto (b,e_i)$ are a frame for $B \times F$. On the other hand, given a frame for a bundle, mapping $\langle b, w \rangle$ to $\sum_i c_i s_i(b)$ where $w = \sum_i c_i e_i$ yields a bundle isomorphism from $B \times F$. The map is clearly a fiberwise isomorphism, and may be seen to be continuous in either direction, as a composition of continuous functions; the map $\sum_i c_i s_i(b) \mapsto c_i$ may be seen to be continuous by trivialization and linear algebra.

Finally we mention another consequence of the base space being paracompact Hausdorff, corollary 2.7.14 of [Spanier], which states that a fiber bundle over such a base space is a fibration. Fibrations are defined in section 27.1.

10. Sheaves. Recall from section 21.4 the contravariant functor $\Omega$ from $\text{Top}$ to $\text{Poset}$ (indeed $\text{Frm}$). This maps a topological space $X$ to the set $\Omega(X)$ of open sets of $X$, ordered by inclusion; and a continuous function $f : X_1 \mapsto X_2$ to the map $U \mapsto f^{-1}[U]$.

If $C$ is a category the functor category $(\text{C}^\text{op})^{\Omega(X)}$, henceforth denoted $\text{PrShf}(X,C)$, is called the category of presheaves on $X$ in $C$. A presheaf assigns to each open $U$ set an object $P(U)$ of $C$. There is also a system of maps (called “restriction maps”) $\rho_{UV} : P(U) \mapsto P(V)$ for pairs $U \supseteq V$, which obey the compatibility requirements that if $U \supseteq V \supseteq W$ then $\rho_{UV} = \rho_{VW} \circ \rho_{UV}$. A morphism from $P_1$ to $P_2$ is a system of maps $P_1(U) \mapsto P_2(U)$, which commute with the restriction maps.

A basic example of a presheaf, which we will call a function presheaf, assigns to each open set $U$ a set (real vector space, etc.) of functions on $U$ (continuous, differentiable, etc.). The restriction map is restriction of functions. Other important examples of $C$ include modules, rings, chain complexes, etc.

As usual, by defining presheaves in any category many facts can be proved in general, sometimes requiring restrictions on $C$. In particular, limits and colimits in $\text{PrShf}(X,C)$ are taken “pointwise” (see section 19.1), which in this case means on each open set. If $C$ is complete, cocomplete, Abelian, or exact then $\text{PrShf}(X,C)$ is. When $C$ is exact, images and coimages are taken pointwise.

For the rest of this section we assume that $C$ is complete, cocomplete, and exact, although various facts holds under weaker restrictions.

Suppose $P$ is a function presheaf; suppose $U = \cup \{ U_i \}$. If $f \in P(U)$, and $f_i = \rho_{U_i,U}(f)$, then the $f_i$ are readily seen to satisfy the compatibility restrictions $\rho_{U_i,U \cap U_j}(f_i) = \rho_{U_j,U \cap U_i}(f_j)$. On the other hand, given $f_i \in P(U_i)$ satisfying these restrictions, there is a unique $f \in P(U)$ such that $f_i = \rho_{U_i,U}(f)$. Indeed, given $x \in U$, let $i$ be such that $x \in U_i$ and define $f(x) = f_i(x)$; the compatibility restrictions ensure that $f(x)$ is independent of the choice of $i$.

In any concrete category, with requirements just as stated above, a presheaf is defined to be a sheaf if the compatibility restrictions imply the existence of a unique $f$. A function presheaf is a sheaf. The requirements may be stated for an arbitrary category, as follows (see [Mitchell]).

Given a family $\{ U_i \}$ of open sets, let $U = \cup \{ U_i \}$. Let $u : P(U) \mapsto \times_i P(U_i)$ be the arrow induced by the arrows $\rho_{U_i,U}$. Let $s : \times_i P(U_i) \mapsto \times_{ij} P(U_i \cap U_j)$ be the arrow induced by the arrows $\rho_{U_i,U \cap U_j}$. Let $t : \times_i P(U_i) \mapsto \times_{ij} P(U_i \cap U_j)$ be the arrow induced by the arrows $\rho_{U_j,U \cap U_i}$. Then $P$ is a sheaf provided
u is the equalizer of s and t for all families \( \{U_i\} \).

Note that \( su = tu \), because \( \rho_{U_i, U_j \cap U_i} \circ \rho_{U_j, U'} = \rho_{U_j, U_j \cap U_i} \circ \rho_{U_i, U_j \cap U_i} \). For a concrete category, the category-theoretic and elementwise definitions are equivalent. Indeed, a family \( \{f_i\} \) is just an element \( \tilde{f} \) of \( \times_i \mathcal{P}(U_i) \), and the family is compatible if \( s(f) = t(f) \); remaining details are left to the reader.

Let \( \mathcal{S}(X, C) \) be the full subcategory of \( \mathcal{P}\mathcal{S}(X, C) \), whose objects are the sheaves. The limit of a diagram in \( \mathcal{P}\mathcal{S}(X, C) \), whose objects are sheaves \( \mathcal{P}_a \), is a sheaf, because each equalizer requirement holds for the limit since it holds for all the \( \mathcal{P}_a \). A colimit in \( \mathcal{P}\mathcal{S}(X, C) \) need not be a sheaf (to quote [Harts], “if \( \phi \) is a morphism of sheaves . . . the presheaf cokernel and presheaf image of \( \phi \) are in general not sheaves”). There is a category-theoretic construction of colimits in \( \mathcal{S}(X, C) \), though.

For \( x \in X \) the open sets \( U \subseteq X \) such that \( x \in X \) form a filtered subset of \( \Omega(X) \), and via a presheaf \( P \) give rise to a direct system in \( C \), with the restriction maps as morphisms (smaller sets are higher in the order). There is a functor \( \mathcal{S}\mathcal{t}lk_x \) from \( \mathcal{P}\mathcal{S}(X, C) \) to \( C \) (indeed a colimit functor on a fixed index category), which assigns to \( P \) the direct limit of the direct system. The object \( \mathcal{S}\mathcal{t}lk_x(P) \) is called the stalk of \( P \) at \( x \).

As an example of a stalk, consider a function sheaf. An element of the stalk is called “germs”.

**Theorem 22.** Suppose \( C \) is a suitable concrete category, and \( P \in \mathcal{P}\mathcal{S}(X, C) \). Then there is a universal arrow \( \mu : P \rightarrow \tilde{P} \) from \( P \) to the forgetful functor from \( \mathcal{S}(X, C) \) to \( \mathcal{P}\mathcal{S}(X, C) \), that is, such that given any arrow \( \alpha : P \rightarrow Q \) where \( Q \) is a sheaf, there is a unique arrow \( \beta : \tilde{P} \rightarrow Q \) with \( \alpha = \beta \mu \).

**Proof:** Let \( P_x \) denote \( \mathcal{S}\mathcal{t}lk_x(P) \), and let \( K \) denote the disjoint union of the \( P_x \). Suppose \( V \in \Omega(X) \). Let \( \rho_{V_x} \) denote the colimit arrow from \( P(V) \) to \( P_x \). For \( g \in P(V) \) let \( \tilde{g} : V \rightarrow K \) be the function where \( \tilde{g}(x) = \rho_{V_x}(g) \). For \( U \in \Omega(X) \) the elements of \( \tilde{P}(U) \) are those \( f : U \rightarrow K \) satisfying the following restriction.

\( \star \) For \( x \in U \), there is a \( V \in \Omega(X) \) with \( x \in V \subseteq U \), and a \( g \in P(V) \), such that \( f|V = \tilde{g} \). It is readily verified that if \( U_1 \subseteq U_2 \) and \( f \in \tilde{P}(U_2) \) then \( f|U_1 \in \tilde{P}(U_1) \). Thus, we may let \( \tilde{P} \) be the function sheaf with the \( \tilde{P}(U) \) as just given. The map \( g \mapsto \tilde{g} \) is an arrow \( \mu_V : P(V) \rightarrow \tilde{P}(V) \), and these arrows commute with the restriction maps, yielding \( \mu \). Suppose \( \alpha : P \rightarrow Q \) where \( Q \) is a sheaf. Given \( f \in \tilde{P}(U) \), and \( x \in U \), let \( V_x \) and \( g_x \) be the \( V \) and \( g \) as in requirement (\( \star \)); then \( U = \bigcup_x V_x \). Since \( Q \) is a sheaf we may define \( \beta(f) \) to be that \( h \) where \( h|V_x = \alpha(g_x) \).

The sheaf \( \tilde{P} \) of the theorem is called the associated sheaf. It may be constructed in a category satisfying certain axioms, called an \( F \)-category; see [Mitchell]. The suitability requirement on \( C \) is that \( \tilde{P}(U) \) may be made an object in \( C \) by pointwise operations; examples include CRng, Ab, R-Mod, etc.

Suppose \( C' \) is a full subcategory of \( C \), and the inclusion functor has a left adjoint \( F \) (for example, as just shown, the associated sheaf functor from \( \mathcal{P}\mathcal{S}(X, C) \) to \( \mathcal{S}(X, C) \)). Under such circumstances, \( C' \) is said to be a reflective subcategory of \( C \). Note that if \( a \) is an object of \( C' \) then \( F(a) \) is isomorphic to \( a \). Note also that a component \( \mu_c \) of the unit maps \( c \rightarrow F(c) \).

Suppose \( \alpha_i : d_i \rightarrow c \) is a colimit cone in \( C \), where the objects \( d_i \) are in \( C' \). Letting \( \beta_i : d_i \rightarrow F(c) \) is a cone in \( C' \). In fact it is a limit cone; if \( \beta_i' : d_i \rightarrow c' \) is any other cone in \( C' \), there is a unique map \( h : c \rightarrow c' \) in \( C \) determined by the \( \beta_i' \), and a unique map \( F(c) \rightarrow c' \) determined by \( h \). (It already follows that \( F(c) \) is a colimit object, from the fact that \( F \) is a left adjoint.)

There is a category \( \mathcal{P}\mathcal{S}(C) \) whose objects are those in \( \mathcal{P}\mathcal{S}(X, C) \) for varying \( X \). A morphism from \( P_1 \) on \( X \) to \( P_2 \) on \( Y \) consists of a continuous function \( f : X \rightarrow Y \), and maps \( P_2(U) \rightarrow P_1(f^{-1}[U]) \) which commute with the restriction maps.

Suppose \( f : X \rightarrow Y \) is a continuous function and \( P \) is a presheaf in \( \mathcal{P}\mathcal{S}(Y, C) \). Since \( f^{-1}[U_1] \supseteq f^{-1}[U_2] \) if \( U_1 \supseteq U_2 \), there is a presheaf \( Q \) in \( \mathcal{P}\mathcal{S}(X, C) \) where \( Q(U) = P(f^{-1}[U]) \); the restriction from \( Q(U_1) \) to \( Q(U_2) \) is just the appropriate restriction map of \( P \). Indeed there is a functor from \( \mathcal{P}\mathcal{S}(Y, C) \) to \( \mathcal{P}\mathcal{S}(X, C) \),
with object map mapping $P$ to $Q$; again a component of a morphism from $Q_1$ to $Q_2$ is just an appropriate component of the morphism from $P_1$ to $P_2$. $Q$ is called the direct image of $P$.

Again supposing $f : X \mapsto Y$ is a continuous function, let $\text{Dir}$ denote the direct image functor from $\text{PrShf}(X,C)$ to $\text{PrShf}(Y,C)$. Let $P$ denote an object of $\text{PrShf}(X,C)$, and $Q$ an object of $\text{PrShf}(X,C)$. We leave it to exercise 13 to show the following. For fact 3, we assume that $\text{Shf}(X,C)$ is a reflective subcategory of $\text{PrShf}(X,C)$ for all $X$ (which holds for concrete $C$ for example).

1. $\text{Dir}$ has a left adjoint, called the inverse image, and denoted $\text{Inv}$. $\text{Inv}(Q)(U)$ is the direct limit of the direct system $\{Q(V) : V \supseteq f[U]\}$.
2. The direct image if a sheaf is a sheaf.
3. The direct image functor from $\text{Shf}(X,C)$ to $\text{Shf}(Y,C)$ has a left adjoint. The object corresponding to $Q(U)$ is the sheaf associated to $\text{Inv}(Q)(U)$.

**11. General manifolds.** Manifolds which are locally isomorphic to a topological vector space $E$ other than $\mathbb{R}^n$ or $\mathbb{C}^n$ have various applications. As usual, as restrictions are relaxed some aspects of the theory remain valid, and others do not. For one example, infinite dimensional Lie groups may be defined, and have been of increasing interest; however their theory is more complicated and is an active area of research.

Suppose $E$ is a topological vector space. We may define a manifold with “model space” $E$ to be a topological space $X$, an open cover $\{U_i\}$ of $X$, and maps $\phi_i : U_i \mapsto E$ such that $\phi_i$ is a homeomorphism to an open subset of $E$.

In specific applications, $E$ might be required to be a restricted type of topological vector space. Examples include

- complete normed linear space over a complete field;
- real or complex vector space with a complete translation-invariant metric;
- real or complex locally convex topological space.

In the above cases, a notion of differentiation of a map may be defined (indeed this has been done in the first case in section 6), and hence various notions of smooth manifold, by imposing a smoothness requirement on the transition maps, which are defined as in the Euclidean case. The notion of a maximal atlas for a smooth manifold is defined as in the Euclidean case.

Also, restrictions might be imposed on $X$, such as that it be Hausdorff.

Suppose $\{V_i\}$ is an open cover of a topological space $X$, and each $V_i$ is a manifold with model space $E$. Suppose that the atlases are compatible, meaning that any two charts from their union are compatible.

Then as is readily verified, the union of the atlases is an atlas on $X$. The coproduct (disjoint union) of manifolds with model space $E$ is a special case.

Let $C$ be a category of model spaces. The manifolds with model space some $E$ in $C$ are a full subcategory of $\text{Top}$. A category of smooth manifolds is generally not full; a morphism is a continuous function $f : X \mapsto Y$, such that $\psi_j \circ f^{-1}_{i} : \phi_i[f^{-1}[V_j] \cap U_i] \mapsto \psi_j[V_j]$ is smooth for all $i,j$.

A finite product usually exists in such a category of manifolds. The model space is the product of the model spaces, the sets of the open cover are the products of open sets of the individual covers, and the maps are the product maps.

As in the Euclidean case, an open subset of a manifold is a manifold, with the same model space and obvious charts. This is not an example of what is usually meant by a submanifold; this is defined as follows. Suppose $X$ is a manifold with model space $E$, $Y$ is a subspace of $X$, and $F$ is a sub-topological vector space of $E$. $Y$ is said to be a submanifold of $X$, with model space $F$, if for every $y \in Y$ there is a chart $\phi$, with domain $U$, such that $y \in U$, and $\phi|Y \cap U = \phi[U] \cap F$. We also require that $U \cap F$ be open in $F$ when $U$ is open in $E$; this is the case for example if $E$ has the product topology of $F$ and another subspace. Such being the case, $Y$ is a manifold with model space $F$, and charts $\phi|Y \cap U$ “inherited” from $X$. For example,
the \( n \)-sphere \( S^n \) is a submanifold of \( \mathbb{R}^{n+1} \).

For some constructions, differentiability is required. We will consider some such in the next section, only for Euclidean manifolds. Other constructions require the inverse function theorem, or the Hahn-Banach theorem.

12. Tensors on manifolds. The “directions” at a point \( x \in \mathbb{R}^n \) constitute a copy of \( \mathbb{R}^n \), and the totality of these for varying \( x \) may be considered to be a trivial bundle with base space \( \mathbb{R}^n \) and total space \( \mathbb{R}^{2n} \). This bundle is called the tangent bundle on \( \mathbb{R}^n \). If \( U \) is an open subset of \( \mathbb{R}^n \) the tangent bundle on \( U \) is just the bundle \( U \times \mathbb{R}^n \).

If \( X \) is an \( n \)-dimensional Euclidean, although the same construction applies more generally, the tangent bundles on the sets \( U_\alpha \) of the open cover can be pieced together using lemma 20, provided \( X \) is at least \( C^1 \). We use \( \Upsilon_\alpha \) for the local trivializations, etc.; for \( x \in U_\alpha \cap U_\beta \) let \( v_\beta \alpha x = (\phi_\beta \phi_\alpha^{-1})'(\phi_\alpha(x)) \). Using the chain rule,

\[
v_\gamma \alpha x = (\phi_\gamma \phi_\alpha^{-1})'(\phi_\alpha(x)) = (\phi_\gamma \phi_\beta^{-1})'(\phi_\beta(x))(\phi_\beta \phi_\alpha^{-1})'(\phi_\alpha(x)) = v_\gamma \beta x v_\beta \alpha x.
\]

Thus, the hypotheses of lemma 20 are satisfied. Using a common notation, we write \( T(X) \) for the \( n \)-plane bundle on \( X \) just constructed; it is called the tangent bundle.

If \( X \) is a \( C^p \) \( n \)-manifold where \( p \geq 1 \) then \( T(X) \) is in fact a \( C^{p-1} \) \( 2n \)-manifold. Indeed, the map \( (\phi_\alpha \times i) \circ \Upsilon_\alpha \), i.e., \([\alpha, x, w] \mapsto (\phi_\alpha(x), w)\), is a chart mapping \( p^{-1}[U_\alpha] \) to \( \phi_\alpha(U_\alpha) \times \mathbb{R}^n \). The transition function from index \( \alpha \) to index \( \beta \) takes \( (\phi_\alpha(x), w) \) to \( (\phi_\beta(x), v_\beta \alpha x(w)) \), and is \( C^{p-1} \).

\( X \rightarrow T(X) \) is the object map of a functor from the category of \( C^p \) \( n \)-manifolds to the category of \( C^{p-1} \) \( 2n \)-manifolds. For \( f : X \rightarrow Y \) \( T(f) \) may be described as follows. Suppose \( x \in U_\alpha \) and \( f(x) \in \psi_\beta \); then \([\alpha, x, w] \) maps to \([\beta, f(x), (\phi_\beta f \phi_\alpha^{-1})'(\phi_\alpha(x))(w)]\), where \( \phi_\alpha \) (\( \psi_\beta \)) are the charts on \( X \) (\( Y \)). Remaining details are left to the reader.

If \( x \in X \) and \( p : T(X) \rightarrow X \) is the bundle projection, the fiber \( p^{-1}(x) \) is called the tangent space of \( X \) at \( x \). It may be given as the set of equivalence classes \([\alpha, x, w] \). The notation \( T_x(X) \) is commonly used for the tangent space. It is isomorphic to \( \mathbb{R}^n \), by an isomorphism which depends in a specified manner on \( \alpha \). This property is characteristic of vectorial quantities defined on a manifold. A vector field on \( M \) may be defined as a section \( s : X \rightarrow T(M) \). A smooth vector field is simply one where \( s \) is a sufficiently smooth map of manifolds.

To describe the tangent vector transformation law more explicitly, some facts about partial derivatives are required. These may be found in various calculus texts; a brief discussion is given here. If \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), and \( \lim_{t \rightarrow 0}(f(x+te_i)-f(x))/t \) exists, it is called the \( i \)th partial derivative of \( f \) at \( x \); the symbol \( \partial f/\partial x_i \) is used to denote it. It is just the derivative of \( f \), considered as a function of \( x_i \) alone, with the other variables held constant.

If \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \) then \( f = (f_1, \ldots , f_m) \) for some functions \( f_j : \mathbb{R}^n \rightarrow \mathbb{R} \) (the components of \( f \)). We leave it to the exercise 14 to show the following, for \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \).

1. If \( f' \) exists at \( x \) then in the matrix giving the column vector \( \Delta y = f'(x)(\Delta x) \) as a function of the column vector \( \Delta x \), the entry in row \( j \) and column \( i \) is \( \partial f_j/\partial x_i \). (Recall from the definition of \( f' \) that \( \Delta y \approx f(x+\Delta x)-f(x) \), in that the difference approaches 0 faster than \( \Delta x \).

2. If \( f \) is \( C^1 \) at \( x \) then each \( \partial f_j/\partial x_i \) (considered as a function of the point \( x \)) is.

3. If \( \partial f_j/\partial x_i \) is \( C^1 \) at \( x \) for all \( i, j \) then \( f \) is \( C^1 \) at \( x \), and the matrix of \( f' \) has these as components.

4. If \( m = n \), if \( f \) is bijective and \( C^1 \) at \( x \) then the matrix of \( f' \) is nonsingular, and its inverse is the matrix of \( (f^{-1})' \) at \( f(x) \).

Let \( M \) be the matrix of partials of \( v_\beta \alpha x \). Writing \( \phi_\alpha \) indifferently for the vector \( \phi_\alpha(x) \) in \( \mathbb{R}^n \), \( \phi_\beta \) is a function of \( \phi_\alpha \), and \( M_{ij} = \partial \phi_\beta/\partial \phi_\alpha \). From \(([\alpha, x, v_\alpha]) = ([\beta, x, v_\beta]) \) iff \( v_\beta = v_\beta \alpha x(v_\alpha) \) we conclude that the
components of a vector of the tangent bundle transform under a change of chart, according to

\[ v^j_\alpha = \sum_i \frac{\partial \beta_j}{\partial \alpha_i} v^i_\beta. \]

It is readily seen that this is contravariant with respect to the change of basis from \([\alpha, x, e_i]\) to \([\beta, x, e_i]\) for \(T_x(X)\). This transformation law is thus considered to be the transformation law for contravariant vectors on a manifold.

The bundle whose fiber is the mixed tensors of type \(\langle k, l \rangle\) over \(\mathbb{R}^n\) may readily be constructed similarly; we use \(T^k_l(X)\) to denote it, so that \(T(X) = T^1_0(X)\). Indeed, the bundle whose fiber is the entire mixed tensor algebra is readily constructed.

Writing \(V\) for \(\mathbb{R}^n\) and \(V^*\) for the dual space, recall that the components of a tensor in \(V^*\) are written with respect to the dual basis \(\{X_j\}\). From this it follows that the transition functions for the bundle with fiber \(V^*\) should be \(v^{-1}_{\alpha x}\). A transition function for the tensor product space is the tensor product of the transition functions on the components of the product. Using the Einstein convention for indices, the component transformation law is

\[ (v_\alpha)_i^{j_k} = \sum_{i_1 \cdots i_k} \frac{\partial \phi_{\beta j_k}}{\partial \phi_{\alpha i_1}} \frac{\partial \phi_{\alpha p_1}}{\partial \phi_{\beta q_1}} \cdots \frac{\partial \phi_{\alpha p_k}}{\partial \phi_{\beta q_k}} (v_{\beta x})_i^{i_1 \cdots i_k}. \]

In section 27.3 differential operators will be defined, and it will be shown that tangent vectors can be viewed as such. In fact, the space of differential operators may be shown to be isomorphic to the tangent space; see [Spivak]. There are various other constructions of the tangent bundle.

Let \(g\) be a \(T^2_2(X)\) tensor field; \(g\) is said to be symmetric if at each \(x\), \(v_{ij} = v_{ji}\) in some (and hence any) chart. A pseudo-Riemannian metric on \(X\) is a symmetric \(T^2_0(X)\) tensor field, which is non-degenerate and has the same signature at all \(x \in X\). If it is positive definite it is a Riemannian metric. Metric tensors are important in applications of tensors on manifolds; see [Spivak] for example.

13. Homotopy. Let \(I\) denote the closed unit interval \([0, 1]\) in \(\mathbb{R}\). Two functions \(f_1, f_2 : X \mapsto Y\) in \(\text{Top}\) are said to be homotopic if there is a function \(F : X \times I \mapsto Y\) such that for all \(x \in X\), \(F(x, 0) = f_0(x)\) and \(F(x, 1) = f_1(x)\) (if \(f_0\) may be “continuously transformed” to \(f_1\)). We say that \(f_0\) is homotopic to \(f_1\) via \(F\). The notion of relative homotopy is also of interest; if \(f_0, f_1 : X \mapsto Y\) are homotopic via \(F, W \subseteq X\), and \(F(x, t) = F(x, 0)\) for \(x \in W\), the homotopy is said to be relative to \(W\).

**Lemma 23.** The relation of homotopy is an equivalence relation on any \(\text{Hom}(X, Y)\) in \(\text{Top}\). If \(f_0, f_1 : X \mapsto Y\) are homotopic and \(g_0, g_1 : Y \mapsto Z\) are also, then \(g_0 f_0, g_1 f_1 : X \mapsto Z\) are homotopic.

**Proof:** Given a continuous function \(f\) let \(F(x, t) = f(x)\); then \(f\) is homotopic to itself via \(F\). If \(f_0\) is homotopic to \(f_1\) via \(F\) let \(G(x, t) = F(x, 1 - t)\); then \(f_1\) is homotopic to \(f_0\) via \(G\). If \(f_0\) is homotopic to \(f_1\) via \(F\) and \(f_1\) is homotopic to \(f_2\) via \(G\) let \(H(x, t) = F(x, 2t)\) if \(t \leq 1/2\) and let \(H(x, t) = G(x, 2t - 1)\) if \(t \geq 1/2\); then \(f_2\) is homotopic to \(f_0\) via \(H\). If \(f_0\) is homotopic to \(f_1\) via \(F\) if \(g_0\) is homotopic to \(g_1\) via \(G\) then \(g_0 f_0\) is homotopic to \(g_0 f_1\) via \(g_0 F\) and \(g_0 f_1\) is homotopic to \(g_1 f_1\) via \(G(f_1(x), t)\).

**Theorem 24.** There is a category \(\text{TopH}\) whose objects are the topological spaces and whose arrows are the equivalence classes of continuous functions under homotopy.

**Proof:** By the lemma \([f] \circ [g]\) may be defined to be \([f \circ g]\); it follows easily that \(\circ\) is associative and \([i_X]\) is the identity for \(X\).
TopH is an example of a category which is not concrete. One reason for its importance is that functors of interest on Top often induce a functor on TopH.

If \( p, q \) are paths in a topological space \( X \) with \( q(0) = p(1) \), their concatenation is defined to be the path \( r \) where where \( r(t) = p(2t) \) if \( t \leq 1/2 \) and \( r(t) = q(2t - 1) \) if \( t \geq 1/2 \). Letting \( pq \) denote the composition, by arguments as in lemma 23 one readily verifies the following.

- Homotopy of paths relative to the endpoints is an equivalence relation.
- If \( p' \) is homotopic to \( p \) and \( q' \) is homotopic to \( q \) then \( p'q' \) is homotopic to \( pq \).

**Theorem 25.** There is a functor from Top to Grpd where if \( G_X \) is the groupoid for \( X \), the points of \( G_x \) are the points of \( X \), and the arrows from \( x \) to \( y \) are the equivalence classes of paths from \( x \) to \( y \), under the relation of homotopy relative to the endpoints. The groupoid morphism for a continuous function \( f : X \rightarrow Y \) maps the point \( x \in G_X \) to \( f(x) \in G_Y \); the class \([p]\) of paths maps to \([f \circ p]\).

**Proof:** The “composition” of two path classes is the class of the concatenation of any two representatives; note the reversal of order in the notation, \( g \circ f = fg \). The associativity of \( \circ \) in \( G_X \) follows because \((fg)h\) is homotopic to \( f(gh) \), via the function \( H(t, s) \) defined as follows. Let \( a = (s + 1)/4 \) and \( b = (s + 2)/4 \);

\[
H(t, s) = f((t/a)(b-a)) \quad \text{for} \quad 0 \leq t \leq a, \quad H(t, s) = g((t-a)/(b-a)) \quad \text{for} \quad a \leq t \leq b, \quad \text{and} \quad H(t, s) = h((t-b)/(1-b)) \quad \text{for} \quad b \leq t \leq 1.
\]

If \( t = x \) is the constant path from \( x \) to \( f(t) \), and if \( t \) is homotopic to \( f \), via functions left to the reader (e.g. for \( f \) let \( a = (1-2)/2 \)). Thus, \( G_X \) is a category. If \( f \) is a path from \( x \) to \( y \) and \( f' \) its reverse, where \( f'(t) = f(1-t) \), then \( f'f \) is homotopic to \( \iota_x \) via \( H \), where \( H(t, s) = f((1-s)/2) \) if \( t \leq 1/2 \) and \( H(t, s) = f((1-s)(2-2t)) \) if \( t \geq 1/2 \). Similarly \( f'f \) is homotopic to \( \iota_y \). Thus, \( G_X \) is a groupoid. The claims about the groupoid morphism are obvious, noting that if \( p_1 \) is homotopic to \( p_2 \) then \( f_{p_1} \) is homotopic to \( f_{p_2} \).

The groupoid \( G_X \) is called the “fundamental groupoid” of \( X \). It is readily verified that in a groupoid, for any object \( x \) \( \hom(x,x) \) is a group with the operation of composition. If there is a morphism from \( x \) to \( y \) then the groups \( \hom(x,x) \) and \( \hom(y,y) \) are isomorphic. In particular for any \( x \in X \) the path classes from \( x \) to \( x \) form a group, called the fundamental group of \( X \) at \( x \); and points in the same path component of \( X \) have isomorphic fundamental groups.

It is convenient to introduce the category \( \text{Top}_* \) of “pointed” topological spaces, where an object is a pair \( \langle X, x \rangle \) where \( X \) is a topological space and \( x \in X \); and a morphism from \( \langle X, x \rangle \) to \( \langle Y, y \rangle \) is a continuous function \( f : X \rightarrow Y \) such that \( f(x) = y \). The map from a pointed space to the fundamental group at the point is the object map of a functor from \( \text{Top}_* \) to Grp.

By facts already proved, if \( f_0, f_1 : \langle X, x \rangle \rightarrow \langle Y, y \rangle \) are homotopic then their images under the fundamental group functor are the same. Letting \( \text{TopH}_* \) be the category where the arrows are homotopy classes, the fundamental group functor can be defined on this category. Two pointed topological spaces are said to be of the same homotopy type if there is an equivalence between them in \( \text{TopH}_* \); in this case they have isomorphic fundamental groups. For example, a space is called contractible if has the same homotopy type as a one point space; in this case its fundamental group is trivial. \( R^n \) is an example of a contractible space.

A path connected topological space is called simply connected if its fundamental group is trivial.

Given a pointed space \( \langle X, x \rangle \), the loop space \( \Omega \) is defined to be the maps from \( I \) to \( X \), which map 0 and 1 to \( x \), equipped with the compact-open topology. By what we have shown, equipping \( \Omega \) with composition of paths makes \( \Omega \) a group object in \( \text{TopH} \). The constant path may be taken as a base point, making \( \Omega \) a pointed space. \( \Omega \) is the object map of a functor from \( \text{Top}_* \) to \( \text{Top}_* \), where for \( f : X_1 \rightarrow X_2 \), \( \Omega(f) \) maps \( g \) to \( fg \).

In a category with finite products, let \( G \) be an object and suppose \( m \in \hom(G \times G, G) \). Define \( \cdot_H : \hom(H,G) \times \hom(H,G) \rightarrow \hom(H,G) \) as \( f_1 \cdot_H f_2 = m\langle f_1, f_2 \rangle \) where \( \langle f_1, f_2 \rangle \) is the induced map to \( G \times G \). Similarly to Yoneda’s lemma, the \( \cdot_H \) are the components of a natural transformation from \( K \) to
Hom(−, G) where \( K(H) = \text{Hom}(H, G) \times \text{Hom}(H, G) \); and given such a natural transformation there is a unique \( m \) such that it arises in this way, namely \( \pi_1 \circ \cdot \pi_2 \) where \( \pi_1, \pi_2 \) are the projections. Indeed, since \( \langle \pi_1, \pi_2 \rangle = \iota_{G \times G} \) we must have \( \pi_1 \circ \cdot \pi_2 = m \cdot m' = m \); remaining details are left to the reader.

A morphism between groups in a category is defined to be a morphism \( f : G_1 \to G_2 \) between the objects, such that the following diagram is commutative. In set this is exactly the same as a group homomorphism.

### Theorem 26
Suppose \( C \) is a category with finite products.

a. Suppose \( G \) is a group object in \( C \), with multiplication \( m \). For each object \( H \) of \( C \) define an operation on \( \text{Hom}(H, G) \) where \( f_1 \cdot f_2 = m(f_1, f_2) \), where \( \langle f_1, f_2 \rangle : H \to G \times G \) is the induced map. With this operation \( \text{Hom}(H, G) \) is a group; and the map \( \text{Hom}(H_2, G) \to \text{Hom}(H_1, G) \) induced by a morphism \( H_1 \to H_2 \) is a group homomorphism.

b. Given group structures on the sets \( \text{Hom}(H, G) \) for each \( H \), which satisfy the conclusions of part 1, there is a unique multiplication \( m \) on \( G \) producing these as in part a.

c. If \( G_1, G_2 \) are group objects a morphism \( f : G_1 \to G_2 \) is a group morphism iff the induced natural transformation from \( \text{Hom}(−, G_1) \) to \( \text{Hom}(−, G_2) \) (enriched with the group operation) has group homomorphisms as components.

The dual statements (that a cogroup structure on \( G \) corresponds to a system of group structures on the sets \( \text{Hom}(G, −) \)) hold for cogroups.

**Proof:** For part a, \( (f_1 \cdot f_2) \times f_3 = (m \times \iota)(f_1, f_2, f_3) \) and \( f_1 \times (f_2 \cdot f_3) = (\iota \times m)(f_1, f_2, f_3) \). Associativity of \( \cdot \) follows from that of \( m \). Let \( T \) be a terminal object, and let \( \bar{e} = e_\tau \) where \( \tau : H \to T \). Then \( f_1 \cdot \bar{e} = m(\iota \times e)(f_1, \tau) \). That \( \bar{e} \) is a right identity follows since \( e \) is, and similarly on the left. That \( if \) is a right inverse follows because \( m(f, if) = m(\iota, i)f = e_\tau f = \bar{e} \). For the last claim, \( (f_2 h, g_2 h) = (f_2, g_2)h \); it follows that \( f_1 \cdot g_1 = (f_2 \cdot g_2)h \). For part b, the hypotheses imply that the \( \cdot_H \) are the components of a natural transformation as above, so \( m = \pi_1 \circ \cdot \pi_2 \) induces the transformation; it remains to show that \( m \) is a multiplication. Considering the natural transformation square for \( m \times \iota \), \( m(m \times \iota) = m \pi_1 \pi_2 \). Considering the square for the canonical equivalence, \( m \pi_1 \pi_2 = m \pi_{12} \pi_3 \), and by another push \( m \pi_{12} \pi_3 = (\pi_1 \pi_2) \pi_3 \). Similarly \( m(m \times \iota)m = \pi_1 \pi_2 \pi_3 \). Associativity of \( m \) thus follows, using that of \( \cdot \). Let \( e_H \) denote the identity of \( \text{Hom}(H, G) \); we claim that \( E_T \) is an identity for \( m \). By hypothesis \( e_H = e_\tau \tau \) where \( \tau \) is the unique map from \( H \) to \( T \). Considering the square for \( \iota \times e_T \), and noting that \( \pi_0(\iota \times e_T) = e_T \pi_2 = e_T \pi = e_{G \times T} \), it follows that \( m(\iota \times e_T) = \pi_1 \). Similarly \( e_\tau \) is an identity on the left. Let \( i \) be the inverse of \( \iota \) in \( \text{Hom}(G, G) \); we claim that \( i \) is an inverse for \( m \). This follows by considering the square for \( \iota \iota \) and showing \( m(\iota, i) = e_G \). For part c, first note that \( (f \times f)(g_1, g_2) = (g_1, g_2)(f \times f) \); this can be seen by considering the cone from \( H \) to the diagram for \( f \times f \). Suppose \( f_m = m_2(f \times f) \); then \( f(g_1 \cdot g_2) = f(g_1, g_2) = m_2(f \times f)(g_1, g_2) = m_2(f g_1, g_2) = f g_1 \cdot f g_2 \). Conversely if \( f(g_1 \cdot g_2) = f g_1 \cdot f g_2 \) then \( f_m = f(\pi_1, \pi_2) = f \pi_1 \cdot f \pi_2 = (\pi_1 f \times f) \cdot (\pi_2 f \times f) = (\pi_1 \cdot \pi_2)(f \times f) = m_2(f \times f) \).

Say that a functor \( F \) to \( D \) is a functor to groups if \( F(X) \) is always a group object in \( D \), and \( F(f) \) a morphism between groups. We have shown that \( G \) is a group object iff \( \text{Hom}(−, G) \) is a functor to groups.

**Corollary 27.** Suppose \( \langle L, R, \alpha \rangle \) is an adjunction. If \( R \) is a functor to groups then \( L \) is a functor to cogroups, and conversely.
Proof: If $R$ is a functor to groups then given $H$, $\text{Hom}(H, R(-))$ is a group, and so $\text{Hom}(L(H), -)$ is. Remaining details are left to the reader.

In Top, let $I^Z$ denote the $\text{Hom}(I, Z)$ equipped with the compact-open topology. By corollary 21.35, the correspondence $g \mapsto \tilde{g}$ where $g(x, y) = \tilde{g}(x)(y)$ maps $\text{Hom}(X \times I, Z)$ bijectively to $\text{Hom}(X, Z^I)$, yielding an adjunction between $- \times I$ and $-^I$. In $\text{Top}_*$, only maps in $h \in Z^I$ where $h(0) = h(1) = z$ are considered, and $\Omega(Z)$ is a subspace of $Z^I$. In addition under a map of $\text{Hom}(X, \Omega(Z))$ the base point $x_0$ of $X$ maps to $t \mapsto z_0$ where $z_0$ is the base point of $Z$; thus, such a map corresponds to a map $g$ where $g(x, y) = \tilde{g}(x)(y)$. Thus $X \times I$ map be replaced by the space $S(X)$, which is the quotient of $X \times I$ with $x_0 \times I, X \times 0$, and $X \times 1$ identified. $S$ is the object map of a functor (the “suspension functor”) from $\text{Top}_*$ to $\text{Top}_*$, where for $f : Z_1 \to Z_2, S(f)$ is the map induced on the quotients by $f \times \iota$.

The functor $\Omega$ on $\text{Top}_*$ induces a functor on $\text{TopH}_*$. That is, if $f_1$ and $f_2$ are homotopic then $\Omega(f_1)$ and $\Omega(f_2)$ are homotopic; this follows readily using lemma 23. We use $\Omega$ to denote this functor as well. Similarly the suspension functor induces a functor on $\text{TopH}_*$, also denoted $S$, because products of homotopic maps are homotopic. Finally, the reader may verify that the adjunction between $\Omega$ and $S$ in $\text{Top}_*$ induces an adjunction in $\text{TopH}_*$.

**Theorem 28.** In $\text{TopH}_*$ $\Omega(X)$ is a group object. Also, $S(X)$ is a cogroup object.

Proof: The proof of the first claim is left to the reader; or see [Spanier]. The second claim follows by corollary 27.

We state the following without proof; proofs may be found in [Spanier]. Recall from chapter 17 that $S^n$ is used to denote the $n$th homotopy group functor.
- $S^{n+1}$ is homeomorphic to $S(S^n)$.
- $\pi_1$ is naturally equivalent to the fundamental group functor.
- $\pi_n$ is a functor to groups.
- If $n \geq 2$ then $\pi_n$ is a functor to Abelian groups.

**14. Homology of simplicial complexes.** In section 19.4 the homology of abstract simplicial complexes was discussed. This provides a homology functor for certain topological spaces by equipping the abstract simplicial complexes with a topology in a natural manner. This homology functor has properties which allow deriving important basic theorems about such spaces.

Suppose $S$ is an abstract simplex (i.e., a finite set). Let $\bar{S}$ be the set of formal linear combinations $\sum_{v \in S} \alpha_v v$, where $S$ is a simplex, $\alpha_v \in [0, 1]$, and $\sum_{v \in S} \alpha_v = 1$. If the vertices of $S$ are instantiated as points in some Euclidean space, then $\bar{S}$ is the convex hull; however the abstract definition is more general. If $K$ is an abstract simplicial complex then $\bar{K}$ equals the union of the $\bar{S}$ over the simplexes $S$ of $K$.

For $S$ with $|S| = n$, choosing linearly independent points in $\mathcal{R}^n$ for the vertices $\nu \in S$ yields an identification of $\bar{S}$ with a subset of $\mathcal{R}^n$, by which $\bar{S}$ may be given a topology; indeed it is a metric space with metric $\sqrt{\sum_{v \in S}(\alpha_v - \beta_v)^2}$. The collection $\{\bar{S}\}$ satisfies the hypotheses of theorem 17.5, yielding a topology on $\bar{K}$ coherent with those of the $\bar{S}$, which are closed subspaces.

We will call a space of the form $\bar{K}$, where $K$ is an abstract simplicial complex, a simplicial complex, although usage varies. Many authors require a simplicial complex to be given as a subset of some Euclidean space. Although we will not prove it, a simplicial complex may be given as a subspace of some Euclidean space iff it is countable, the simplexes have bounded dimension, and each vertex belongs to only finitely many simplexes (see [Spanier]).

The map $K \mapsto \bar{K}$ is readily seen to be the object map of a functor from ASC to Top. A simplicial map $f : K \to L$ induces a continuous function $\tilde{f} : \bar{K} \to \bar{L}$ by linearity, that is, $\tilde{f}(\sum_i \alpha_v v) = \sum_i \alpha_v f(v)$. This is
clearly continuous on each simplex, and hence continuous.

In many important applications $K$ is finite. In this case an embedding of $\bar{K}$ in $R^n$ may readily be given, where $n$ is the cardinality of $V$; simply map each $v \in V$ to a distinct standard unit vector, and extend by linearity. It is clear that $\bar{K}$ is compact. Conversely if $\bar{K}$ is compact then $\bar{C}$ is finite (again see [Spanier]).

We will mostly be concerned with finite simplicial complexes, although some general facts will be given. Examples of topological spaces which are homeomorphic to a finite simplicial complex include the closed unit ball, the sphere $S^n$, and the torus (which can be given either as an embedding of a 2-manifold in $R^3$, or as a quotient space of the closed square in $R^2$). According to [Hatcher], a triangulation (homeomorphism to a simplicial complex) of the torus requires at least 14 triangles; and a generalization of simplicial complexes (Δ-complexes) allows reducing this to 2. Simplicial complexes are by far more commonly used in developing the theory; also, ad-hoc methods for improving efficiency can be considered subsequently.

For a vertex $v$ let $D_v = \{ \alpha \in K : \alpha_v > 0 \}$. If $S$ is a simplex then $S \cap D_v$ is open in $S$, and since $\bar{K}$ has the coherent topology $D_v$ is open.

**Lemma 29.** A simplicial complex is paracompact and Hausdorff.

**Proof:** Suppose $C$ is a simplicial complex with vertex set $V$. Suppose $K$ is a cover of $\bar{K}$, and let $K' = \{ U \in D_v : U \in K, v \in V \}$. Then $K'$ is a cover refining $K$. Further it is locally finite, since for $p = \sum_v \alpha_v v, p \in D_v$ only if $\alpha_v > 0$. If $p_1, p_2$ are distinct points in $\bar{K}$ they may be considered as vectors in $R^l$ for a suitable $l$. Let $U_i$ for $i = 1, 2$ be open sets in $R^l$ separating them. Let $V_i = \cup_{\sigma \in C} U_i \cap S$. Then $V_i, i = 1, 2$, are open sets in $\bar{K}$ separating $p_1$ and $p_2$.

By theorem 17.13, a simplicial complex is normal.

Let $N_n$ denote $\{0, \ldots , n\}$. Then serves as a “standard” $n$-simplex. A singular simplex in a topological space $X$ is defined to be a continuous function $\sigma : N_n \mapsto X$ for some $n$. We let $\bar{C}_n(X)$ be the free Abelian group generated by the $n$-simplexes. As remarked in section 19.4, the free $R$-module may be considered for a commutative ring $R$; and this may be viewed as a tensor product. Also, $\bar{C}_n(X) = 0$ for $n < 0$.

For $0 \leq i \leq n$ let $\mu^n_i$ denote the map from $\bar{S}_{n-1}$ to $\bar{N}_n$ induced by the map $j \mapsto j$ if $j < i$, and $j \mapsto j + 1$ if $j \geq i$. This maps the standard $n-1$-simplex to a face of the standard $n$-simplex, namely the one missing vertex $i$. Define the group homomorphism $\bar{\partial}_n : \bar{C}_n(X) \mapsto \bar{C}_{n-1}(X)$ by its action on the generators, namely

$$\bar{\partial}_n(\sigma) = \sum_{i=0}^k (-1)^i \sigma \mu^n_i.$$ 

It is readily seen that $\bar{\partial}_{n-1}(\bar{\partial}_n) = 0$ (each $n-2$ face occurs twice, with opposite sign).

Thus, for $X$ in Top the “singular chain complex” $\bar{C}(X)$ in Ab-Ch has been defined. This is the object map of a functor. A continuous function $f : X \mapsto Y$ induces a map $\sigma \mapsto f\sigma$ on the basis elements, and hence from $\bar{C}_n(X)$ to $\bar{C}_n(Y)$ by linearity. These maps in turn yield a chain map from $\bar{C}(X)$ to $\bar{C}(Y)$; indeed $\bar{\partial}_n(\sigma)$ maps to $\sum_{i=0}^k (-1)^i f\sigma \mu^n_i$ which equals $\bar{\partial}_n(f\sigma)$. In particular, homeomorphic topological spaces have isomorphic singular chain complexes.

The homology functor $\bar{H}$ is derived from the chain functor $\bar{C}$ by the general machinery of chapter 19. In the remainder of the section a proof will be given that $H_n(K)$ and $\bar{H}_n(K)$ are naturally isomorphic. In particular, different triangulations of a space yield the same homology groups. This is clearly a very useful fact. We omit a discussion of its basic consequences, which can be found in [Spanier] or any other introductory algebraic topology text. [Wallace] is a comprehensive introduction which does not use category theory.

An outline of the proof is helpful.

1. Augmentation and reduced complexes are defined.
2. Ordered chain complexes are defined.
3. The reduced ordered chain complex of a simplex is shown to be acyclic.
4. A natural transformation from the ordered to simplicial chains is given, and shown to induce an equivalence on the homology chains.
5. The reduced singular chain complex of a simplex is shown to be acyclic.
6. A natural transformation $\theta$ from the ordered to singular chains is given, and shown to induce an equivalence on the homology chains for a simplex.
7. Barycentric subdivision is defined, and properties proved.
8. The excision theorem (lemma 38) is proved.
9. The homotopy theorem (lemma 40) is proved.
10. A strengthened version of the excision theorem is proved.
11. The natural transformation $\theta$ is shown to induce an equivalence on the homology chains for any finite complex.

If $C$ is a chain complex an augmentation is a map $\epsilon : C_0 \rightarrow \mathbb{Z}$ such that $\epsilon \partial_1 = 0$ (this is a special case of a notion encountered in chapter 19). The reduced chain complex $C^r$ is defined to have $C^r_0 = \text{Ker} (\epsilon)$, and $C^r_n = C_n$ for $n > 0$.

**Lemma 30.** If $\epsilon$ is an augmentation for a chain complex $C$ then $H(C)$ is naturally equivalent to the chain obtained from $H(C^r)$ by replacing $H_0(C^r)$ by $H_0(C^r) \oplus \mathbb{Z}$.

**Proof:** Since $C_0$ is a free group, $C_0$ is isomorphic to $C_0^r \oplus \mathbb{Z}$. Noting that $Z^r_n = Z_n$ for $n > 0$, and $B^r_n = B_n$ for all $n$, the lemma follows.

If $\epsilon$ ($\epsilon'$) is an augmentation for $C$ ($C'$) a chain map $\tau : C \rightarrow C'$ is said to preserve augmentation if $\epsilon \tau = \epsilon$. In this case, there is an induced chain map $\tau^r : C^r \rightarrow C'^r$. Further, under the isomorphisms from $C_0$ to $C_0^r \oplus \mathbb{Z}$ and $C_0'$ to $C_0'^r \oplus \mathbb{Z}$, the induced map $C_0^r \oplus \mathbb{Z} \rightarrow C_0'^r \oplus \mathbb{Z}$ may be taken as $\tau^r \oplus \epsilon_Z$.

Define the chain functor $\hat{C}$ on ASC, where the basis elements of $\hat{C}_n$ are the “ordered simplexes” $\langle v_0, \ldots, v_n \rangle$. Authors vary on whether repetitions are allowed; we will allow them. The boundary maps are given by

$$\partial_n(\langle v_0, \ldots, v_n \rangle) = \sum_{i=0}^{n} (-1)^i \langle v_0, \ldots, \hat{v}_i, \ldots, v_n \rangle,$$

where the notation $\hat{v}_i$ for deleting $v_i$ is used.

We leave the following to the reader.
- The chain functors $\hat{C}$ and $\bar{C}$ both preserve monics, so that, as discussed in section 19.3, the relative homology chains $\tilde{H}(X, X')$ and $\tilde{H}(X, X')$ for $X' \subseteq X$ may be defined.
- The map determined by $\langle v \rangle \mapsto 1$ for any $v$ is an augmentation for $\hat{C}(K)$, for any nonempty $K$ in ASC. For a nonempty $X$ in Top, let $\sigma_v \in C_0(X)$ be the map taking 0 to $v$. The map determined by $\sigma_v \mapsto 1$ is an augmentation for $\tilde{C}(X)$.

**Lemma 31.** If $K$ is the complex of a simplex (i.e., the subsets of some nonempty finite set $S$) then $\hat{C}^r(K)$ is chain contractible.

**Proof:** Choose $u \in K$, and let $h_n(\langle v_0, \ldots, v_n \rangle) = \langle u, v_0, \ldots, v_n \rangle$. One verifies that for $n > 0$, $\partial_n h_n + h_{n+1} \partial_n = \iota$. The kernel of the augmentation is those chains $c = \sum_i n_i \langle v_i \rangle$ where $\sum_i n_i = 0$. For such chains, $\partial_1 h_0(c) = c$.

As observed in section 19.3, it follows that $\tilde{H}_n^r(K) = 0$, whence $\tilde{H}_0(K) = \mathbb{Z}$ and $\tilde{H}_n(K) = 0$ for $n > 0$. The maps $\langle v_0, \ldots, v_n \rangle \mapsto [v_0, \ldots, v_n]$ are clearly the components of a chain map $\eta$ from $\hat{C}$ to $C$, where $C$ is the oriented chain functor to Ab-Ch introduced in section 19.4.
**Lemma 32.** $H(\eta)$ is a natural equivalence.

**Proof:** It suffices to show that $H(\eta)$ is an equivalence for any $K$. Well-order the vertices of $K$. For brevity let $\nu$ denote the sequence $\langle v_0, \ldots, v_n \rangle$. Let $\zeta$ be the chain map from $C(K')$ to $\hat{C}(K)$, mapping $[v_0, \ldots, v_n]$ to $\langle v_{\pi(0)}, \ldots, v_{\pi(n)} \rangle$ where $\pi$ is the permutation reordering $\nu$ to agree with the well-order; and the sign is $sg(\pi)$. Then $\eta_\nu = \nu$, whence $H_n(\eta)H_n(\zeta) = \nu$. It suffices to show that $\zeta\eta$ is chain homotopic to $\nu$; for then $H_n(\zeta)H_n(\eta) = H_n(\zeta\eta) = \nu$ by theorem 19.10. Maps $h_n : \tilde{C}n \mapsto \tilde{C}_{n+1}$ must be constructed, such that $\zeta - \nu = \tilde{d}_{n+1}h_n + h_n - 1\tilde{d}_n$. We will also ensure that in $h_n(\nu)$, all ordered simplexes involve only $v_0, \ldots, v_n$. It is readily verified that $h_0$ may be taken as $0$. For $n > 0$

$$\tilde{d}_n(\zeta_{\nu} - \nu - h_{n-1}\tilde{d}_n) = \tilde{d}_n\zeta_{\nu} - \tilde{d}_n - (\zeta_{n-1}\nu_{\nu-1} - \nu - h_{n-2}\tilde{d}_{n-1})\tilde{d}_n = 0.$$  

It follows from lemma 31 that $(\zeta_{\nu} - \nu - h_{n-1}\tilde{d}_n)(\nu)$ is the image of a chain of the required form; $h_n(\nu)$ may be chosen as any such.

Exercise 15 gives an alternative proof of lemma 32. We remark that many authors use $\eta$, to denote $H(\eta)$ for a chain map $\eta$. We have used $H(\eta)$ to keep the notation to a minimum. Confusion with the use of $H(X)$ for $H(C(X))$ is avoided, because in the latter case $X$ is an object in the category of interest, in this case Top.

A subspace $X$ of Euclidean space $\mathbb{R}^k$ for some $k$ is said to be star-shaped from the origin. If $K$ is the complex of a simplex then $\tilde{K}$ (considered as embedded in Euclidean space) is convex, hence star-shaped from any point of $\tilde{K}$.

**Lemma 33.** If $X$ is star-shaped then $\tilde{C}^r(X)$ is chain contractible.

**Proof:** W.l.o.g. $X$ may be assumed to be starlike from the origin. A singular $n$-simplex may be written as a function $\sigma(t_0, \ldots, t_n)$ where $0 \leq t_i \leq 1$ and $\sum t_i = 1$. Given such, define $\sigma' = h_n(\sigma)$ to satisfy $\sigma'(t_0, 1, \ldots, t_{n+1}) = (1 - t_0)\sigma(t_1/(1 - t_0), \ldots, t_{n+1}/(1 - t_0))$ if $t_0 < 1$, and $\sigma'(1, 0, \ldots, 0) = 0$. By the hypotheses $\sigma' \in \tilde{C}_{n+1}(X)$. The remainder of the proof is similar to that of lemma 31.

Given a sequence $\nu = \langle v_0, \ldots, v_n \rangle$ of vertices of a simplex of a simplicial complex $K$, let $\sigma_{\nu} : \tilde{N}_n \mapsto K$ be the linear map induced by the map $i \mapsto v_i$ for $i \in N_n$. The map $\nu \mapsto \sigma_{\nu}$ is readily verified to induce a natural transformation $\theta$ from the functor $\tilde{C}(K)$ from ASC to Ab-Ch, to the composite functor $\tilde{C}(\{K\})$. Further, $\theta$ is readily verified to preserve augmentation.

**Lemma 34.** If $K$ is the complex of a simplex then $H(\theta) : \tilde{H}(K) \mapsto \tilde{H}(\tilde{K})$ is an isomorphism.

**Proof:** By remarks following lemma 30, $H(\theta^r) \otimes \mathbb{Z}$ is the map from $\tilde{H}^r(K) \otimes \mathbb{Z}$ to $\tilde{H}^r(\tilde{K}) \otimes \mathbb{Z}$ induced by the isomorphisms. Since $\tilde{C}^r(K)$ and $\tilde{C}^r(\tilde{K})$ are both acyclic, the lemma follows.

In accordance with the definition of ASC, a subcomplex $L$ of an abstract simplicial complex $K$ is a subset of $K$; its vertex set is $\cup L$.

**Lemma 35.** Suppose $K_1$ and $K_2$ are subcomplexes of an abstract simplicial complex $K$.

a. $\tilde{C}(K_1 \cap K_2) = \tilde{C}(K_1) \cap \tilde{C}(K_2)$.

b. $\tilde{C}(K_1 \cup K_2) = \tilde{C}(K_1) + \tilde{C}(K_2)$.

c. $\tilde{C}(K_1 \cup K_2)/\tilde{C}(K_2)$ is isomorphic to $\tilde{C}(K_1)/\tilde{C}(K_1 \cap K_2)$.

d. $\tilde{H}(K_1 \cup K_2, K_2)$ is isomorphic to $\tilde{H}(K_1, K_1 \cap K_2)$.

**Proof:** From $K_1 \cap K_2 \subseteq K_1, K_2 \subseteq K_1 \cup K_2$ and the fact that $\tilde{C}$ respects inclusion, $\tilde{C}(K_1 \cap K_2) \subseteq \tilde{C}(K_1 \cap \tilde{C}(K_2)$ and $\tilde{C}(K_1) + \tilde{C}(K_2) \subseteq \tilde{C}(K_1 \cup K_2)$. If $\nu \in C(K_1)$ and $\nu \in C(K_2)$ then $\nu \in C(K_1 \cap C(K_2)$ If $\nu \in C(K_1 \cup K_2)$ then $\nu \in \tilde{C}(K_1)$ or $\nu \in \tilde{C}(K_2)$. Part c is an immediate consequence of the second Noether isomorphism theorem (section 8.1). Part d follows immediately from the definition.
It is not true in general that for subspaces $X_1, X_2$ of a topological space, $\hat{H}(X_1 \cup X_2, X_2)$ is isomorphic to $\hat{H}(X_1, X_1 \cap X_2)$ (see [Spanier]). A sufficient condition is given in lemma 38; the notion of barycentric subdivision is used in the usual proof of this.

If $\{v_0, \ldots, v_n\}$ is an abstract simplex the barycenter of its topological space is defined to be $\sum (1/(n + 1))v_i$. Given an abstract simplicial complex $K$, let $K_\beta$ be the abstract simplicial complex defined as follows. The vertices are the barycenters $v_S$ of the simplexes $S$ of $K$. A simplex of $K_\beta$ is a set $\{v_{S_0}, \ldots, v_{S_n}\}$, where $|S_i| = i + 1$ and the $S_i$ form an ascending chain. Figure 1 gives an example (the node label encodes the subset in binary notation).

![Figure 1](image)

A vertex of $K_\beta$ may be mapped to a corresponding point of $\tilde{K}$ in an obvious manner, namely $\sum_i \alpha_i v_{S_i}$ maps to $\sum \alpha_i \sum_{w \in S_i} (1/(|S_i| + 1))w$. Extending by linearity yields a triangulation of $\tilde{K}$, such that the image of each simplex of $K_\beta$ is a subset of a simplex of $\tilde{K}$.

**Lemma 36.** If $K$ is an $n$-simplex in Euclidean space, $L$ is an $n$-simplex of its barycentric subdivision, and $d_K$ ($d_L$) is the diameter of $K$ ($L$) then $d_L \leq (n/(n + 1))d_K$.

**Proof:** Since $|v - \sum_i t_i v_i| \leq \sum_i t_i |v - v_i| \leq \max |v - v_i|$, $d_L$ equals the maximum distance between two vertices of $L$. Given two such, $v_S$ and $v_T$, where $S \subset T$ and $|T| = t$, $|v_S - v_T| \leq \max_{e \in S} |v_i - v_T|$. Also, $|v_i - \sum_{j \in T} v_j| \leq (1/(t + 1)) \max_{j \in T} |v_i - v_j|$. Thus, $|v_S - v_T| \leq (t/(t + 1))d_S$, and $(t/(t + 1)) \leq (n/(n + 1))$.

The barycentric subdivision may be used to define a natural transformation $\beta$ from $\bar{C}$ to $C$. If $\nu = \langle v_0, \ldots, v_n \rangle$ where $v_i \in \bar{N}_n$ let $\lambda_\nu : \bar{N}_n \mapsto \bar{N}_n$ be the linear map induced by $i \mapsto v_i$. For a simplex $\sigma : \bar{N}_n \mapsto X$, chains will be considered of the form $\bar{C}(\sigma)(c)$ where $c$ is a chain of some $\lambda_\nu$'s; this has $\sigma$ composed on the left with each term in the chain of $\lambda_\nu$'s. The notation is simplified by letting $\nu$ denote indifferently $\lambda_\nu$. Note that $\bar{\partial}$ may be expressed as $\bar{\partial}(v_0, \ldots, v_n) = \sum_i (-1)^i \langle v_0, \ldots, \hat{i}, \ldots, v_n \rangle$.

We introduce the following further notation.

- Let $v \oplus \nu$ denote $\langle v, v_0, \ldots, v_n \rangle$, and extend the notation to chains as usual (i.e., by applying the operation to the generators).

- For $\nu = \langle v_0, \ldots, v_n \rangle$ let $v_\nu$ denote $v_{\langle v_0, \ldots, v_n \rangle}$.

- Let $\iota$ denote $\langle 0, \ldots, n \rangle$, the identity map on $\bar{N}_n$.

- Let $\sigma \cdot c$ denote $\bar{C}(\sigma)(c)$, $c$ composed with $\sigma$ on the left.

We first define $\beta_n$ for linear chains to $\bar{N}_n$ by recursion on $n$, by letting $\beta_0(\nu) = \nu$, and for $n > 0 \beta_n(\nu) = v_\nu \oplus \beta_{n-1}(\bar{\partial}_n(\nu))$. For $\sigma \in \bar{C}_n(X)$ let $\beta_n(\sigma) = \sigma \cdot \beta_n(\iota)$. Define also the map $\gamma_n : \bar{C}_n(X) \mapsto \bar{C}_{n+1}(X)$ by recursion on $n$, by letting $\gamma_0(\langle v \rangle) = \langle v, v \rangle$ and for $n > 0 \gamma_n(\nu) = v_\nu \oplus \langle v - \gamma_n - 1(\bar{\partial}_n(\nu)) \rangle$. This chain is a “prism” from $\nu$ to $\beta_n(\nu)$, as the next lemma shows. Finally for $\sigma \in \bar{C}(X)$ let $\gamma_n(\sigma) = \sigma \cdot \gamma_n(\iota)$.

**Lemma 37.**

a. The maps $\beta_n : \bar{C}_n(X) \mapsto \bar{C}_n(X)$ are the components of a chain map.
b. For \( f : X \to Y \), \( \bar{C}_n(f) \) commutes with \( \beta \), so that the chain maps of part a are the components of a natural transformation from \( \bar{C} \) to \( \bar{C} \).

c. The maps \( \gamma_n : \bar{C}_n(X) \to \bar{C}_{n+1}(X) \) are the components of a chain homotopy from \( \iota \) to \( \beta \).

d. For \( f : X \to Y \), \( \bar{C}_n(f) \) commutes with \( \gamma \).

e. \( \gamma^{(m)} = \sum_{i<m} \gamma^i \beta^i \) is a chain homotopy from \( \iota \) to \( \beta^m \).

**Proof:** We first prove part a for linear chains to \( \bar{N}_n \), by induction on \( n \). For \( n = 0 \), \( \bar{\partial}\beta(\nu) = \bar{\partial}(\nu) = 0 \) and \( \beta\bar{\partial}(\nu) = \beta(0) = 0 \). For \( n > 0 \),

\[
\bar{\partial}\beta(\nu) = \bar{\partial}(v_\nu \otimes \bar{\partial}(\nu)) = \bar{\partial}(\nu) - v_\nu \otimes \bar{\partial}\beta(\nu) = \beta\bar{\partial}(\nu) - v_\nu \otimes \bar{\partial}\beta(\nu) = \beta\bar{\partial}(\nu).
\]

We already know that for linear chains to \( \bar{N}_n \), \( \bar{\partial}(\sigma \cdot \iota) = \sigma \cdot \bar{\partial}(\iota) \). We claim that \( \beta(\sigma \cdot \iota) = \sigma \cdot \beta(\iota) \); indeed \( \beta(\sigma \cdot \nu) = \sigma \cdot \nu \cdot \beta(\iota) = \sigma \cdot \beta(\nu) \).

Part a follows, since

\[
\bar{\partial}\beta(\sigma) = \bar{\partial}(\beta(\sigma)) = \partial(\sigma \cdot \beta(\iota)) = \sigma \cdot \bar{\partial}(\beta(\iota)) = \beta(\sigma \cdot \bar{\partial}(\iota)) = \beta(\partial(\sigma)) = \beta(\bar{\partial}(\sigma)).
\]

For part b, suppose \( f : X \to Y \); then \( \bar{C}(f)(\beta(\sigma)) = f \cdot \beta(\iota) = \beta(\bar{C}(f)(\sigma)) \). We first prove part c for linear chains to \( \bar{N}_n \), by induction on \( n \). For \( n = 0 \), \( \beta = \iota \) and \( \bar{\partial}\gamma = 0 \), so \( \iota - \beta = \iota + \gamma \bar{\partial} \) follows. For \( n > 0 \), first, using the induction hypothesis, \( \bar{\partial}(\iota - \gamma \bar{\partial}) = \bar{\partial} \iota - \gamma \bar{\partial} \bar{\partial} = \bar{\partial}(\iota - \beta - \gamma \bar{\partial} \bar{\partial}) = \beta \bar{\partial} \). Then \( \bar{\partial}\gamma(\nu) = (\bar{\partial}(v_\nu \otimes (\nu - \gamma \bar{\partial}(\nu))) = \nu - \gamma \bar{\partial}(\nu) - v_\nu \otimes \bar{\partial}(\iota - \gamma \bar{\partial}(\nu)) = \nu - \gamma \bar{\partial}(\nu) - v_\nu \otimes \bar{\partial}\beta(\nu) = \nu - \gamma \bar{\partial}(\nu) - \beta(\nu) \), the last step following by the definition of \( \beta \), \( \beta(\sigma) \) follows in general, similarly to part a. Part d follows similarly to part b. For part e, \( \bar{\partial}\gamma^{(m)} + (\gamma^{(m)})\bar{\partial} = \sum_{i<m}(\bar{\partial}\gamma^i \beta^i + \gamma^i \bar{\partial}\beta^i) = \sum_{i<m}(\bar{\partial}\gamma^i + \gamma^i \bar{\partial})\beta^i = \sum_{i<m}(\iota - \beta)^{\beta^i} = \iota - \beta^m \).

Recall the notation \( \langle X, X' \rangle \) for topological pairs introduced in chapter 17. The notation \( \langle Y, Y' \rangle \subseteq \langle X, X' \rangle \) denotes \( Y \subseteq X \) and \( Y' \subseteq X' \). There are two different ways of writing certain inclusions of pairs, (1) \( \langle X_2, X_1 \cap X_2 \rangle \subseteq \langle X_1 \cup X_2, X_1 \rangle \) and (2) \( \langle X - Z, Y - Z \rangle \subseteq \langle X, Y \rangle \) for \( Z \subseteq Y \subseteq X \). Given an inclusion (1), letting \( X = X_1 \cup X_2, Y = X_1 \), and \( Z = X_2 \) (meaning \( X - X_2 \)) yields an inclusion (2). On the other hand, given an inclusion (2), letting \( X_1 = X \) and \( X_2 = Z^c \) yields an inclusion (1).

Assuming the above to be the case, using exercise 17.3 one verifies that \( X_1^{\text{int}} \cup X_2^{\text{int}} = X_1 \cup X_2 \) (where the interiors are taken in \( X_1 \cup X_2 \)) if \( Z^\text{cl} \subseteq Y^{\text{int}} \). Also, (1) \( \bar{H}(X, Y) \) is isomorphic to \( \bar{H}(X - Z, Y - Z) \) if \( (2) \bar{H}(X_1 \cup X_2, X_1) \) is isomorphic to \( H(X_1 \cup X_2, X_1) \). Various terminology is used in conjunction with the above facts. If (1) holds we say that \( Z \) can be excised. If (2) holds then \( \{X_1, X_2\} \) is called an exciscive couple.

**Lemma 38.**

a. For subspaces \( Z \subseteq Y \subseteq X \), if \( Z^{\text{cl}} \subseteq Y^{\text{int}} \) (in \( X \)), then \( \bar{H}(X, Y) \) is isomorphic to \( \bar{H}(X - Z, Y - Z) \).

b. For subspaces \( X_1, X_2 \), if \( X_1 \cup X_2 = X_1^{\text{int}} \cup X_2^{\text{int}} \) (in \( X_1 \cup X_2 \)) then \( \bar{H}(X_1 \cup X_2, X_1) \) is isomorphic to \( H(X_1, X_1) \cap X_2 \).

**Proof:** There is a map \( k : \bar{C}(X - Z)/\bar{C}(Y - Z) \to \bar{C}(X)/\bar{C}(Y) \) induced by the inclusions, namely \( k(c + \bar{C}(Y - Z)) = c + \bar{C}(Y) \). We claim that under the hypotheses, \( H(k) \) is an isomorphism. One verifies that \( H(k) \) acts by mapping \( c + B \) (where \( c \in \bar{C}(X - Z) \), \( c \in \bar{C}(Y - Z) \) and \( B = \bar{C}(X - Z) + \bar{C}(Y - Z) \)) to \( c + B' \) (where \( B = \bar{C}(X) + \bar{C}(Y) \)). Suppose \( \sigma \) is an \( n \)-simplex in \( X \); then \( \sigma^{-1}[X^{\text{int}}] \) is an open cover of \( \bar{N}_n \) (where \( X_1 = Y, X_2 = Z^c \)). Let \( \epsilon \) be a Lebesgue number for the cover. Let \( m \) be less such that \( (n/(n + 1))^m d < \epsilon \) where \( d \) is the diameter of \( \bar{N}_n \). Then \( \beta^m(\sigma) \) is a chain such that the image of every simplex is either in \( Y^{\text{int}} \) or \( (Z^c)^{\text{int}} \). It is readily verified that \( \bar{C}(Y) \) is closed under the action of the maps \( \beta \) and \( \gamma \) defined above. If \( c_1 \in \bar{C}(X) \) and \( \bar{C}(c_1) \in \bar{C}(Y) \) then \( c_1 - \beta(c_1) = \bar{C}(c_1) + \gamma(\bar{C}(c_1)) \), so \( c_1 - c_2 \in \bar{C}(X) + \bar{C}(Y) \). From all this, a chain \( c_2 \) homologous to \( c_1 \) in \( \bar{H}(X, Y) \) can be found, such that each simplex of \( c_2 \) is either in \( Y^{\text{int}} \) or \( (Z^c)^{\text{int}} \). Let \( c_3 \) be the part in \( (Z^c)^{\text{int}} \); then \( c_3 \) is homologous
to $c_2$ in $\tilde{H}(X, Y)$, $c_3 \in \tilde{C}(X - Z)$, and $\partial(c_3) \in \tilde{C}(Y - Z)$. This shows that $H(k)$ is surjective. Suppose $c_1 \in \tilde{C}(X - Z)$, and $c_1 \in B'$, say $c_1 = \partial(c_2) + c_3$ where $c_2 \in \tilde{C}(X)$ and $c_3 \in \tilde{C}(Y)$. As before, we may assume $c_2 = c_4 + c_5$ where $c_4 \in Y^{\text{int}}$ and $c_5 \in (Z^c)^{\text{int}}$. From $c_1 - \partial(c_3) = \partial(c_4) + c_3$ we conclude that both sides are in $\tilde{C}(Y - Z)$. It follows that $c_1 \in B$, which shows that $H(k)$ is injective. Part a is thus proved, and part b follows by remarks preceding the lemma.

**Lemma 39.** For a topological space $X$, and $i = 0, 1$, let $I_i : X \mapsto X \times [0, 1]$ denote the map $(x) \mapsto (x, i)$. The singular chain maps induced by $I_0$ and $I_1$ are chain homotopic.

**Proof:** Given a singular $n$-simplex $\sigma : \tilde{S}_n \mapsto X$, for $0 \leq i \leq n$ let $E_i\sigma : \tilde{S}_{n+1} \mapsto X \times [0, 1]$ be the $n + 1$-simplex induced by the vertex map $j \mapsto (\sigma(j), 0)$ for $0 \leq j \leq i$, and $j \mapsto (\sigma(j-1), 1)$ for $i < j \leq n + 1$. The chain $\sum_{i=0}^n (-1)^i E_i\sigma$ is called the prism of $\sigma$; we use $p_n$ to denote the map taking an $n$-simplex to its prism, extended by linearity to $\tilde{H}_n(X)$. We claim that $I_1\sigma - I_0\sigma = \tilde{\partial}_n p_n(\sigma) + p_{n-1} \tilde{\partial}_n(\sigma)$, proving the lemma. Let $F_\sigma$ denote $\mu_\sigma$; then $\tilde{\partial}_{n+1} p_n(\sigma) = \sum_{i=0}^n \sum_{j=0}^{n+1} (-1)^{i+j} F_j E_i \sigma$, and $p_{n-1} \tilde{\partial}_n(\sigma) = \sum_{i=0}^n \sum_{j=0}^n (-1)^{i+j} E_j F_i \sigma$. One verifies that $E_i F_\sigma$ equals $F_{j+1} E_i \sigma$ if $i < j$, and $F_j E_{i+1} \sigma$ if $i \geq j$; for $1 \leq i \leq n$ $F_j E_i \sigma$ equals $F_j E_{i-1} \sigma$; $F_0 E_0 \sigma$ equals $I_1\sigma$; and $F_{n+1} E_n \sigma$ equals $I_0\sigma$.

**Lemma 40.** If $f_0, f_1 : X \mapsto Y$ are homotopic then the singular chain maps induced by them are chain homotopic.

**Proof:** This follows by theorem 19.10, and the fact that if $h$ is a homotopy from $f_0$ to $f_1$ then $f_i = h \circ I_i$.

In particular, the induced maps on the homology chains are the same, and spaces of the same homotopy type have the same homology chains. Morphisms $f,g : (X, X') \mapsto (Y, Y')$ of topological pairs are said to be homotopic if there is a morphism $h : (X \times [0, 1],X' \times [0, 1]) \mapsto (Y, Y')$, which is a homotopy from $f$ to $g$.

Preceding fact can be generalized to morphisms of topological pairs; see exercise 16.

A subspace $X' \subseteq X$ is called a strong deformation retraction if there is a continuous function $r : X \times [0, 1] \mapsto X$ (called a strong deformation retraction), such that $r(x, 0) = x$, $r(x, 1) \in X'$, and $r(x, t) = x$ for $x \in X'$ and $0 \leq t \leq 1$.

**Lemma 41.** Suppose $X_1, X_2$ satisfy the hypotheses of lemma 38.b, and $X'_2$ is a strong deformation retraction of $X_2$ with $X_1 \cap X'_2 \neq \emptyset$ and $X_1 \cup X'_2 = X_1 \cup X_2$. Then $\tilde{H}(X_1 \cup X'_2, X'_2)$ is isomorphic to $\tilde{H}(X_1, X_1 \cap X'_2)$.

**Proof:** By lemma 40, $\tilde{H}(X'_2)$ is isomorphic to $\tilde{H}(X_2)$. By the long exact sequence and the five lemma, $\tilde{H}(X_1 \cup X'_2, X'_2)$ is isomorphic to $\tilde{H}(X_1 \cup X'_2, X_2)$. It is immediate that $X_1 \cap X'_2$ is a strong deformation retract of $X_1 \cap X_2$, whence by the same argument $\tilde{H}(X_1, X_1 \cap X'_2)$ is isomorphic to $\tilde{H}(X_1, X_1 \cap X_2)$. The lemma follows.

**Lemma 42.** Suppose $K_1, K_2$ are subcomplexes of $K$. Then $\tilde{H}(K_1 \cup K_2, K_2)$ is isomorphic to $\tilde{H}(K_1, K_1 \cap K_2)$.

**Proof:** Given a pair $(K, L)$ of complexes, say that $L$ is a full subcomplex of $K$ if whenever $S$ is a simplex of $K$, and the vertices of $S$ are all vertices of $L$, then $S$ is a simplex of $L$. We claim that for any pair $(K, L)$, $L_\beta$ is a full subcomplex of $K_\beta$. Indeed, if $v_S$ is a vertex of $L_\beta$ then the vertices of $S$ are in $L$, so the simplexes $T$ of $K_\beta$ for which $S$ is the largest among the $S_i$ of the $v_S$. Of $T$ are in $L_\beta$. Supposing $L$ to be a full subcomplex of $K$, let $N$ be the complex whose vertices are those not in $L$, and whose simplexes are those of $K$ with vertices in this set. For any simplex of $K$, its vertices in $L$ yield a simplex in $L$, and similarly for $N$. We claim that $L$ is a strong deformation retract of $K - N$. Indeed, a point $p$ of $K - N$ may be written as $\sum_{v \in A} e_v + \sum_{v \in B} e_v$, where $A (B)$ are vertices in $L (N)$. It suffices to consider the case where both $A$ and $B$ are nonempty. Let $p_L$ be $\sum_{v \in A} e_v + \sum_{v \in B} e_v$. The strong deformation retraction may be given by $r(p, t) = tp_L + (1 - t)p$ for $p \in K - L - N$. With $K = K_1 \cup K_2$ and $L = K_2$, let $V$ be the subspace of $K_1 \cup K_2$ just constructed. Then $K_1 - K_2 \subseteq K_1^{\text{int}}$, and $K_2 \subseteq V$, so the lemma follows by lemma 41.
**Theorem 43.** On the finite simplicial complexes, \( \theta \) induces a natural equivalence on the homology chains.

**Proof:** It suffices to show that the map \( \tilde{H}_n(K) \rightarrow \tilde{H}_n(K) \) induced by \( \theta \) is injective for all \( n \), for all \( K \); this may be proved by induction on \( |K| \). The case \( |K| = 0 \) is trivial. Suppose \( |K| = 1 \); then \( K \) is a simplex and the claim follows by lemma 34. Otherwise, let \( S \) be a maximal simplex. Let \( K_1 \) be \( K \) with \( S \) removed. Let \( K_2 \) be the subsets of \( X \). \( \theta \) is an isomorphism on \( K_1 \) and \( K_1 \cap K_2 \) by induction, and on \( K_2 \) by lemma 34. Considering the section

\[
\cdots \rightarrow \tilde{H}_{n+1}(K_1 \cap K_2) \rightarrow \tilde{H}_{n+1}(K_1) \rightarrow \tilde{H}_n(K_1, K_1 \cup K_2) \rightarrow \tilde{H}_n(K_1 \cap K_2) \rightarrow \tilde{H}_n(K_1) \cdots
\]

of the long exact sequence, \( H(\theta) \) gives a transformation to the like section for singular homology. All vertical arrows but the middle are isomorphisms, and so the middle one is by lemma 19.6. By lemma 35 \( \tilde{H}(K_1 \cup K_2, K_2) \) is isomorphic to \( \tilde{H}(K_1, K_1 \cap K_2) \), and by lemma 42 the like fact holds for singular homology. Further, as is readily verified (or see [Spanier]), these isomorphisms are induced by inclusions. Thus in the section

\[
\cdots \rightarrow \tilde{H}_{n+1}(K_1 \cup K_2, K_2) \rightarrow \tilde{H}_n(K_2) \rightarrow \tilde{H}_n(K_1 \cup K_2) \rightarrow \tilde{H}_n(K_1 \cap K_2, K_2) \rightarrow \tilde{H}_{n-1}(K_2) \cdots
\]

of the long exact sequence, with transformation \( H(\theta) \) to the like section for singular homology, all vertical arrows but the middle are isomorphisms. The theorem follows by the five lemma.

The theorem is in fact true for all simplicial complexes. The argument involves taking a direct limit of the finite subcomplexes, and may be found in [Spanier].

**Exercises.**

1. Suppose \( T \) is a terminal object in a category with finite products. Let \( \tau : A \rightarrow T \) be the map, and let \( (\iota, \tau) \) be the induced map from \( A \) to \( A \times T \). Show that \( \pi_1 \) is a two-sided inverse. Hint: That \( \pi_1 \) is a left inverse is immediate. Let \( g \) denote \( (\iota, \tau)\pi_1 \); then \( \pi_1 g = \pi_1 \) and \( \pi_2 g = \tau = \pi_2 \), so by uniqueness of maps to the product, \( g = \iota \).

2. Show that if the topology of a nontrivial TVS over \( F \) is Hausdorff the topology of \( F \) must be. Hint: The map \( a \rightarrow ax, x \neq 0 \), is continuous.

3. Show that \( |x|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} \) is a norm on the product of normed linear spaces. Letting \( |x|_1 = \sum |x_i| \), show that \( |x|_2 \leq |x|_1 \leq \sqrt{n}|x|_2 \). Hint: For the triangle inequality proceed as in chapter 10. Use the Cauchy-Schwarz inequality for \( \mathcal{R} \), and the hypotheses \( |x_i + y_i| \leq |x_i| + |y_i| \).

4. a. Show that the Hahn-Banach theorem holds if for any subspace \( Y \subseteq X \), linear functional \( f : Y \rightarrow F \), and \( x \in X - Y \), there is a value \( \xi \in F \) such that \( |\xi - f(y)| \leq |f||x - y| \) for all \( y \in Y \).

b. Show that the condition of part a holds for \( \mathcal{R} \).

c. Show that the Hahn-Banach theorem holds for \( \mathcal{C} \).

d. Show that the condition of part a hold for a field with a non-Archimedean absolute value which is spherically complete, provided the norm on \( X \) is ultrametric.

Hint: For part a, letting \( f(x) = \xi \) and extending by linearity,

\[
f(y) + \alpha \xi = |\alpha| f(-\frac{y}{\alpha}) + \xi \leq |\alpha||f||(x - y)| = |f||y + \xi|.
\]

That is, \( f \) can be extended to \( x \). The linear functionals on subspaces \( Z \subseteq X \) with \( Y \subseteq Z \) extending \( f \) without increasing the norm are inductively ordered by approximation (inclusion); using Zorn’s lemma and the preceding fact \( f \) can be extended to \( X \). For part b, from \( f(y_1 - y_2) \leq |f||y_1 - y_2| \) it follows that \( f(y_1) - |f||x - y_1| \leq f(y_2) + |f||x - y_2| \). Thus there is a \( \xi \) such that for all \( y \in Y \)

\[
f(y) - |f||x - y| \leq \xi.
\]
\[ \xi f(y) + |f||x - y|. \] For part c, let \( g \) be the real part of \( f \), extend \( g \) to \( X \), and let \( F(x) = g(x) - ig(ix) \).

Show that \( F \) preserves addition, multiplication by real scalars, and multiplication by \( i \). Show that \( F \) extends \( f \) (because the real part of \( iz \) is the negative of the imaginary part of \( z \)). Clearly \( |g| \leq |F| \). Let \( \alpha \) be such that \( |\alpha| = 1 \) and \( |F(x)| = \alpha F(x) \); then \( |F(x)| = g(\alpha x) \leq |g|\alpha|x| = |g||x| \).

For part d, \( |f(y_1 - y_2)| \leq |f||y_1 - y_2| \leq |f| \text{max}(|x - y_1|, |x - y_2|) \). This shows that \( B_{\supset}(y_1, f|x-y_1|, f|x-y_2|) \) is nonempty, and the claim follows by results in section 4.

5. Show that if the Hahn-Banach theorem holds in \( F-\text{NLS} \) then the map \( x \mapsto \phi_x \) from \( V \) to \( V^* \) is a homeomorphism; if in addition for all \( x \in X \) there exists \( a \in F \) such that \( |a| = |x| \) the map is an isometry. Hint: Suppose the extra condition holds. Define \( f \) on \( Fx \) by letting \( f(x)^2 = a \) where \( |x| = |a| \), and extend it to all of \( X \), resulting in a functional with \( |f| = 1 \) and \( |f(x)| = |x| \). The claim follows since if \( |\phi_x(f)| \leq r|f| \) then \( |x| \leq r \). In any case there is a \( t \in \mathbb{R} \) with \( 0 < t \leq 1 \) such that for all \( x \) there is an \( a \) with \( t|x| \leq |a| \leq |x| \).

6. Suppose \( X \) and \( Y \) are topological spaces with \( Y \) Hausdorff, and \( f : X \to Y \) is continuous. Show that the graph of \( f \) is a closed subset of \( X \times Y \). Hint: If \( (x, y) \) is a point not in the graph, choose disjoint open sets \( U_1, U_2 \) with \( y \in U_1 \) and \( f(x) \in U_2 \). Then \( f^{-1}(U_1) \times U_1 \) and \( f^{-1}(U_1) \times U_2 \) are disjoint.

7. Suppose \( X \) and \( Y \) are in \( F-\text{CLNS} \), and \( f \) is a linear transformation from \( X \) to \( Y \) whose graph is a closed subset of \( X \times Y \). Then \( f : X \to Y \) is continuous. Hint: By hypothesis the graph of \( f \) is a closed subset of \( X \times Y \) equipped with the norm \( |x| + |y| \). It is therefore complete. The projection onto \( X \) is bijective and linear from the graph to \( X \). By the open mapping theorem there is some \( b \) such that \( |x| + |f(x)| \leq b|x| \), whence \( f \) is continuous.

8. Suppose \( X \) is a normed linear space over \( \mathbb{R} \), and the parallelogram law holds. Let \( x * y = (1/4)(|x + y|^2 - |x - y|^2) \); show that this defines an inner product. Show that its norm is the original norm. Hint: Clearly \( * \) is symmetric. Adding and subtracting \( |x - y + z|^2 \) on the left and using the polarization law, \( |x + y + z|^2 - |x + y - z|^2 = 2(|x + z|^2 + |y|^2) - 2(|x|^2 + |y - z|^2) \). Exchange \( x \) and \( y \) and add the two equations. Conclude that \( (x + y) * z = x * z + y * z \). Use induction to conclude that \( nx * y = n(x + y) \) for \( n \) and integer.

Conclude that \( qx * y = q(x + y) \) for \( q \) a rational. Use continuity to conclude that \( rx * y = r(x * y) \) for \( r \) a real. The last claim is immediate.

9. Prove the complex version of exercise 8, where \( x * y = (1/4)(|x + y|^2 - |x - y|^2 + i(|x + iy|^2 - |x - iy|^2)) \).

Hint: Verify that \( y * x = (x * y)^* \). Show that \( (x + y) * z = x * z + y * z \) by splitting it into real and imaginary parts and proceeding as in exercise 8. Conclude as before that \( rx * y = r(x + y) \) for \( r \) a real. Verify that \( ix * y = i(x * y) \).

10. Show that \( \ell_2 \) is a Hilbert space, in either the real or complex case (indeed, more generally if in each component \( x_i \) is from some Hilbert space \( X_i \)). Hint: The sequences form a vector space with the componentwise operations; let \( X \) be those where \( \sum_i x_i \) converges. Given sequences \( x_i \) and \( y_i \) in \( X \), by the parallelogram law \( |x_i + y_i|^2 \leq 2(|x_i|^2 + |y_i|^2) \), so the componentwise sum is in \( X \). Clearly \( X \) is closed under scalar multiplication. By the Cauchy-Schwarz inequality, \( |x_i * y_i| \leq |x_i||y_i| \leq (x_i^2 + y_i^2)/2 \); it follows that \( \sum_i x_i * y_i \) exists. Defining this to be the inner product of the sequences, the inner product is linear in the first argument since it is in each position, and absolutely convergent series are rearrangeable; similarly it is symmetric (or Hermitian). It is positive definite because \( \sum_i x_i * x_i \) is 0 only if each term is 0. Given a Cauchy sequence of sequences, it is a Cauchy sequence in each component, so converges to a limit in each component. The sequence of limits is in \( X \), and the original Cauchy sequence converges to it in \( X \).

11. Show that an orthonormal set \( S \) in a Hilbert space \( X \) is maximal iff its span is dense. Hint: If \( S \) is maximal every element in \( X \) is the limit of the sequence of partial sums of an infinite series. If \( S \) is not maximal there is a vector orthogonal to its span, so at positive minimum distance.

12. Show that a Hilbert space \( X \) is separable iff it has a countable (or finite) maximal orthonormal set. Hint: Let \( Q \) be the rationals, or if \( F = \mathbb{C} \) the complex numbers \( q_1 + q_2i \) where \( q_j \) is rational. If \( S \) is a
countable maximal orthonormal set let $T$ be the finite linear combinations of elements of $S$ with coefficients from $Q$. Then $T$ is a countable dense subset of $X$. If $T$ is a countable dense subset, enumerate it as $t_i$.

Construct a set $S = \{s_n\}$ inductively, by letting at stage $n$ $i$ be least such that $t_i$ is not in the span of $\{s_j : j < n\}$; obtain $s_n$ by orthonormalizing as in section 10.8. $S$ is orthonormal, and its span is dense in $X$.

13. Prove the claims about $\text{Inv}$ made at the end of section 10. Hint: Consider the functors $F_i \in C^{J_i}$ where $J_i = \{V : f[U_i] \subseteq V\}$, for $i = 1, 2, 3$, where $U_1 \supseteq U_2 \supseteq U_3$. The map induced by the inclusion $J_1 \subseteq J_3$ is the composition of those induced by $J_1 \subseteq J_2$ and $J_2 \subseteq J_3$. This shows that $\text{Inv}(Q)$ is a presheaf. Given components $\alpha_U : Q_1(U) \rightarrow Q_2(U)$ of a morphism of presheaves on $X$, the arrows $\alpha_V$ for $\{V : f[U] \subseteq V\}$ determine an arrow $\beta_U : \text{Inv}(Q_1)(U) \rightarrow \text{Inv}(Q_2)(U)$. Further, these maps are the components of a morphism from $\text{Inv}(Q_1)$ to $\text{Inv}(Q_2)$. This shows that $\text{Inv}$ is a functor. Suppose that $Q$ is a presheaf on $Y$. Then $\text{Dir}(\text{Inv}(Q))(U)$ is the direct limit of $\{Q(V) : V \supseteq f[f^{-1}[U]]\}$. This gives an arrow $\mu_U : Q(U) \rightarrow \text{Dir}(\text{Inv}(Q))(U)$. These arrows are the components of a morphism $\mu$ (which will be seen to be the unit). This gives a map $\alpha \mapsto \text{Dir}(\alpha)\mu$ from $\text{Hom}(\text{Inv}(Q), P)$ to $\text{Hom}(Q, \text{Dir}(P))$. Given $\beta \in \text{Hom}(Q, \text{Dir}(P))$, $U \in \Omega(X)$, and $V \in \Omega(Y)$ with $f[U] \subseteq V$, $\beta_V : Q(V) \rightarrow P(f^{-1}[V])$, and the is an arrow of the presheaf $P$ from $P(f^{-1}[V])$ to $P(U)$. The composition of these two maps is a map from $\text{Hom}(Q, \text{Dir}(P))$ to $\text{Hom}(\text{Inv}(Q), P)$, which is inverse to the above map. Further the systems of maps are natural. This shows that $\text{Inv}$ is the left adjoint. Let $Q$ denote $\text{Dir}(P)$. Suppose $U = \cup\{U_i\}$ in $Y$, and let $u, s, t$ be in the definition of a sheaf in an arbitrary category. It follows from $f^{-1}[U] = \cup\{f^{-1}[U_i]\}$, and the fact that $P$ is a sheaf, that $u$ is the equalizer of $s$ and $t$. This shows that $Q$ is a sheaf. The third fact follows by composition of natural transformations.

14. Prove the claims made about partial derivatives made in section 12. Hint: For any $u \in \mathbb{R}^n$ Let $h : \mathbb{R} \rightarrow \mathbb{R}^m$ be defined by $h(t) = x + tu$. By the chain rule, $(f \circ h)'(0) = f'(x)u$. This vector in $\mathbb{R}^m$ is called the directional derivative of $f$ at $x$, in the direction $u$. It equals $\lim_{t \rightarrow 0}(f(x + tu) - f(x))/t$. Facts 1 and 2 follow. For fact 3, let $\delta_i$ be the vector which is $(\Delta x)_j$ for $j \leq i$, and 0 for $j > i$. It suffices to show that

$$\sum_i(f(x + \delta_i) - f(x + \delta_{i-1}) - \frac{\partial f}{\partial x_i}\Delta x_i)$$

approaches 0 faster than $\Delta x$. This follows by continuity and lemma 18. Fact 4 follows by the chain rule and the fact that the matrix of the derivative of the derivative of the identity function is the identity matrix.

15. In the notation of lemma 32, given a sequence $\nu = \nu_0 \ldots \nu_n$ of vertices let $s_0 \ldots s_n$ be the sequence, reordered to agree with a chosen well-order on the vertices. Let $h_n(\nu) = \sum_{i=0}^n(-1)^{t_i}s_0 \ldots s_i\nu_i$, where $t_i = i + \sum_{j<i}|\{k : j < v_k > s_j\}|$, and $v_i$ is $\nu$ with the elements $s_0 \ldots s_{i-1}$ removed (where multiple occurrences of an element are indexed from left to right). Show that $h$ is a chain homotopy from $\eta\zeta$ to $\iota$.

Hint: The term of $h_{n-1}\tilde{\partial}_n(\nu)$ obtained by deleting $v_i$ and adding $s'_0, \ldots s'_j$ in front (where the sequence $s'_j$ is the sequence $s_j$, with $v_i$ deleted) equals up to sign the term of $\tilde{\partial}_{n+1}h_n(\nu)$ obtained by adding $s_0 \ldots s_j$ in front and deleting $v_i$. The remaining terms cancel in pairs, except for $\nu + \eta(\zeta)\nu$.

16. Show that if $f, g : \langle X, X' \rangle \rightarrow \langle Y, Y' \rangle$ are homotopic morphisms of topological pairs then the induced maps on the relative singular homology chains $\tilde{H}(X, X')$, $\tilde{H}(Y, Y')$ are the same. Hint: Suppose $c$ is a relative cycle in $C_n(X)$ (i.e., its boundary is in $C_{n-1}(X')$). Then $\tilde{\partial}_n(c) \in C_n(X')$, so $p_{n-1}\tilde{\partial}_n(c) \in C_n(X' \times I)$, where $p_n$ is as in the proof of lemma 42. So by the proof of lemma 42, $I_1c - I_0c$ is a relative boundary.
25. Algebraic geometry.

1. Introduction. An algebraic curve in the real plane $\mathbb{R}^2$ is defined to be any set $\{ (x, y) : p(x, y) = 0 \}$ where $p(x, y)$ is a polynomial in two variables with real coefficients. Algebraic curves have fascinated mathematicians since the time of the Greeks.

More generally if $F$ is a field an algebraic set in $F^n$ is any set of the form $V(P)$ where $P$ is a set of polynomials in $n$ variables with coefficients in $F$, and $x \in V(P)$ iff $p(x) = 0$ for all $p \in P$. Although some facts may be given for arbitrary $F$, the theory when $F$ is algebraically closed is more straightforward. For this reason, if $\bar{F}$ is the algebraic closure of $F$ the algebraic set defined in $\bar{F}^n$ is considered to give the “geometry” of the set, and the points in $F^n$ to be of special interest, even for polynomials with coefficients in (“defined over”) $F$.

To obtain a “complete picture” of it, an algebraic set may be enlarged in another way, by considering projective $n$-space (to be defined below) rather than affine $n$-space $F^n$.

Algebraic geometry proceeds by developing algebraic tools to study the properties of algebraic sets. Many results can be proved by brute force computations with equations, but introducing abstract tools can make proofs more tractable and transparent, and as the subject progresses they become increasingly more necessary. In the 1960’s the subject underwent a revolution with the introduction of scheme theory. The 1970’s saw another revolution, with the development of a computational aspect of the subject and a return to the use of concrete methods.

For the original purposes of the subject the field $F$ is the complex numbers. Much goes through unchanged if $F$ is allowed to be any algebraically closed field of characteristic 0. It has become important to consider arbitrary characteristic; in some cases complications arise, and characteristic 0 is assumed to avoid considering them. Unless explicitly specified characteristic 0 will not be assumed. Recall that $F$ is infinite (Theorem 7.5), and that as a consequence the only polynomial in $F[x_1, \ldots, x_n]$ which is 0 everywhere is the 0 polynomial.

Some topics in this chapter belong more properly to commutative algebra. However their use in algebraic geometry is a fundamental one; further they are sensibly covered here in the organization of the text. Theorems from commutative algebra may not require algebraic closure, and this will be stated if so.

2. Noetherian topological spaces. The algebraic sets comprise the closed sets of a topological space, of a particular type; basic properties of such spaces can be given axiomatically. Recall that a topological space (or subspace) is irreducible if it is nonempty and is not the union of two proper closed subsets.

Lemma 1. For a nonempty topological space $X$, the following are equivalent.

a. $X$ is irreducible.

b. Any two nonempty open sets in $X$ intersect.

c. If $\{K_i\}$ is a finite closed cover of $X$ then $X = K_i$ for some $i$.

d. If $U \subseteq X$ is nonempty and open then $U^\text{cl} = X$.

e. If $U \subseteq X$ is open then $U$ is connected.

Proof: That a$\Rightarrow$b follows because $X = K_1 \cup K_2$ iff $\emptyset = U_1 \cap U_2$, where $U_i = K_i^c$. That c$\Rightarrow$a is trivial; the other direction is proved by induction on the number $n$ of sets in the cover. If $n = 1$ then c is trivial; otherwise $X = K_1$ or $X = K_2 \cup \cdots \cup K_n$ and it follows by induction. To see that b$\Rightarrow$d, observe that there is a nonempty open subset disjoint from $U$ iff $U^\text{cl} \neq X$. Finally, if $U_1, U_2$ are disjoint nonempty open subsets of $U$ then b is violated, so b$\Rightarrow$e. Conversely if $U_1, U_2$ are disjoint nonempty open sets then $U_1 \cup U_2$ is not connected.

Lemma 2. Let $X$ be a nonempty topological space and $S$ a nonempty subset.

a. If $X$ is irreducible and $S$ is open then $S$ is irreducible.
b. $S$ is irreducible iff $S^\dagger$ is.

c. The irreducible subsets of $X$ are inductively ordered, so there are maximal ones. These are closed and cover $X$.

d. If $S$ is irreducible and $\{K_i\}$ is a finite closed cover of $S$ then $S \subseteq K_i$ for some $i$.

e. If $\{U_i\}$ is a finite open cover of $X$, with each $U_i$ nonempty, then $X$ is irreducible iff each $U_i$ is.

f. If $T \subseteq S$ is nonempty then if $T$ is irreducible in $S$ then $T$ is irreducible; if $S$ is closed the converse holds.

g. If $f : X \rightarrow Y$ is continuous and $S$ is irreducible then $f[S]$ is irreducible.

**Proof:** For part a, if $U_1, U_2$ are disjoint open subsets of $S$ they are disjoint open subsets of $X$. For part b, if $U_1, U_2$ are open, the statements $U_1 \cap S \not= \emptyset$ and $U_1 \cap U_2 \cap S \not= \emptyset$ are true iff they are true with $S$ replaced by $S^\dagger$. For part c, suppose for a contradiction that $C$ is a chain of irreducible subsets, $U_1, U_2$ are open, $U_i \cap (\cup C) \not= \emptyset$, and $U_1 \cap U_2 \cap (\cup C) = \emptyset$. Then for $i = 1, 2$ $U_i \cap S \not= \emptyset$ for some $S \subseteq C$. Let $S$ be the larger of $S_1, S_2$; then $U_i \cap S = \emptyset$, and also $U_1 \cap U_2 = \emptyset$, which is impossible because $S$ is irreducible. That a maximal irreducible subset is closed follows by part b. Each point is irreducible, so is contained in a maximal irreducible subset. For part d, by lemma 1.c $S = S \cap K_i$ for some $i$. For part e, if $X$ is irreducible then each $U_i$ is by part a. Conversely, $\cap_i U_i$ is a nonempty open subset of $U_i$ so is irreducible. Any nonempty open subset $V$ intersects some $U_i$, so intersects $\cap_i U_i$; any two such thus intersect. For part f, let $T_1, T_2$ denote closed sets; if $T = T_1 \cup T_2$ then $T = (T_1 \cap S) \cup (T_2 \cap S)$, so either $T = T_1 \cap S$ or $T = T_2 \cap S$, so either $T = T_1$ or $T = T_2$. If $S$ is closed and $T = (T_1 \cap S) \cup (T_2 \cap S)$ then either $T = T_1 \cap S$ or $T = T_2 \cap S$ because $T_1 \cap S$ is closed. For part g, if $U_1$ and $U_2$ are disjoint nonempty open subsets of $F[S]$ then $S \cap f^{-1}[U_1]$ and $S \cap f^{-1}[U_2]$ are disjoint nonempty open subsets of $S$.

A topological space $X$ is called Noetherian if the closed subsets satisfy the descending chain condition. A collection of sets $\{Y_i\}$ is called irredundant (or subset-free) if for no $i, j$ is $Y_i \subseteq Y_j$. For any finite collection $\{Y_i\}$ of sets there is an irredundant subcollection $\{Z_i\}$ with $\cup \{Y_i\} = \cup \{Z_i\}$. Indeed, if the collection is redundant simply repeat the step of choosing $Y_i \subseteq Y_j$ and discarding $Y_i$.

**Theorem 3.** Suppose $X$ is a Noetherian topological space.

a. Every subspace $Y \subseteq X$ is Noetherian.

b. $X$ is compact.

c. Every nonempty closed subset $Y \subseteq X$ is a union of finitely many irreducible closed subsets $\{Y_i\}$. There is a unique irredundant such collection, namely the maximal irreducible subsets of $Y$.

d. If $\{Y_i\}$ is a finite cover of a topological space $Y$ and each $Y_i$ is Noetherian then $Y$ is Noetherian.

**Proof:** For part a, if $K_i \cap Y$ ($K_i$ closed) is a descending chain in $Y$ then $K_i$ is a descending chain in $X$. For part b, if $C$ is an open cover inductively choose $U_i \in C$ which is not a subset of $\cup_{j<i} U_j$; eventually the process must terminate. For part c, suppose $Y$ were a minimal closed subset which is not a union of finitely many closed irreducible subsets; $Y$ can then be irreducible nor a union of proper closed subsets. If $\{Y_i\}$ and $\{Z_i\}$ are two irredundant collections, then $Y_1$ is a subset of some $Z_i$ and $Z_i$ is a subset of some $Y_i$; it follows that $Z_i = Y_1$, and renumbering we may assume $Z_1 = Y_1$. Replace $Y$ by $(Y - Y_1)^{\dagger}$ and continue inductively, noting that $(Y - Y_1)^{\dagger} = \bigcup_{i>1}(Y_i - Y_i)^{\dagger} = \bigcup_{i>1} Y_i$. Each $Y_i$ must be maximal, and each maximal subset of $Y$ must appear. For part d, if $K_j$ ($K_j$ closed) is a descending chain in $Y$, for each $i$ $K_j \cap Y_i$ is eventually constant; and so $K_j$ is eventually constant.

This section concludes with some further definitions from topology which are occasionally useful in algebraic geometry, and other subjects. Suppose $Y$ is a subspace of a topological space $X$. The map $y \mapsto y^{\dagger}$ maps subsets of $Y$ to closed subsets of $X$, and $x \mapsto x \cap Y$ maps closed subsets of $X$ to subsets of $Y$. Clearly both maps are order preserving, and for $y \subseteq Y$ and closed $x \subseteq X$, $y \subseteq x \cap Y$ iff $y^{\dagger} \subseteq x$. Thus, these maps comprise a Galois adjunction between the subsets of $Y$ and the closed subsets of $X$, ordered by inclusion.
The constructible subsets of a topological space are defined to be the ring of sets generated by the open sets. A subset of a topological space is called locally closed if it is the intersection of a closed set and an open set. The intersection of two locally closed sets is locally closed, and it readily follows that a set is constructible iff it is a finite union of locally closed sets.

Clearly a subset is locally closed iff it is open in a closed set, or closed in an open set. Also, it follows by the above Galois adjunction that $Z$ is locally closed iff $Z$ is open in $Z^{\text{cl}}$ ($Z = K \cap U$ iff $Z = Z^{\text{cl}} \cap U$, $K$ closed).

If $Z \subseteq Y \subseteq X$, and if $Z$ is locally closed (equaling $K \cap U$) then $Z$ is locally closed in $Y$ (equaling $(K \cap Y) \cap (U \cap Y)$). If $Z$ is locally closed in $Y$ (equaling $(K_1 \cap Y) \cap (U_1 \cap Y)$) and $Y$ is locally closed (equaling $K_2 \cap U_2$) then $Z$ is locally closed (equaling $(K_1 \cap K_2) \cap (U_1 \cap U_2)$).

3. Affine algebraic sets. Let $F$ be an algebraically closed field, and $n$ and integer; $F^n$ is also known as $n$-dimensional affine space. As on previous occasions, we let $x$ denote an $n$-tuple $x_1, \ldots, x_n$. Let $V$ be the map from subsets of $F[x]$ to subsets of $F^n$, where for $P \subseteq F[x]$, $x \in V(P)$ iff $p(x) = 0$ for all $p \in P$. A set of the form $V(P)$ will be called an algebraic set in $F^n$. An affine algebraic set is an algebraic set in $F^n$ for some $n$, although other terminology is used.

Let $I$ be the map from subsets of $F^n$ to subsets of $F[x]$, where for $p \in F[x]$ and $X \subseteq F^n$, $p \in I(X)$ iff $p(x) = 0$ for all $x \in X$. Observe that $I(X)$, the set of all polynomials which vanish everywhere on $X$, is an ideal.

The map $V$ is order preserving from the subsets of $F[x]$ ordered by reverse inclusion to the subsets of $F^n$ ordered by inclusion (decreasing the polynomial set increases the zero set). The map $I$ is order preserving in the opposite direction (increasing the zero set decreases the polynomial set). Given $P \subseteq F[x]$ and $X \subseteq F^n$, $X \subseteq V(P)$ iff $p(x) = 0$ for $p \in P$ and $x \in X$, iff $I(X) \supseteq P$.

The preceding paragraph shows that $V$ and $I$ form a Galois adjunction. The following facts are readily verified, either from the adjunction or directly.

- $X \subseteq V(I(X))$, $I(V(P)) \supseteq P$, $V(I(V(P))) = V(P)$, $I(V(I(X))) = I(X)$, $V(\cup_i P_i) = \cap_i V(P_i)$, and $I(\cup_i X_i) = \cap_i I(X_i)$.
- A subset $X$ of $F^n$ is algebraic iff $X = V(I(X))$.
- If $I$ is the ideal generated by $P$ then $V(I) = V(P)$.
- $I(\emptyset) = F[x]$, $I(F^n) = \{0\}$, and $I(X)$ is a proper ideal if $X \neq \emptyset$.
- $V(\emptyset) = F^n$, and $V(F[x]) = \emptyset$.

Finally, by Corollary 20.36, $V(I)$ is nonempty if $I$ is a proper ideal, and $V(P) = \emptyset$ iff $P$ contains a nonzero constant polynomial.

**Theorem 4.** $V(P_1) \cup V(P_2) = V(P_1P_2)$.

**Proof:** If $x \in V(P_1)$ then $p(x) = 0$ for all $p \in P_1$, so $p(x) = 0$ for any multiple of any $p \in P_1$, and a fortiori $V(P_1) \subseteq V(P_1P_2)$; similarly $V(P_2) \subseteq V(P_1P_2)$. Suppose $p_1p_2(x) = 0$ for all $p_1 \in P_1$ and $p_2 \in P_2$, and $p_1(x) \neq 0$ for some $p_1 \in P_1$; then $p_2(x) = 0$ for all $p_2 \in P_2$.

If $P_1$ and $P_2$ are ideals $P_1P_2$ may be replaced by their product. As a corollary, any finite union of algebraic sets in $F^n$ is algebraic. By this and preceding facts the algebraic sets are the closed sets of a topology on $F^n$, called the Zariski topology. Since the closed sets of the Zariski topology are the range of $V$, it follows by properties of Galois adjunctions that $X^{\text{cl}} = V(I(X))$. Also by properties of Galois adjunctions, $I$ is strictly monotone on the algebraic sets. Since there is no ascending chain of ideals in $F[x]$ there is no descending chain of algebraic sets in $F^n$; that is, the Zariski topology is Noetherian.

An algebraic set in $F^n$ is said to be irreducible if it is an irreducible closed set in the Zariski topology, that is, cannot be written as a union of two proper algebraic subsets. Since the Zariski topology is Noetherian,
every algebraic set can be written uniquely as the union of an irredundant collection of irreducible algebraic subsets. The zero set of \( \{xz,yz\} \) in \( \mathbb{R}^3 \) is a good example; it is the union of the \( xy \) plane and the \( z \) axis. Irreducible algebraic sets have properties which often make them more suitable than general ones.

The Zariski topology on an affine algebraic set \( V \) is defined to be the subspace topology. On topological grounds the closed sets are the affine algebraic sets \( W \) which are subsets of \( V \). The following further facts are readily verified, where \( V \) denotes an affine algebraic set.

- The sets \( V(p) \) comprise a base for the closed sets of the Zariski topology on \( F^n \).
- A nonempty open subspace \( U \) of \( V \) is dense in \( V \) (lemma 1.d).
- If two continuous functions \( f_1, f_2 : V \to F^n \) agree on a nonempty open subspace \( U \subseteq V \), then they agree on \( V \); this follows since their difference \( f_1 - f_2 \) is 0 on a closed subset of \( V \) containing \( U \).

An open subset of an affine algebraic set is called a quasi-affine algebraic set. Note that these are exactly the locally closed subsets of \( F^n \).

**Theorem 5.** \( \text{Rad}(I(X)) = I(X) \), and \( I(V(I)) = \text{Rad}(I) \).

**Proof:** If \( p \in I(X) \) and \( q^n = p \) then clearly \( p \in I(X) \); this shows that \( \text{Rad}(I(X)) \subseteq I(X) \), and since \( I \subseteq \text{Rad}(I) \), \( \text{Rad}(I(X)) = I(X) \). From this and \( I \subseteq I(V(I)) \), \( \text{Rad}(I) \subseteq \text{Rad}(I(V(I))) = I(V(I)) \). The inclusion \( \text{Rad}(I) \subseteq I(V(I)) \) is just theorem 20.37.

Recall from section 20.8 that an ideal \( I \) is radical if \( \text{Rad}(I) = I \). Thus, the Galois adjunction yields a bijection between the algebraic subsets of \( F^n \), and the radical ideals in \( F[x] \).

**Theorem 6.** The Galois adjunction yields a bijection between the irreducible algebraic subsets of \( F^n \), and the prime ideals in \( F[x] \).

**Proof:** If \( p_1, p_2 \in I(V) \) then \( V \subseteq V(p_1, p_2) \), so \( V = V \cap V(p_1, p_2) = V \cap (V(p_1) \cup V(p_2)) \). If \( p_i \not\in V \) for \( i = 1, 2 \) then \( V = V \cap V(p_i) \neq V \). This shows that if \( V \) is irreducible then \( I(V) \) is prime. If \( V \) is reducible, say \( V = V_1 \cup V_2 \) where \( V_i \not\subseteq V \), then \( I(V_i) \not\supseteq I(V) \), so for \( i = 1, 2 \) there is a polynomial \( p_i \in I(V_i) \setminus I(V) \). Since \( I(V) = I(V_1) \cap I(V_2) \), \( p_1, p_2 \in I(V) \), and \( I(V) \) is not prime.

By theorem 20.32, theorem 20.26, and lemma 20.28, in any Noetherian ring a proper radical ideal \( I \) is the intersection of finitely many prime ideals. In the case of \( F[x] \), the prime ideals may be taken as those corresponding to the subsets of \( V(I) \) in its irredundant cover by irreducible algebraic subsets.

The Galois adjunction clearly yields a bijection between the points (minimal irreducible closed subsets) of \( F^n \), and the maximal ideals in \( F[x] \). By lemma 20.33, the latter are those generated by \( x_1 - a_1, \ldots, x_n - a_n \), for some point \( a \).

If \( V \) is an affine algebraic set it is natural to ask what properties the quotient ring \( F[x]/I(V) \) of \( F[x] \) by the ideal \( I(V) \) has. We use \( F[V] \) to denote this ring; the following are readily verified.

- Under the canonical epimorphism \( \phi : F[x] \to F[V] \), ideals of \( F[V] \) are in bijective correspondence with ideals of \( F[x] \) which contain \( I(V) \). Radical (prime, maximal) ideals of \( F[V] \) correspond to like ideals.

The radical ideals correspond to those arising from the algebraic sets which are contained in \( V \).

- \( F[V] \) is a Noetherian ring.
- \( F[V] \) is finitely generated as an \( F \)-algebra, by \( x_1 + I(V), \ldots, x_n + I(V) \).
- \( F[V] \) is reduced.
- \( F[V] \) is an integral domain iff \( V \) is irreducible.
- The canonical map \( F \to F[x] \to F[V] \) is injective, provided \( V \) is nonempty.
- \( F[V] \) is a field iff \( I(V) \) is a point \( \{a\} \). In this case, the kernel of the evaluation map \( p \to p(a) \) is \( I(V) \), and the canonical isomorphism \( F[x]/I(V) \to F \) is the inverse to the canonical map given above.

A polynomial \( p \in F[x] \) determines a function from \( F^n \) to \( F \); such a function is called a polynomial function. For an affine algebraic set \( V \), a function \( f : V \to F \) is called a polynomial function if it is the
restriction of a polynomial function on \( F^n \). The polynomial functions on \( V \) may be identified with \( F[V] \), since two polynomials determine the same function on \( V \) iff their difference is in \( I(V) \). The functions \( x_i \) are called the coordinate functions, and \( F[V] \) is called the coordinate ring.

A function \( f : F^n \to F^m \) is called a polynomial map if it is a polynomial function in each coordinate. For affine algebraic sets \( V \subseteq F^n \) and \( W \subseteq F^m \), a function \( f : V \to W \) is called a polynomial map if it is the restriction of a polynomial map from \( F^n \) to \( F^m \). It is readily verified that the composition of polynomial maps is a polynomial map. Thus, with these as morphisms the affine algebraic sets over \( F \), that is, those with \( V \subseteq F^n \) for some \( n \), form a category; we may assume that \( n \) can be determined from \( V \).

The map \( V \to F[V] \) is the object map of a functor from the category of affine algebraic sets over \( F \), to the category of \( F \)-algebras. The image of the polynomial map \( f : V \to W \) maps \( F[W] \) (considered as polynomial functions) to \( F[V] \) by composition on the right with \( f \); this yields a functor by a well-known argument.

It is readily verified that a polynomial map \( f : V \to W \) is continuous when \( V \) and \( W \) are equipped with their Zariski topologies. The converse does not hold. The simplest counterexample is \( F \); the algebraic subsets of \( F \) are \( F \) itself, and finite subsets. Any bijection of \( F \) is thus continuous in the Zariski topology.

**Theorem 7.** The functor \( V \to F[V] \) is an equivalence of categories, to the full subcategory of the commutative \( F \)-algebras, of reduced and finitely generated algebras.

**Proof:** For each reduced finitely generated \( F \)-algebra \( A \), choose a set of generators; this specifies an isomorphism from \( A \) to \( F[x]/I \) for an ideal \( I \). Since \( A \) is reduced \( I \) is radical, whence \( I = I(V(I)) \), when \( F[x]/I = F[x]/I(V(I)) = F[V(I)] \). Writing \( V(A) \) for \( V(I) \), the map \( A \to V(A) \) is the object map of a functor. First, given \( \phi : F[W] \to F[V] \) where \( W \subseteq F^m \), for \( 1 \leq i \leq m \) let \( p_i \in F[x_1, \ldots, x_m] \) be a polynomial such that \( \phi(x_i + I(W)) = p_i + I(V) \). Let the image of \( \phi \) be the polynomial map from \( V \) to \( W \) with the \( p_i \) as components. The image of a morphism \( \phi : A \to B \) is the image of the corresponding map from \( F[V(A)] \) to \( F[V(B)] \). Suppose \( \phi : F[U] \to F[V] \psi : F[V] \to F[W] \). Under the image of \( \phi \) \( x_1 \) maps to \( p_i \). Under the image of \( \psi p_i \) maps to \( p_i(q_1, \ldots, q_l) \) where \( W \subseteq F^l \). This is exactly the composition of the multivariate image maps \( u \to V \to U \). Thus, \( A \to V(A) \) respects composition for maps of this type; it readily follows that \( A \to V(A) \) respects composition in general. We may assume that \( V(F[V]) = V \). By definition there are isomorphisms \( A \to F[V(A)] \). Also, \( f : F[x]/I \to F[x]/J \) maps to itself under the composed functor \( F[V(A)] \), since the latter map \( x \to I + J \) maps to the composition of \( x \) with the polynomial map \( V(J) \to V(I) \), that is, to \( p_i + J \). It follows that the system of isomorphisms \( A \to F[V(A)] \) is natural. The theorem follows by lemma 21.11.

This theorem shows, among other things, that the polynomial maps are a class of interest. Also, two affine algebraic sets are equivalent by polynomial maps iff their coordinate rings are isomorphic. The category of affine algebraic sets possesses a product; the corresponding coordinate ring is the tensor product of the coordinate rings of the factors (exercise 1).

An affine subspace (see section 22.1) of \( F^n \) is the set of solutions to a system of linear equations. As such, it is an affine algebraic set. We define an affine linear map from \( F^n \) to \( F^m \) to be one of the form \( x \mapsto Mx + t \), where \( M \) is an \( m \times n \) matrix and \( t \in F^m \) (recall from section 23.7 that these include the isometries in the case of Euclidean space).

The affine subspaces, with the restrictions of the affine linear maps, form a subcategory of the affine algebraic sets. From linear algebra one easily sees that an affine subspace of \( F^n \) is equivalent by an affine linear map to \( F^k \) for some \( k \leq n \). Since \( I(F^k) = \{0\} \), and \( \{0\} \) is a prime ideal in the integral domain \( F[x] \), \( F^n \) is irreducible. It follows that \( F^k \subseteq F^n \) is irreducible, and hence that any affine subspace is irreducible.

For a polynomial \( p \), let \( D_p = \{ x : p(x) \neq 0 \} \). Then \( D_p \) is the complement of \( V(\{p\}) \), and the sets \( D_p \) comprise a base for the open sets.
If $V$ is irreducible then $F[V]$ is an integral domain. We use $F(V)$ to denote the field of fractions, which is called the field of rational functions of $V$. A quotient $p/q$ of polynomials defines a function on the basic open $D_q$ of $F^n$. This function is continuous on its domain (since $\{ x : (p/q)(x) = c \}$ equals $\{ x : (p-cq)(x) = 0 \}$). The use of these functions involves various complications. Further discussion is given in section 6; here we note that $p_1/q_1$ and $p_2/q_2$ represent the same element of $F(V)$ iff $p_1q_2 - p_2q_1 \in I(V)$.

Another example of relations between an affine algebraic set and its coordinate ring is the following. The argument shows that for any ideal $I$, if $F[X]/I$ is a finite dimensional vector space over $F$ then $V(I)$ is finite. The converse is also true; see [Fulton].

**Theorem 8.** An affine algebraic set $V$ is finite iff $F[V]$ is a finite dimensional vector space over $F$, and in this case the size equals the dimension.

**Proof:** Suppose $a_j$ for $1 \leq j \leq t$ are distinct points in $F^n$. Let $i(j,k)$ be a coordinate index where $a_{ji} \neq a_{kj}$. Let $p_j = \prod_{k \neq j} \frac{x_{i(j,k)} - a_{k,i(j,k)}}{a_{j,i(j,k)} - a_{k,i(j,k)}}$. Then $p_j(a_k) = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta function. If $I$ is an ideal in $F[x]$ and the points $a_j$ are in $V(I)$ then the elements $p_j + I$ of $F[x]/I$ are linearly independent, since if $\sum_j c_j p_j + I = 0$ then $\sum_j c_j p_j \in I$, so $c_k = \sum_j c_j p_j(a_k) = 0$. Suppose in addition that $I$ is radical, $\overline{V(I)} = \{ a_1, \ldots, a_t \}$, and $q \in F[x]$; let $r = q - \sum_j q(a_j)p_j$. Then $r(a_k) = 0$ for all $k$, so $r \in I(\{ a_1, \ldots, a_t \}) = I(V(I)) = I$, so $q + I$ is in the span of the $p_j + I$. We have shown that if $V$ is finite then the dimension of $F[V]$ equals $|V|$, and if $V$ is infinite then $F[V]$ is infinite dimensional, and the theorem follows.

4. **Noether normalization.** The Noether normalization theorem is a ring-theoretic version of the fact that an extension field is algebraic over a transcendence base. It holds for any field, but versions may be given with restrictions on the field and a strengthened conclusion. We give the version for any field.

**Lemma 9.** Suppose $F$ is a field, $p \in F[x]$, and $e$ is greater than the degree of any variable in $p$. For $1 \leq i < n$ let $y_i = x_i - x_n^e$. Then $p = q(y_1, \ldots, y_{n-1}, x_n)$ where $q$ is monic as a polynomial in $x_n$.

**Proof:** If $x_1^{d_1} \cdots x_n^{d_n}$ is a monomial of $p$, when written in terms of $y_1, \ldots, y_{n-1}, x_n$ its highest degree term is $x_n^d$ where $d = d_1e + \cdots + d_{n-1}e^{n-1}$. The $d_i$ are just the digits in the $e$-adic notation for $f$. Thus, the highest degree terms from the monomials are powers of $x_n$, of distinct degree for distinct monomials, and the lemma follows.

**Theorem 10.** Suppose $A$ is a finitely generated $F$-algebra over a field $F$, and an integral domain. Then there is a transcendence base $\{ a_1, \ldots, a_n \} \subseteq A$ for $A_{\mathbb{A}^n}$ over $F$, such that $A$ is integral over $F[a_1, \ldots, a_n]$.

**Proof:** Let $\{ a_1, \ldots, a_n \}$ be a subset of $A$ with least $n$, such that $A$ is integral over $F[a_1, \ldots, a_n]$. Suppose $p(a_1, \ldots, a_n) = 0$ where $p$ is a nonzero polynomial with coefficients in $F$. Then $q(b_1, \ldots, b_{n-1}, a_n) = 0$, where $b_i = a_i - a_i^e$, and $e$ and $q$ are as in lemma 9. This shows that $A$ is integral over $F[b_1, \ldots, b_{n-1}]$, contradicting the existence of $p$.

As a corollary, the transcendence degree of $A_{\mathbb{A}^n}$ equals the maximum size of an algebraically independent subset of $A$. This may be called the transcendence degree of $A$.

5. **Dimension.** Being a geometric object, one suspects that an algebraic set has a dimension. There are several equivalent definitions of the dimension of an affine algebraic set, each with its own advantages. The Krull dimension, or simply dimension, of a commutative ring can be defined, and justified by some axioms (see [Eisenbud]), to be the largest $n$ such that there is a chain $P_0 \subset \cdots \subset P_n$ of prime ideals; or $\infty$ if no such $n$ exists. In section 20.9 we have already defined dimension 0 (every prime ideal is maximal) and
dimension 1 for an integral domain (every nonzero prime ideal is maximal). The dimension of the trivial
ring must be defined by convention; one such is $-\infty$.

The dimension of a topological space $V$ can be defined to be the the largest $n$ such that there is a chain
$V_0 \supset \cdots \supset V_n$ of irreducible closed subsets of $V$. For an affine algebraic set $V$, this is readily seen to equal
the Krull dimension of $F[V]$, using the correspondence of prime ideals of $F[V]$ with those of $F[x]$ containing
$I(V)$. Other types of dimension are defined for a topological space, and this one is sometimes called the
combinatorial dimension. Again, by convention the dimension of the empty set is defined to be $-\infty$.

The Krull dimension has additional properties for Noetherian rings, although it is not necessarily finite.
A counterexample due to Nagata may be found in [AtiMac]. It will follow from results of this section that
the dimension of a finitely generated $F$-algebra is at most $n$ where $n$ is the number of generators. As a
consequence, the dimension of an affine algebraic set in $F^n$ is at most $n$.

We use $\text{Kdim}(R)$ to denote the Krull dimension of a commutative ring $R$. More standard notation in
algebraic geometry is to use dim for the Krull dimension, and $\dim_F$ For the dimension of a vector space over $F$. We also use $\text{Kdim}(X)$ for the combinatorial dimension of a topological space $X$; as already noted for an
affine algebraic set $V$ $\text{Kdim}(V) = \text{Kdim}(F[V])$. Finally, we use $\text{Trdeg}$ to denote the transcendence degree.

**Lemma 11.** Suppose $A$ is a finitely generated $F$-algebra over a field $F$, and an integral domain. If $P$ is a
nontrivial prime ideal of $A$ then $\text{Trdeg}(A/P) < \text{Trdeg}(A)$.

**Proof:** Suppose $a_1 + P, \ldots, a_n + P$ are algebraically independent, and $b \in P$ is nonzero. Suppose
$p(a_1, \ldots, a_n, b) = 0$; then $p(a_1, +P, \ldots, a_n + P, 0) = 0$ in $A/P$, so $p(x_1, \ldots, x_n, 0)$ must be the zero poly-
nomial, so $p$ is a polynomial in $x_{n+1}$ with 0 constant term. Since $A$ is an integral domain $p$ may be supposed
irreducible, so $p$ equals $x_{n+1}$. This contradicts the assumption $b \neq 0$, showing that $p$ does not exist.

**Lemma 12.** Suppose $F$ is a field; then $\text{Kdim}(F[x]) = n$.

**Proof:** For $0 \leq i \leq n$ let $P_i$ be the ideal generated by $x_1, \ldots, x_i$. $P_i$ is a prime ideal, because $p \in P_i$ iff, as a
polynomial with coefficients in $F[x_{i+1}, \ldots, x_n]$, the constant term is 0. Further these ideals form a chain, with
the inclusions being proper. Thus, $\text{Kdim}(F[x]) \geq n$. $\text{Trdeg}(F[x])$ equals $n$, because $F[x]$ is the free algebra.
If $P_0 \subset \cdots \subset P_m$ is any chain of prime ideals in $F[x]$ then $n > \text{Trdeg}(F[x]/P_0) > \cdots > \text{Trdeg}(F[x]/P_m)$,
whence $m \leq n$ by lemma 11 and the correspondence of prime ideals under canonical epimorphism.

In particular, the Krull dimension of an affine algebraic set equals its dimension as a linear space. In
the remainder of the section, we will show among other things that for an irreducible affine algebraic set $V$,
$\text{Kdim}(V)$ equals the transcendence degree of $F(V)$ over $F$, a fact known as the dimension theorem.

In a commutative ring, the height of a prime ideal $P$ is defined to be the largest $n$ such that there is a
chain $P_0 \subset \cdots \subset P_n = P$ of prime ideals. Recall from section 20.7 that for a multiplicative subset $S \subseteq R$
there is a bijective correspondence between the prime ideals of $RS$, and those of $R$ disjoint from $S$. It follows
that the height of $P$ equals the dimension of $RP$.

Krull’s principal ideal theorem states the following. Suppose $R$ is a Noetherian ring. Suppose $r \in R$ is
neither a zero divisor nor a unit. Let $P$ be a minimal prime ideal containing $r$. Then $P$ has height 1. This
theorem is a centerpiece of commutative algebra. Proofs may be found in [AtiMac], [Eisenbud], [Jacobson],
or [Matsumura]. Unfortunately, the proof is long and involves introducing many additional definitions.

Fortunately there is a specialized version which suffices for some purposes in algebraic geometry, and
has a shorter proof; we give this, following [Nielsen].

Suppose $R$ is an integral domain, $S \subseteq R$ is a multiplicative subset, and $R^* (R^*_S)$ is the field of fractions
of $R$ ($R_S$). As observed in section 6.4 there is a canonical embedding of $R_S$ in $R^*$, and by lemma 6.7.c this
induces a homomorphism from $R^*_S$ to $R^*$; indeed, it maps $(r_1/s_1)/(r_2/s_2)$ to $(r_1s_2)/(r_2s_1)$. This map is
readily verified to be an isomorphism.

361
Recall from section 6.5 that if $R$ is a commutative ring, and $s \in R$ is not nilpotent, $R_s$ denotes the ring of fractions where the denominator is a power of $s$. If $R$ is an integral domain then the restriction is met for any nonzero $s$, and $R_s$ is an integral domain contained in the field of fractions. As noted in section 20.7, there is a bijective correspondence between the prime ideals of $R_s$, and those of $R$ not containing $s$. For another fact of interest, $R_s$ is isomorphic to $R[x]/[sx - 1]$, where as in chapter 6 $[S]$ denotes the ideal generated by $S$ (exercise 7).

Recall from section 15.2 the definition of the norm $N_{E:F}$ of a finite extensions of fields.

**Lemma 13.** Suppose $A$ is an integral domain, $F$ is the field of fractions, and $E \supseteq F$ is a finite extension. If $e \in E$ is integral over $A$ then so is $N_{E:F}(e)$.

**Proof:** As noted in section 20.6, the irreducible polynomial $p$ for $e$ over $F$ may be assumed to be monic. As noted in section 15.2, $N_{E:F}(e)$ is the product of powers of the roots of $p$. The lemma follows by corollary 20.18.

**Lemma 14.** Suppose $R$ is a Noetherian ring. Then there are only finitely many minimal prime ideals in $R$.

**Proof:** Let $S$ be the set of ideals $I$ such that $R/I$ has infinitely many minimal prime ideals. Since $R$ is Noetherian, if $S$ is nonempty then it contains a maximal element $I$. We may thus assume that $R$ is such that for any ideal $I$ in $R$, there are only finitely many prime ideals containing $I$. $R$ is not an integral domain, because the only minimal prime ideal in an integral domain is $\{0\}$. Choose nonzero $a, b$ with $ab = 0$. Any minimal prime must contain either $a$ or $b$, since it contains 0. There are only finitely many minimal prime ideals containing $a$, and only finitely many containing $b$. This is a contradiction.

**Theorem 15.** Suppose $A$ is a finitely generated $F$-algebra over a field $F$, and an integral domain. Suppose $a \in A$ is neither zero nor a unit. Let $P$ be a minimal prime ideal containing $a$. Then $\text{Trdeg}(A/P) = \text{Trdeg}(A) - 1$.

**Proof:** By lemma 13 applied to the quotient, there are only finitely many minimal prime ideals containing $a$, say $P_1, \ldots, P_t$ with $P = P_1$. Choose $s_i \in P_i - P$ for $i \geq 2$, and let $s = s_2 \cdots s_n$; then $s \in P_2 \cdots P_n - P$. Since the field of fractions of $A_s$ is isomorphic to that of $A$, $\text{Trdeg}(A_s) = \text{Trdeg}(A)$. Similarly, $\text{Trdeg}(A_s/P_s) = \text{Trdeg}(A/P)$. In $A_s$, there is only one minimal prime ideal containing $a$, because any minimal prime ideal in $A_s$ containing $a$ does not contain $s$, so arises from a prime ideal in $A$ containing $a$ but not $s$, which cannot be a minimal prime ideal containing $a$, so the ideal in $A_s$ cannot be minimal. Thus, we may assume that there is only one prime ideal containing $a$. By theorem 10 there are algebraically independent elements $t_1, \ldots, t_n$ such that $A$ is integral over $B = F[t_1, \ldots, t_n]$. As noted in section 20.7, $B/(B \cap P)$ is a subring of $A/P$, so it suffices to show that $\text{Trdeg}(B/(B \cap P)) \geq n - 1$, for then $\text{Trdeg}(A/P) \geq n - 1$, whence by lemma 11 $\text{Trdeg}(A/P) = n - 1$. Let $A^* (B^*)$ be the field of fractions of $A (B)$. As observed in section 20.6, $A^*$ is an algebraic extension of $B^*$. Since it is finitely generated, it follows using corollary 20.17 that it is a finite extension. Let $e$ denote the dimension, and let $N$ denote $N_{B^*:A^*}$. Suppose $c \in P \cap B$. By theorem 20.25, $c^e = aa$ for some $a \in A$ and $r \in N$. Also, $N(c) = c^e$; so $N(c^r) = c^e r = N(a) N(a)$. Since $N(a) \in B^*$ by lemma 13, and $c^e r \in B$, $N(a) \in B^*$. Also by lemma 13, $N(a) \in A$. As observed in section 7.3, $B$ is a factorial domain. By theorem 20.19, $B$ is integrally closed. Thus, $N(a) \in B$. We have shown that $P \cap B \subseteq \text{Rad}(bB)$ where $b = N(a)$; by theorem 20.25 equality holds, and $B \cap P$ is the unique prime ideal if $B$ containing $b$. Considering the unique factorization of $b$ in the factorial domain $B$, the minimal prime ideals containing $b$ are principal ideals, generated by prime elements of $B$, namely the prime factors of $b$. In particular $P \cap B$ equals $dB$ for some such $d$, which is an irreducible polynomial in the $t_i$. Further, this polynomial cannot be a constant, else $d \in F$, so $P \cap B \subseteq F$, so $P \cap B \subseteq \{0\}$, so $d = 0$, so $a = 0$, contradicting the hypotheses. Suppose w.l.g. that $d$ has nonzero degree in $t_n$. Then no polynomial in $t_1, \ldots, t_{n-1}$ can be a multiple of $d$, which shows that $B/dB$ contains $n - 1$ algebraically independent elements.
COROLLARY 16. Suppose $A$ is a finitely generated $F$-algebra over a field $F$, and an integral domain. Then all maximal chains of prime ideals have length $\text{Trdeg}(A)$.

PROOF: The proof is by induction on $d = \text{Trdeg}(A)$. If $d = 0$, then $A$ is an integral extension of $F$, so by theorem 20.21.e $A$ is a field, and the only prime ideal is $\{0\}$. Suppose $P_0 \subset \cdots \subset P_e$ is a chain of prime ideals. It follows by lemma 11 that $e \leq d$. If the chain is maximal then $P_e$ is a minimal prime ideal. By the correspondence of prime ideals under canonical epimorphism, theorem 14, and the induction hypothesis applied to $A/P_0$, $e = d$.

So far it has not been necessary to assume that $F$ is algebraically closed. If $F$ is algebraically closed, it follows that the length of any maximal chain of irreducible closed subsets of an irreducible affine algebraic set $V \subseteq F^n$ has length $\text{Kdim}(V)$.

Many authors define the dimension as the maximum of the dimensions of the irreducible components, and the dimension of an irreducible affine algebraic set as the transcendence degree. An affine algebraic set is said to be of pure dimension $d$ if all its irreducible components have dimension $d$.

In the case $F = \mathbb{C}$, the dimension of an affine algebraic set is that of the complex algebraic set over $\mathbb{C}$. In simple cases (for example affine spaces) this equals what one would like for the dimension of the real part over $\mathbb{R}$.

6. Projective algebraic sets. Let $P^n$ denote the quotient of $F^{n+1} - \{0\}$ by the equivalence relation, $x \equiv y$ iff $y = rx$ for some $r \in F^\neq$. $P^n$ ($n$-dimensional projective space) is the lines through the origin in $F^{n+1}$, although as in $F^n$ its elements are called points. A vector $\langle x_1, \ldots, x_{n+1} \rangle$ in the class of a point is said to constitute homogeneous coordinates for the point. We let $x$ denote a vector in $F^{n+1}$, and by abuse of notation consider such an element of $P^n$.

If $x \in P^n$ and $p \in F[x]$, $x$ is said to be a zero of $p$ if $p(rx) = 0$ for all $r \in F^\neq$. We introduce an $N$-grading of $F[x]$, by letting the homogeneous subspace of degree $i$ be those polynomials where each monomial is of total degree $i$. Let $p_{(i)}$ denote the monomials of degree $i$ in $p$. If $x$ is a 0 of $p$ then $q(r) = \sum_i r^i p_{(i)}(r)$ is identically 0, so each $p_{(i)}(x)$ is 0; in particular the constant term $p_{(0)}$ equals 0. Also, if $p$ is homogeneous then $x$ is a zero of $p$ iff $p(x) = 0$.

As in the affine case, for $Q \subseteq F[x]$ let $V(Q)$ be the set of $x \in P^n$ such that $x$ is a 0 of $p$ for all $p \in Q$. A set of the form $V(Q)$ will be called a projective set (in $P^n$). For $X \subseteq P^n$ let $I(X)$ be the set of $p \in F[x]$ such that $x$ is a zero of $p$ for all $x \in X$.

The following are readily verified.
- $I(X)$ is a homogeneous ideal in $F[x]$, by theorem 18.6 and the above noted fact.
- Given $P \subseteq F[x]$ and $X \subseteq F^n$, $X \subseteq V(P)$ iff $I(X) \supseteq P$, whence $V$ and $I$ form a Galois adjunction.
- $V(Q_1) \cup V(Q_2) = V(Q_1Q_2)$; the modifications to the proof of theorem 4 are straightforward.
- The algebraic subsets of $P^n$ comprise the closed sets of a topology on $P^n$, again called the Zariski topology; it is Noetherian.
- A homogeneous ideal is proper iff none of its members has a nonzero constant term.
- The ideal $I$ generated by the polynomials $x_1, \ldots, x_{n+1}$ is called the irrelevant ideal; $V(I) = \emptyset$ for this ideal, since the only affine zero is 0.
- If $I$ is homogeneous then $\text{Rad}(I)$ is homogeneous (exercise 5).

If $X$ is an set in $P_n$ let $\text{Acon}(X)$ be the set in $F^{n+1}$ of all the points belonging to the points of $X$, together with the 0 vector. When necessary $I_a$ and $V_a$, and $I_p$ and $V_p$, are used to distinguish between the affine and projective operations. It is readily verified that if $X \subseteq P^n$ is nonempty then $I_a(\text{Acon}(X)) = I_p(X)$; and if $I$ is a homogeneous ideal such that $V_p(I) \neq \emptyset$ then $\text{Acon}(V_p(I)) = V_a(I)$. Also, $V$ is algebraic iff $\text{Acon}(V)$ is algebraic.
THEOREM 17. Suppose $I$ is a homogeneous ideal in $F[x]$.
(a) $V(I) = \emptyset$ iff for some integer $N$, $x_1^N \in I$ for $1 \leq i \leq n+1$.
(b) If $V(I) \neq \emptyset$ then $I(V(I)) = \text{Rad}(I)$.

PROOF: Let $I_+$ denote the irrelevant ideal. Then $V_p(I) = \emptyset$ iff $V_o(I) \subseteq \{0\}$ iff $I_+ \subseteq I_o(V_o(I))$ iff $I_+ \subseteq \text{Rad}(I)$, and part a follows. For part b, $I_p(V_p(I)) = I_o(A\text{con}(V_p(I))) = I_o(V_o(I)) = \text{Rad}(I)$.

It follows that $I$ and $V$ establish a bijection between algebraic subsets of $P^n$, and homogeneous radical ideals other than the irrelevant ideal in $F[x]$.

Let $\alpha_i$ be the map from $F^n$ to $F^{n+1}$ where $\alpha_i((x_1, \ldots, x_n)) = (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$ (with obvious conventions when $i$ is 1 or $n$). Let $\alpha_i$ be the composition with the canonical epimorphism from $F^{n+1}$ to $P^n$. Via any $\alpha_i$, affine space $F^n$ may be viewed as a subspace of projective space $P^n$. Letting $A_i$ denote the image under $\alpha_i$, clearly $A_i$ is an open subspace of $P^n$. Further, the $A_i$ cover $P^n$.

Choosing $i = n + 1$, let $H_\infty$ be the points $\langle x, \ldots, x, 0 \rangle$ of $P^n$, so that $P^n$ is the disjoint union of $A_{n+1}$ and $H_\infty$. This fact illustrates the reason for considering projective space. $H_\infty$, the “points at infinity”, are points added to affine space, one for each direction in affine space, and themselves form a copy of $P^{n-1}$. Adding these points facilitates the statement of facts about algebraic sets.

Any subset $X \subseteq P^n$ has a closure, namely $V(I(X))$. If $V$ is an algebraic set in $F^n$, let $V^\dagger$ denote the closure of $\alpha_{n+1}[V]$. On the other hand, given $V \subseteq P^n$, let $V^\uparrow$ denote the affine algebraic set $\alpha_{n+1}^{-1}(V)$ (equivalently $\alpha_{n+1}^{-1}(A\text{con}(V))$).

If $p \in F[x_1, \ldots, x_n]$ is a polynomial of total degree $d$ let $p^\dagger \in F[x_1, \ldots, x_{n+1}]$ be the degree $d$ homogeneous polynomial obtained by multiplying each degree $d$ term of $p$ by $x^{d-e}$. Given $I \subseteq F[x_1, \ldots, x_n]$ let $I^\dagger$ be the ideal generated by $\{p^\dagger : p \in I\}$; clearly $I^\dagger$ is homogeneous. On the other hand given a homogeneous polynomial $p \in F[x_1, \ldots, x_{n+1}]$ let $p^\dagger \in F[x_1, \ldots, x_n]$ be $p(x_1, \ldots, x_n, 1)$; and for a homogeneous ideal $I \subseteq F[x_1, \ldots, x_{n+1}]$ let $I^\dagger$ be the ideal generated by $\{p^\dagger : p \in I\}$.

LEMMA 18.
(a) For $V \subseteq F^n$, $V^\dagger = V_p(I_p(V)^\dagger)$.
(b) For $V \subseteq P^n$, $V^\dagger = V_o(I_o(V)^\dagger)$.
(c) For $V \subseteq F^n$, $V^\dagger = V$.
(d) For $V \subseteq P^n$, $V^\dagger \subseteq V$, with equality if $V \subseteq A_{n+1}$.
(e) For $V \subseteq F^n$, if $V$ is irreducible then $V^\dagger$ is irreducible.
(f) For $V \subseteq P^n$, if $V$ is irreducible and $V \cap A_{n+1} \neq \emptyset$ then $V^\dagger \subseteq V$; further $V^\dagger$ is irreducible.

PROOF: For part a, it suffices to show that if $\alpha_{n+1}[V] \subseteq W$ then $V_p(I_p(V)^\dagger) \subseteq W$. So suppose the former, and $q \in I(W)$; then $\alpha_{n+1}(x) \in W$, so $\langle x, 1 \rangle$ is a zero of $q$, so $q^\dagger(x) = 0$. This shows that $q^\dagger \in I(V)$. But $q = x_{n+1}^e q^\dagger$ where $e$ is the largest power of $x_{n+1}$ dividing $q$. Thus, $q \in I(V)^\dagger$. We have now shown that $I_p(V) \subseteq I_p(V)^\dagger$, whence $W \supseteq V_p(I_p(V)^\dagger)$. For part b, if $\langle x, 1 \rangle \in A\text{con}(V)$ and $q \in I_p(V)^\dagger$ then $q = q^\dagger(x, 1)$ where $q^\dagger \in I_p(V)$. Also, $\langle x, 1 \rangle$ is a zero of $q^\dagger$, whence $p(x) = 0$. This shows that $V^\dagger \subseteq V_o(I_o(V)^\dagger)$. Suppose $q(x) = 0$ for $q \in I_p(V)^\dagger$. Then $\langle x, 1 \rangle$ is a zero of $q^\dagger$ for $q^\dagger \in I_p(V)$, that is, $\langle x, 1 \rangle \in V_p(I_p(V))^\dagger$, so $\langle x, 1 \rangle \in A\text{con}(V)$. This shows that $V_o(I_p(V)^\dagger) \subseteq V^\dagger$. For part c, clearly if $x \in V$ then $\langle x, 1 \rangle \in V^\dagger$. Suppose $\langle x, 1 \rangle \in V^\dagger$. Suppose $q \in I_o(V)$; then $q^\dagger \in I_o(V)^\dagger$, whence by part a $q^\dagger(x, 1) = 0$, whence $q(x) = 0$. Thus, $x \in V_o(I_o(V)^\dagger) = V$. We have shown that $x \in V$ iff $\langle x, 1 \rangle \in V^\dagger$; and clearly $x \in V^\dagger$ iff $\langle x, 1 \rangle \in V^\dagger$, proving part c. For part d, $\alpha_{n+1}[\alpha_{n+1}^{-1}(V)] \subseteq V$, with equality if $V \subseteq A_{n+1}$; the claim follows by applying $V_p \circ I_p$. For part e, if $V$ is irreducible then $I_p(V)$ is prime; it suffices to show that $I_p(V)^\dagger$ is prime, since then $V^\dagger = V_p(I_p(V)^\dagger)$ is irreducible. Suppose $I$ is prime; using $q = q^\dagger$ and $(q_1 q_2)^\dagger = q_1^\dagger q_2^\dagger$, if $q_1^\dagger q_2^\dagger \in I^\dagger$ then $q_1 q_2 \in I$, so $q_1 \in I$ or $q_2 \in I$, so $q_1^\dagger \in I^\dagger$ or $q_2^\dagger \in I^\dagger$. For part f, if $q \in I_p(V)^\dagger$ and $\langle x, 1 \rangle \in A\text{con}(V)$ then $x \in V^\dagger$, so $q(x) = 0$, so $\langle x, 1 \rangle$ is a zero of $q^\dagger$. This shows that $x$ is a zero of $q^\dagger$ whenever $x \in V = V_p(I_p(V))$. 364
By theorem 17, $q^m \in I_p(V)$ for some integer $m$, and $q \in I_p(V)^1$. It follows that for some $e$, $x_{n+1}^{e+1} q^{m+1} \in I_p(V)$. But $x_{n+1} \notin I_p(V)$, by the hypothesis that $V \cap A_{n+1} \neq \emptyset$, and $I_p(V)$ is prime; thus, $q^1 \in I_p(V)$. We have shown that $I_p(V(q_p(I_0(V)^1))) = I_p(V)^1 \subseteq I_p(V)$, whence $V \subseteq V_p(I_0(V)^1) = V^1$. By part $d$, $V = V^1$. Now suppose $q_1 q_2 \in I_a(V^1)$; by an argument just given, for some $e$, $x_{n+1}^e (q_1 q_2)^m \in I_p(V)$, whence $q_1^1 \in I_p(V)$ or $q_2^1 \in I_p(V)$. But if $q^1 \in I_p(V)$ and $x \in V^1$ then $(x, 1) \in V^1 \subseteq V$. Thus, $(x, 1)$ is a zero of $q$, whence $q(x) = 0$. That is, if $q^1 \in I_p(V)$ then $q \in I_a(V^1)$. Since $q_1^1 \in I_p(V)$ or $q_2^1 \in I_p(V)$, $q_1 \in I_a(V^1)$ or $q_2 \in I_a(V^1)$. This shows that $I_a(V^1)$ is prime, whence $V^1$ is irreducible.

It readily follows that the map $\bar{a}_i$ is a homeomorphic embedding of $F^n$ in $P^n$, with image $A_i$. For further facts concerning these operations, see [Fulton].

As in the affine case, the Zariski topology on a projective algebraic set $V$ is that inherited from $P^n$. The closed sets are the projective algebraic sets which are subsets. A nonempty open subset is dense. Two continuous functions to $F$ or further remarks on the algebraic and may be seen to reduce to what would be desired in specific cases. It is readily verified that with these functions determined by the elements of $\Gamma \subseteq \mathbb{A}$. The quasi-projective algebraic sets comprise a category. For further remarks on the algebraic objects associated with a projective or quasi-projective algebraic set, see [Harts] for example.
Because the relationship of an algebraic set to algebraic objects associated with it is more complicated than in the affine case, it is simplest to define the dimension of a projective algebraic set as the as the length of the longest chain of irreducible closed subsets in the Zariski topology. As in the affine case, it has alternative characterizations.

**Lemma 19.** If $V$ is an irreducible projective algebraic set, and $V^\perp \neq \emptyset$, then $F(V)$ is isomorphic to $F(V^\perp)$.

**Proof:** Suppose $p/q$ is an element of $F(V^\perp)$, where $\deg(p) = s$ and $\deg(q) = t$. Writing $x$ for $x_{n+1}$, let $p' = x^iq^1$ and $q' = x^tp^1$. We claim that $p/q \mapsto p'/q'$ is an isomorphism from $F(V^\perp)$ to $F(V)$, proving the theorem. Given $p_1/q_1$ and $p_2/q_2$, $p'_1/q'_1 \equiv p'_2/q'_2$ iff $p'_1q'_2 - p'_2q'_1 \in I(V)$ iff $x^{(p_1q_2 - p_2q_1)^1} \in I(V)$, where $e = \max(s_1 + t_2, s_2 + t_1) - \min(s_1 + t_2, s_2 + t_1)$. Since $I(V)$ is prime and $x \notin I(V)$, this is so iff $(p_1q_2 - p_2q_1)^1 \in I(V)$. By the proof of lemma 18, this is so iff $p_1q_2 - p_2q_1 \in I(V^\perp)$, proving that the map is well-defined and injective. Given $p/q$ in $F(V)$, $p'/q'$ maps to it, proving the map is surjective.

Recall the Galois adjunction of section 2 between the closed subsets of a topological space $X$ and the subsets of a subspace $Y$. Suppose $Y$ is open, $x \subseteq X$ is closed and irreducible, and $x \cap Y \neq \emptyset$. By lemma 1.d in $x$, if $x' \subseteq X$ is closed, $x' \subseteq x$, and $x' \cap Y = x \cap Y$ then $x' = x$. It follows that $x \cap Y$ is irreducible in $Y$ (if $x \cap Y = (x_1 \cap Y) \cup (x_2 \cap Y)$ then $(x_1 \cup x_2) \cap x = x$). It also follows that $(x \cap Y)^{cl} = x$ (because if $x'$ is the closure then $x' \cap x = x$; and $x$ is closed).

**Lemma 20.**

a. If $Y$ is a subspace of a topological space $X$ then $\text{Kdim}(Y) \leq \text{Kdim}(X)$.

b. If $\{U_i\}$ is an open cover of a topological space $X$ then $\text{Kdim}(X) = \max_i \text{Kdim}(U_i)$.

c. If $X$ is an irreducible topological space, $\text{Kdim}(X) = d$, $Y$ is a closed subspace, and $\text{Kdim}(Y) = d$, then $Y = X$.

**Proof:** By the Galois adjunction, the subsets of $Y$ closed in $Y$ are in bijective correspondence with the closures of the subsets of $Y$. Thus, a chain of closed subsets of $Y$ yields a chain of closed subsets of $X$, by taking the closures in $X$. Further, by lemma 2 if the subsets are irreducible in $Y$ then their closures are irreducible in $X$. This proves part a. Suppose $K_0 \subset \cdots \subset K_n$ is a chain of irreducible closed subsets of $X$. Then $K_0 \cap U_i \neq \emptyset$ for some $i$. By the remarks above $K_0 \cap U_i \subset \cdots \subset K_n \cap U_i$ is a chain of closed irreducible subsets of $U_i$. Part b follows. For part c let $K_0 \subset \cdots \subset K_d \subset Y$ be a chain of irreducible closed subsets of $Y$. $K_i^{cl} = K_i$, and by lemma 2. $K_i$ is irreducible. Thus, the $K_i$ are a chain in $X$, whence $K_d = Y = X$.

**Theorem 21.** If $V$ is an irreducible projective algebraic set then the length of any maximal chain of irreducible closed subsets equals $\text{Trdeg}(F(V))$.

**Proof:** Choose any maximal chain, of length $d$, say; by lemma 20, and the fact that the $A_i$ are an open cover of $P_n$, we may assume the chain restricts to a maximal chain of length $d$ in $V^\perp$. In particular $d = \text{Trdeg}(F(V^\perp))$. The theorem follows by lemma 19.

It may also be shown that $\text{Kdim}(V) = \text{Kdim}(F[V]) - 1$; see [Harts]. Note that $\text{Kdim}(P^n) = n$.

**7. Hilbert functions.** Hilbert functions are defined for graded modules over a graded ring $R$. Many authors consider the ring to be $\mathcal{N}$-graded, but applications in combinatorics (see for example [Stanley]) have led to considering more generally $\mathcal{N}$-graded rings for some $n$, where $\mathcal{N}$ is taken as a commutative monoid with componentwise addition.

For a graded ring $R$, $R$ is Noetherian iff $R_0$ is Noetherian and $R$ is finitely generated as an $R_0$-algebra (exercise 2). If $R_0$ is Artinian then any finitely generated $R_0$-module has finite length (exercise 3). In most applications $R_0$ is a field, whence finitely generated $R_0$-modules are finite dimensional vector spaces. For the rest of the section, unless otherwise stated $R$ is a finitely generated $\mathcal{N}$-graded algebra over a field $F$. 

366
If $M$ is a $\mathbb{Z}^n$-graded module finitely generated over $R$, then for only finitely many $v \leq 0$ is $M_v$ nonzero. Thus these modules are only slightly more general than $\mathcal{N}$-graded modules; and they are of interest. Each homogeneous submodule $M_v$ is finitely generated (exercise 4), so is a finite dimensional vector space.

Under these circumstances, there is a function $H_M : \mathbb{Z}^n \rightarrow \mathcal{N}$, where $H_M(v) = \dim(M_v)$; this function is called the Hilbert function of $M$.

The function giving the dimension of a finite dimensional vector spaces over $F$ has the following “additivity” property. Suppose $N \subseteq M$ is a subspace; then $\dim(M) = \dim(N) + \dim(M/N)$. In fact, the length function has this additivity property on finite length modules over a commutative ring. This follows by lemma 16.12 and the behavior of the inverse of the canonical epimorphism on submodules.

If $0 \rightarrow M_1 \rightarrow \cdots \rightarrow M_t \rightarrow 0$ is an exact sequence of finite dimensional vector spaces, with $f_i : M_i \rightarrow M_{i+1}$ for $1 \leq i < t$, let $N_i = \text{Im}(f_i) = \text{Ker}(f_{i+1})$ (with $N_0 = N_{n+1} = 0$). Then $\dim(M_i) = \dim(N_i) + \dim(N_{i+1})$, from which $\sum_{i=0}^{n} (-1)^i \dim(M_i) = 0$.

For $v \in \mathcal{N}^n$ let $x^v$ denote the monomial $x_1^{v_1} \cdots x_n^{v_n}$. The following theorem is known as the Hilbert-Serre theorem in the case $n = 1$; the generalization to $n > 1$ may be found in [Stanley].

**Theorem 22.** Suppose $R$ is generated as a $\mathcal{N}^n$-graded commutative $F$-algebra by homogeneous generators $g_1, \ldots, g_t$, where $g_i \in R_{v_i}$, and $M$ is a finitely generated $\mathbb{Z}^n$-graded $R$-module. Then $\sum_{v} H_M(v)x^v = x^{-u}p(x)/\prod_{i=1}^{t} (1 - x^{v_i})$, where $u \in \mathcal{N}^n$ and $p \in \mathbb{Z}[x]$.

**Proof:** The proof is by induction on $t$. If $t = 0$ then $M$ is a finite dimensional vector space, so $M_v=0$ except for finitely many $v$. If $t > 0$, multiplication by $g_t$ yields an exact sequence

$$0 \rightarrow K_v \rightarrow M_v \rightarrow M_{v+v_t} \rightarrow L_{v+v_t} \rightarrow 0$$

for each $M_v$. Let $K = \oplus_v K_v$ and $L = \oplus_v L_v$; these are both finitely generated $\mathbb{Z}^n$-graded $R$-modules. Further, for $m \in K_v$, $g_t g_v m = g_t g_v m = 0$, whence $K$ is in fact a module over $F[g_1, \ldots, g_{t-1}]$. Similarly $L$ is a module over $F[g_1, \ldots, g_{t-1}]$. By the exact sequence, $H_K(v) = H_M(v) + H_M(v + v_t) - H_L(v + v_t) = 0$.

Let $\Sigma_M$ abbreviate $\sum_{v} H_M(v)x^v$; multiplying by $x^{v+v_t}$ and adding yields $(1 - x^v)\Sigma_M = \Sigma_L - x^v\Sigma_K$. The theorem follows by induction.

Specializing to the case $n = 1$, $R$ is $\mathcal{N}$-graded and $M$ is $\mathbb{Z}$-graded; also suppose $v_1 = \cdots = v_t = 1$.

Writing $n$ now for the grade, $H_M(n)$ is the coefficient of $x^n$ in $x^{-u}p(x)/(1 - x)^t$ (if $M$ is $\mathcal{N}$-graded $u$ may be taken as $0$, as may be verified from the proof of the theorem). Reducing to lowest terms, $H_M(n)$ is the coefficient of $x^n$ in $x^{-u}q(x)/(1 - x)^t$, where $s \leq t$ and $q = \sum_{i=0}^{r} q_i x^i$, say. Now,

$$\frac{1}{(1-x)^t} = \sum_{k \geq 0} \binom{k + s - 1}{s - 1} x^k$$

(this is an infinite case of the binomial theorem, or may be obtained by successive differentiation of $1/(1-x) = \sum_{k} x^k$). It follows that for $n \geq r - u$, $H_M(n) = H_p_M(n)$ where

$$H_p_M(n) = \sum_{i=0}^{r} q_i \binom{n + u - i + s - 1}{s - 1}.$$
Proof: Suppose $I$ is such an ideal, with $I = \text{Ann}(x)$. Suppose $rs \in I$ and $s \notin I$. Then $rsx = 0$ and $sx \neq 0$, whence $r \in \text{Ann}(sx)$. Since $I \subseteq \text{Ann}(sx)$, by the maximality of $I$, $r \in I$.

A prime ideal which is the annihilator of some (necessarily nonzero) $x \in M$ is called an associated prime ideal. Various additional facts about associated primes may be found in standard commutative algebra references.

Lemma 24. Suppose $R$ is a $\mathcal{N}$-graded commutative ring and $M$ is a $\mathcal{Z}$-graded $R$-module. Suppose $\text{Ann}(x)$ is a prime ideal $P$. Then $P$ is homogeneous.

Proof: Let $x = x_1 + \cdots + x_s$ where the $x_i$ are homogeneous and $x_1$ is of least degree. We prove the lemma by induction on $s$. If $r \in P$, write $r = r_1 + \cdots + r_t$ where the $r_i$ are homogeneous and of increasing degree. Then $r_1 x_1 = 0$. When $s = 1$, $r_1 \in P$ follows, and also $(r_2 + \cdots + r_t)x_1 = 0$; and inductively $r_i \in P$ follows for each $i$. If $s > 1$ let $y = r_1(x_2 + \cdots + x_n)$, and let $I = \text{Ann}(y)$. Then $P \subseteq I$; if $P = I$ then since $y$ is a sum of $s - 1$ homogeneous elements, $P$ is homogeneous by induction. Otherwise choose $q \in I - P$. Then $q y = 0$, so $q r_1 \in \text{Ann}(x) = P$, so $r_1 \in P$. Again we may continue inductively.

Lemma 25. Suppose $R$ is a Noetherian commutative ring and $M$ is a finitely generated $R$-module. Then there is a chain $0 = M_0 \subseteq \cdots \subseteq M_m = M$ of submodules, such that for $i > 0$ $M_i/M_{i-1}$ is isomorphic to $R/P_i$ for some prime ideal $P_i$. If $R$ is $\mathcal{N}$-graded and $M$ is $\mathcal{Z}$-graded then $P_i$ may be taken as homogeneous.

Proof: As noted in section 8.5, $M$ is Noetherian. By lemma 23 there is an $x \in M$ such that $\text{Ann}(x)$ is a prime ideal $P$. As noted in section 8.1, the submodule $M_1 = Rx$ is isomorphic to $R/P_i$ for some prime ideal $P_i$. If $R$ is $\mathcal{N}$-graded and $M$ is $\mathcal{Z}$-graded then $P_i$ may be taken as homogeneous.

Lemma 26. Suppose $V$ is an irreducible affine algebraic set, $K\dim(V) = d$, $p \in F[x]$, $V \not\subseteq V(p)$, and $W$ is an irreducible component of $V \cap V(p)$. Then $K\dim(W) = d - 1$. The same is true in the projective case.

Proof: Let $P = I(V)$; then $I(W)$ is a prime ideal $Q \subseteq F[x]$ containing $P$. Letting $\bar{p}$ denote the image under the canonical epimorphism, $\bar{Q}$ is a prime ideal in $F[V]$. Also, $\bar{p} \in \bar{Q}$, and by the hypothesis that $V \not\subseteq V(p)$, $\bar{p} \neq 0$. If $\bar{Q}$ were not minimal then $W$ would not be maximal. Thus, by theorem 15 $\bar{Q}$ has height 1, indeed by corollary 16 $K\dim(W) = d - 1$. The projective case follows by choosing $i$ with $V \cap V(p) \cap A_i \neq \emptyset$.

An algebraic set of the form $V(p)$ for a single polynomial $p$ is called a hypersurface. Some further observations required for the next theorem are as follows.

- If $V_1, V_2$ are projective algebraic sets then $K\dim(V_1 \cup V_2) = \max(K\dim(V_1), K\dim(V_2))$. This follows by lemma 2.d.
- If $I$ is an ideal in a ring commutative ring $R$ then $\text{Ann}_R(R/I) = I$.

Suppose $0 \to M' \to M \to M'' \to 0$ is a short exact sequence of finitely generated $R$-modules, where $R$ is Noetherian.

- $H_f M = H_f M' + H_f M''$, immediately from the definition.
- $H_p M = H_p M' + H_p M''$, since this holds at all sufficiently large $n$.
- $\deg(H_p M) = \max(\deg(H_p M'), \deg(H_p M''))$, since the leading term must have positive coefficient. Recall that the degree of the 0 polynomial is defined to be $-\infty$, whence the claim holds if either or both of $H_p M'$ and $H_p M''$ is 0.
- $\text{Rad}(\text{Ann}(M)) = \text{Rad}(\text{Ann}(M')) \cap \text{Rad}(\text{Ann}(M''))$. One inclusion follows by theorem 20.26 and the fact that $\text{Ann}(M) \subseteq \text{Ann}(M'), \text{Ann}(M'')$. For the other, if $r^n \in \text{Ann}(M')$, $r^m \in \text{Ann}(M'')$, and $x \in M$ then $r^{m+n} x = 0$.

368
Theorem 27. Suppose $F$ is an algebraically closed field and $M$ is a finitely generated $F[x_1, \ldots, x_{n+1}]$-module. Then the degree of $H_{P М}$ equals $Kdim(V_P(Ann(M)))$.

Proof: By the remarks preceding the theorem and lemma 25, we may assume that $M$ is $F[x]/P$ where $P$ is a homogeneous prime ideal, and it suffices to show that $H_{P_M}$ equals $Kdim(V(P))$. The proof is by induction on $d = Kdim(V(P))$. If $d = 0$ then $P$ is the irrelevant ideal, so $V(P) = \emptyset$, so $Kdim(V(P)) = -\infty$; also, $M = F$, so $H_{P_M}$ is eventually 0, so $H_{P_M} = 0$, and its degree is $-\infty$. If $d > 0$ then $P$ is not the irrelevant ideal; suppose that $x_i \notin P$. The map $p + P \mapsto x_i p + P$ from $M$ to $M$ is injective; let $M''$ be the cokernel. By the additivity of dimension, $H_{P_M}(l) = H_{P_M}(l-1) + H_{P_M''}(l)$. Letting $W = V(Ass(M''))$, we claim that $Kdim(W) = d - 1$. Given this, by induction the degree of $H_{P_M''} = d - 1$. It follows by the remarks preceding theorem 7.8 that the degree of $H_{P_M}$ equals the degree of $H_{P_M''}$, plus 1, i.e., $d$. To prove $Kdim(W) = d - 1$, let $H_i$ be the hyperplane $x_i = 0$. Letting $\phi$ denote the composition of multiplication by $x_i$ with the canonical epimorphism $F[x] \mapsto M$, $p \in F[x]$ is in the kernel of $\phi - x_i q \in P$ for some $q \in F[x]$: that is, the kernel of $\phi$ is $P + x_i F[x]$, so $Ann(M'') = P + x_i F[x]$, so $V(Ann(M'')) = V(P \cup x; F[x]) = V(P) \cap V(x_i F[x]) = V(P) \cap H_i$. If $V(P) \subseteq H_i$ were true, then $x_i F[x] \subseteq P$ would be, contradicting the assumption that $x_i \notin P$. Thus, $Kdim(W) = d - 1$ by lemma 26.

In particular the degree of the Hilbert polynomial of $F[V]$ equals the dimension of $V$. There is a definition of a Hilbert polynomial in the affine case, but it requires generalizing the theory; see [CLO].

The Hilbert polynomial of $F[V]$ for a projective algebraic set $V$ yields further information, in addition to $Kdim(V)$. If $V \neq \emptyset$ the degree of $V$ may be defined as $d! a_d$, and denoted deg($V$). By theorem 7.8, this is an integer, indeed positive since the value of the polynomial is always nonnegative and the polynomial is not 0.

The value at $l$ of the Hilbert function for $P^n$ equals the number of monomials in $n + 1$ variables, of degree $l$. This equals $\binom{l + n}{n}$ (exercise 8). As a polynomial in $l$, this has leading coefficient $1/n!$, so $P^n$ has degree 1.

Suppose $V = V(p)$ is a projective hypersurface, where $p \in F[x_1, \ldots, x_{n+1}]$ is a homogeneous polynomial of degree $d$. Let $M$ be the cokernel of the map $q \mapsto pq$ from $F[x]$ to $F[x]$. Then $H_F(x)(l) = H_{F[x]}(l - d) + H_{F[x]}(l)$, from which $H_{P_M}(l) = \binom{l + n}{n} - \binom{l - d + n}{n}$. In the numerator of the right side, the coefficient of $l^n$ equals $1 - 1 = 0$, and the coefficient of $l^{n-1}$ equals $(n + \cdots) - ((n - d) + \cdots) = nd$; thus, the degree equals $d$.

8. Bezout’s theorem. The basic form of Bezout’s theorem is a statement about curves in the projective plane. It has a variety of generalizations; we will give one from [Harts].

If $R$ is a Noetherian commutative ring and $M$ is a finitely generated $R$-module, a minimal prime ideal of $M$ is defined to be a minimal element of $\{ P : P$ prime, $Ann(M) \subseteq P \}$. Recalling the notion of an associated prime mentioned in section 7, a minimal prime ideal is an associated prime ideal; we omit a proof, which can be found in [Eisenbud] or [Matsumura].

Lemma 28. Suppose $R$ is a Noetherian commutative ring, $M$ is a finitely generated $R$-module, $0 = M_0 \subset \cdots \subset M_m = M$ is a chain of submodules, such that for $i > 0 M_i/M_{i-1}$ is isomorphic to $R/P_i$ for some prime ideal $P_i$. If $P$ is a prime ideal then $Ann(M) \subseteq P$ iff $P_i \subseteq P$ for some $i > 0$. In particular, the minimal prime ideals of $M$ are the minimal prime ideals among $\{ P_1, \ldots, P_n \}$. The number of times that a minimal prime ideal $P$ of $M$ occurs as some $P_i$ is independent of the chain, equaling the dimension of $M_P$ as a vector space over the residue class field of the ring $R_P$.

Proof: By a remark preceding theorem 27, and induction, $Rad(Ann(M)) = \cap_{i > 0} Rad(Ann(M_i/M_{i-1}))$. Thus, $Rad(Ann(M)) = \cap_{i > 0} P_i$, and by remarks preceding lemma 20.29, $Rad(Ann(M)) = P_i$ for some $i$. For this $i$, $Ann(M) \subseteq P$ iff $Rad(Ann(M)) \subseteq P$ iff $P_i \subseteq P$. Suppose $P$ is a minimal prime ideal, and $S = R - P$. Since $R \mapsto R_P$ is a flat homomorphism, the $M_{iP}$ form a chain, and $M_{iP}/M_{i-1,P}$ is isomorphic
to \((M_i/M_{i-1})_P\). If \(P \neq P_i\) then since \(P_i\) and \(P\) are minimal, there is some \(s_0 \in P_i - P\). Since \(s_0(r + P_i) = 0\) in \(R/P_i\), every fraction in \((R/P_i)_P\) is equivalent to 0/1. Thus, in this case, \(M_iP\) equals \(M_{i-1,P}\). If \(P = P_i\), \((R/P)_P\) is the residue class field \(K\) of \(R/P\). It follows that \(M_P\) is the \(R_P\)-module \(K^m\) where \(m\) is the number of times that \(P\) occurs.

In fact the dimension of \(M_P\) over the residue class field \(K\) of \(R/P\) equals the length of \(M_P\) as an \(R/P\) module (exercise 10). The dimension of \(M_P\) is called the multiplicity of \(M\) at \(P\). Suppose \(V\) is an irreducible projective algebraic set, \(\text{Kdim}(V) = d\), \(p \in F[x]\), \(V \nsubseteq V(p)\), and \(W\) is an irreducible component of \(V \cap V(p)\). Let \(M = F[x]/(I(V) + I(V(P)))\); then similarly to an argument in the proof of theorem 27, \(V(\text{Ann}(M)) = V \cap V(p)\). Since \(\text{Kdim}(W) = d - 1\) by lemma 26, \(I(V)\) is a minimal homogeneous prime ideal of \(M\). The intersection multiplicity of \(V\) and \(V(p)\) along \(W\), denoted \(i(V, V(p); W)\), is defined to be the multiplicity of \(M\) at \(I(W)\).

**Theorem 29.** Suppose \(V\) is an irreducible projective algebraic set, \(p \in F[x]\), \(V \nsubseteq V(p)\), and \(W_1, \ldots, W_s\) are the irreducible components of \(V \cap V(p)\). Then \(\sum_{i=1}^s i(V, V(p); W_i) \deg(W_j) = \deg(V) \deg(V(p))\).

**Proof:** Letting \(M\) be the cokernel of the map \(q \mapsto pq\) from \(F[V]\) to \(F[V]\), by the usual argument \(\text{Hilb}_M(l) = \text{Hilb}_{F[V]}(l) - \text{Hilb}_{F[V]}(l - d)\). If \(\deg(p) = d\), and \(V\) has dimension \(r\) and degree \(e\), the leading coefficient of the right side equals \(de/(r - 1)!\). Taking a chain \(0 = M_0 \subseteq \cdots \subseteq M_q = M\) as in lemma 25, with \(M_i/M_{i-1}\) isomorphic to \(F[x]/P_j\) for \(i > 0\), \(\text{Hilb}_M = \sum_j \text{Hilb}_{F[x]/P_j}\). Since \(M = F[x]/(I(V \cap V(P)))\), \(\text{Hilb}_{F[x]/P_j}\) has degree \(r - 1\) iff \(P_j\) is a minimal prime among \(\{P_j\}\), and is \(I(W_i)\) for one of the \(W_i\). The leading coefficient of \(\text{Hilb}_{F[x]/P_j}\) for such a \(j\) equals \(\deg(W_j)\), and the theorem follows.

A projective plane curve is defined to be a hypersurface in \(P^2\). If \(C_1\) and \(C_2\) are distinct irreducible projective plane curves, their intersection is a finite set of points. Since \(F[x]/I(a)\) for a point \(a\) is \(F\), the Hilbert polynomial is the constant polynomial 1. The following corollary of the theorem, called Bezout’s theorem, is immediate.

**Corollary 30.** If \(C_1\) and \(C_2\) are distinct irreducible projective plane curves, intersecting in points \(a_1, \ldots, a_s\), then \(\sum_{i=1}^s i(C_1, C_2; a_i) = \deg(C_1) \deg(C_2)\).

Suppose \(p_1, p_2\) define projective plane curves \(C_1, C_2\), intersecting at a point \(a\). We have not introduced a notation for the rational functions in the affine case; let \(O(a)\) denote those that are defined at \(a\). The intersection multiplicity is often defined as \(O(a)/[p_1^i, p_2^j]\), where the affine space may be any \(A_i\), or indeed any copy of \(F^n\), containing \(a\). In [Fulton] for example, it is shown that this is the unique definition satisfying certain properties. The two definitions are equivalent (exercise 11).

9. **Prime spectrum of a ring.** The prime ideals of a commutative ring can be made the points of a topological space, in a manner analogous to Stone duality for distributive lattices. Indeed for Boolean rings the resulting space is the same. The prime spectrum is of interest in algebraic geometry, in particular to the definition of schemes, to be given in the next section.

Suppose \(R\) is a commutative ring. To begin with, let \(\text{Spec}(R)\) denote the set of prime ideals of \(R\). For \(S \subseteq R\) let \(V(S) = \{P \in \text{Spec}(R) : S \subseteq P\}\). This is an order preserving map from the subsets of \(R\) ordered by inclusion, to the subsets of \(\text{Spec}(R)\) ordered by reverse inclusion. It is right adjoint to the map \(T \mapsto \cap T\). In particular, \(\cap V(S_i) = V(\cup_i S_i)\).

The following facts hold:
- If \(I = [S]\) then \(V(I) = V(S)\).
- For ideals \(I_1\) and \(I_2\), \(V(I_1 I_2) = V(I_1 \cap I_2) = V(I_1) \cup V(I_2)\). Indeed \(I_1 I_2 \subseteq I_1 \cap I_2 \subseteq I_j\); and if \(P \notin V(I_1) \cup V(I - 2)\) then for some \(r_1, r_2, r_j \in I_j - P\), and \(r_1 r_2 \in I_1 I_2 - P\).
- The sets in the image of $V$ are the closed sets of a topology on $\text{Spec}(R)$. This topology is called the Zariski topology.
- For an ideal $I$, $\cap V(I) = \text{Rad}(I)$; this follows by theorem 20.25.
- The radical ideals and the closed sets are in bijective correspondence under the Galois adjunction.
- The prime ideals correspond to the irreducible closed sets (exercise 12).
- $\text{Spec}(R)$ is a Noetherian space iff $R$ has no infinite ascending chain of prime ideals, in particular if $R$ is Noetherian. $R$ need not be Noetherian, though; see [MacDonald] for a counterexample.
- Since $V(S) = \cap_{r \in S} V(r)$, the sets $V(r)$ comprise a base for the closed sets, whence the sets $V(r)^c$ form a base for the open sets. We use $D_r$ to denote $V(r)^c$.

**Theorem 31.** $\text{Spec}(R)$ is a Coherent space.

**Proof:** Suppose $D_a = \cup_i D_{a_i}$, and let $I$ be the ideal generated by the $a_i$. Then $V(a) \supseteq V(I)$, so $\text{Rad}(a) \subseteq V(I)$, so $a^s \in I$ for some $s$, so $a^s = r_1 a_{i_1} + \cdots + r_t a_{i_t}$ for some $t, r_j, i_j$. It follows that $V(a) = V(a^s) = V(a_{i_1}) \cap \cdots \cap V(a_{i_t})$, and so $D_a = D_{a_1} \cup \cdots \cup D_{a_t}$. This shows that $D_a$ is compact. Since $R = D_1$, $R$ is compact. By the remarks preceding the theorem, $D_{r_1} \cap D_{r_2} = D_{r_1 r_2}$, and so the compact open sets are closed under $\cap$. Also by the foregoing remarks, $P \mapsto V(P)$ is a bijection from $\text{Spec}(R)$ to the irreducible closed sets, so $\text{Spec}(R)$ is sober.

Suppose $\phi : R_1 \mapsto R_2$ is a morphism in $\text{CRng}$, and let $\hat{\phi}$ be the map from $\text{Spec}(R_2)$ to $\text{Spec}(R_1)$, taking $Q \mapsto \phi^{-1}[Q]$. Then $Q \in \hat{\phi}^{-1}[V(I)]$ iff $\phi(Q) \in V(I)$ iff $\phi^{-1}[Q] \supseteq I$ iff $Q \supseteq \phi[I]$ iff $Q \in V(f[I])$. That is, $\hat{\phi}^{-1}[V(I)] = V(f[I])$, which shows $\hat{\phi}$ is continuous and $R \mapsto \text{Spec}(R)$ is the object map of a contravariant functor from $\text{CRng}$ to $\text{Top}$. Since $[\phi[P]]$ need not be prime in general (for example 5 factors in $\mathbb{Z}[i]$), it cannot be concluded that the functor is to $\text{CohTop}$, as defined in section 21.4.

**10. Schemes.** Schemes are a category of mathematical objects which generalize the notion of an algebraic set, and have proved useful in advancing not only algebraic geometry, but other branches as well. We will give only a brief introduction; a standard reference is [Harts]. The study of schemes began in the mid 1950’s.

The category of sheaves on a topological space $X$ in a category $C$ can be generalized, to let $X$ vary. An object is a pair $\langle X, O \rangle$ consisting of a topological space $X$, and a sheaf $O$ on $X$. A morphism from $\langle X, O_X \rangle$ to $\langle Y, O_Y \rangle$ is a pair $(\phi, \psi)$, where $\phi : X \mapsto Y$ is a continuous function, and $\psi$ is a system of morphisms $\psi_U : O_Y(U) \mapsto O_X(\phi^{-1}[U])$, for $U$ open in $Y$, which commute with the restriction maps (as in the following diagram).

$$
\begin{array}{ccc}
O_Y(U) & \xrightarrow{\psi_U} & O_X(\phi^{-1}[U]) \\
\rho_{U,V} & & \rho_{\phi^{-1}[U],\phi^{-1}[V]} \\
O_Y(V) & \xrightarrow{\psi_V} & O_X(\phi^{-1}[V])
\end{array}
$$

The requirement on $\psi$ may also be stated that it be a morphism of presheaves on $Y$ from $O_Y$ to $O_X^\phi$, the direct image of $O_X$ under $\phi$. It is readily verified that this construction yields a category. Henceforth we consider only the case where $C$ is $\text{CRng}$; this category is called the ringed spaces, and we use $\text{RngdSpc}$ to denote it.

We wish to enrich $\text{Spec}$ so that it is a functor to $\text{RngdSpc}$. The construction requires some further observations. If $B$ is a base for the topology of a topological space $X$ we may define a presheaf on $B$ as a functor $\hat{O}$ from $B$, rather than the entire topology $T$. The construction of theorem 24.22 can be modified to transform $\hat{O}$ to a sheaf $O$ on $T$. An alternative is to transform $\hat{O}$ to a presheaf; see [MacDonald] for this.

The modifications to the construction given in the proof of theorem 24.22 are as follows.
1. For $x \in X \{U \in B : x \in U\}$ is filtered, so $\text{Stlk}_x$ may still be defined as the direct limit; we denote it as $\mathcal{O}_x$.

2. In all mentions of $V$, $V$ is taken in $B$.

In particular, $\mu_V$ is defined for $V \in B$, and is a natural system of morphisms from $\mathcal{O}$ to $O$.

In the case of $\text{Spec}(R)$ for a commutative ring $R$, the base is $\{D_s : s \in R\}$. $D_s \supseteq D_t$ iff $V(s) \subseteq V(t)$ iff $\text{Rad}(Rs) \supseteq \text{Rad}(Rt)$ iff $t^k = as$ for some integer $k$ and $a \in R$. The map $r/s^n \mapsto a^r/r/t^k$ from $R_s$ to $R_t$ depends only on $s$ and $t$, and not on $k$ and $a$, because $a^r/r/t^k = a^{r/r}(t^k/n)$ follows from $t^k = as$ and $t^n = as$. Letting $\rho_{st}$ denote this map, we may define a presheaf $\mathcal{O}$ on $B$ by letting $\mathcal{O}(D_s) = R_s$ and $\rho_{D_s,D_t} = D_{st}$. Then $\rho_{st} \rho_{ts} = \rho_{st}$ is readily verified, noting that if $t^k = as$ and $u^l = at$ then $u^k = b^k \cdot s$.

If $P \in \text{Spec}(R)$ then $P \in D_s$ iff $s \not\in P$. We claim that $R_P$ is the direct limit of these $D_s$, where the map $\rho : s : R_s \mapsto R_P$ is $r/s^n \mapsto r/s^n$. This system forms a cone, because if $t^k = as$ then $r/s^n = a^r/r/t^n$. Given any other cone to $R'$, if $r/s \in R_P$ then $s \not\in P$ and $r/s \in R_s$, and the image of $r/s$ in $R'$ is determined. We have thus shown that $\mathcal{O}_P = R_P$.

Writing $\mu_s$ for $\mu_{D_s}$, we next claim that $\mu_s$ is injective. For any $P \in D_s$, $r/s^n = 0$ in $R_P$ iff $tr = 0$ for some $t \notin P$ iff $\text{Ann}(r) \not\subseteq P$. Thus, if $\mu_s(r/s^n) = 0$ then $\text{Ann}(r) \not\subseteq P$ for any $P \in D_s$, so $s \in P$ if $\text{Ann}(r) \subseteq P$, so $s \in \text{Rad}(\text{Ann}(r))$, so $s^{m}r = 0$ for some $m$, so $r = 0$ in $R_s$.

We claim that $\mu_s$ is also surjective. If $f : D_s \mapsto K$ then there is a cover $\{D_t\}$ of $D_s$ such that on $D_t$, $f$ is represented by $g_t$, (i.e., equals $\tilde{g}_t$) where $g_t \in D_t$. By theorem 1, we may take the cover as finite. Say $g_t = r_{ij}/t_{ij}$; since $D_t^{ij} = D_t$, we may assume $k_t = 1$ for all $t$. Since $r_{ij}/t_i$ and $r_{ij}/t_j$ both represent $f$ on $D_{t_i,t_j}$, and $\mu_{t_i,t_j}$ is injective, there is a $k_{ij}$ such that $(t_it_j)^k (r_{ij} - r_{ij}) = 0$. Replacing $k_{ij}$ by $k = \max\{k_{ij}\}$, and $r_{ij}/t_i$ by $r_{ij}/t_{ij}^{k+1}$, we may assume $r_{ij} = t_{ij}$. Since the $D_t$ cover $D_s$, there is an integer $m$ and $b_i \in R$ such that $s^m = \sum b_j r_j$. Then $r_t = \sum b_j r_j t_j = \sum b_j t_j r_j = rs^m$ where $r = \sum b_j t_j$. Thus, $f$ is represented by $r/s^m$ on $D_t$ for every $i$.

We have shown that $O$ is also the “same as” $\mathcal{O}$ on $\{D_t\}$. That is, there is a sheaf $O$ on $\text{Spec}(R)$, which maps $D_s$ to a ring of functions isomorphic to $R_s$. Note that $\mathcal{O}_P = R_P$. From hereon Spec will be considered as mapping $R$ to the object $(\text{Spec}(R), O)$ in $\text{RngdSpc}$.

Spec is in fact a contravariant functor to $\text{RngdSpc}$. First, suppose $\mathcal{O}_i$ is a presheaf on a base for the topology of $X_i$, for $i = 1, 2$. Suppose $\phi : X_1 \mapsto X_2$ is a continuous function, and further if $V$ is a basic open set in $X_2$ then $\phi^{-1}[V]$ is a basic open set in $X_1$. Finally suppose $\psi$ is a system of morphisms $\psi_V : O_2(V) \mapsto O_1(\phi^{-1}[V])$, for basic open $V$. Then a map $\psi$ can be constructed, making $(\phi, \psi)$ a morphism from $(X_1, O_1)$ to $(X_2, O_2)$.

Let $\tilde{\psi}_x$ denote the map from $\mathcal{O}_{2,\phi(x)}$ to $\mathcal{O}_{1x}$ induced by the cone $\{\rho_{1,\phi^{-1}[V],x} \tilde{\psi}_V : \phi(x) \in V\}$. For $f \in O_2(U)$ and $x \in \phi^{-1}[U]$ let $\psi_U(f)(x) = \tilde{\psi}_x(f(\phi(x)))$. It must be verified that $\psi_U(f) \in O_1(\phi^{-1}[U])$. Given $x \in \phi^{-1}[U]$ there is a $V$ with $\phi(x) \in V \subseteq U$, and a $g_2 \in \mathcal{O}_2(V)$, such that $f \mid V = g_2$. Then $x \in \phi^{-1}[V] \subseteq \phi^{-1}[U]$ and $g_2 \psi_V(g_2) \in \mathcal{O}_1(\phi^{-1}[V])$. The reader may verify that $\psi_U(f) \mid \phi^{-1}[V] = \tilde{g}_1$. The reader may also verify that $\psi_V(\tilde{g}_2) = \tilde{g}_1$, from which $\psi$ agrees with $\psi$ on basic open sets.

Suppose $\phi : R_2 \mapsto R_1$ is a ring homomorphism. For each $s \in R_2$ there is an induced homomorphism $\psi_s : R_{2,s} \mapsto R_{1,s}$, mapping $r/s^n$ to $\phi(r)/\phi(s)^n$. Also, $\phi$, where $\hat{\phi}(Q) = \phi^{-1}[Q]$, is a continuous function from $\text{Spec}(R_1)$ to $\text{Spec}(R_2)$; and $\phi^{-1}[D_s] = D_{\phi(s)}$. By the construction just given, $\psi$ can be extended to a natural system of morphisms $\psi$ from the sheaf $O_2$ on $\text{Spec}(R_2)$ to the sheaf $O_1$ on $\text{Spec}(R_1)$. This completes the definition of the contravariant functor $\text{Spec}$ from $\text{CRng}$ to $\text{RngdSpc}$.

Again considering a ring homomorphism $\phi : R_2 \mapsto R_1$, we noted above that in the sheaf $O_1$, the stalks are the local rings $R_{1,P}$ for the prime ideals $P \subseteq R_1$. Letting $\psi_P : R_{2,\phi^{-1}[P]} \mapsto R_{1,P}$ denote the induced map, a diagram chase shows that $\psi_P(r/s) = \phi(r)/\phi(s)$. If $R_1, R_2$ are local rings, with maximal ideals $M_1, M_2$, a homomorphism $\phi : R_2 \mapsto R_1$ is said to be a local homomorphism if $\phi[M_2] \subseteq M_1$. It is readily verified that $\psi_P$ is a local homomorphism.
Given a sheaf \( O \) on a topological space \( X \), and an open subset \( U \subseteq X \), the restriction \( O|_U \) of \( O \) to \( U \) is simply the restriction of the functor to the full subcategory of open subsets of \( U \). More general restrictions can be defined; but when \( U \) is open the stalks for \( x \in U \) remain the same as in \( X \).

A scheme is defined to be a ringed space \( (X, O) \), where \( X \) has an open cover \( \{U_i\} \) such that for each \( i \), \( \langle U_i, O|_U \rangle \) is isomorphic as a ringed space to \( \text{Spec}(R) \) for some commutative ring \( R \). The resemblance to the open cover of \( P^n \) by copies of \( F^n \) is noteworthy; ringed spaces isomorphic to \( \text{Spec}(R) \) for some commutative ring \( R \) are called affine schemes.

Note that the stalks of a scheme are local rings. In the category of schemes, morphisms are defined to be ringed space morphisms, such that the induced map on any stalk is a local homomorphism. By abuse of notation, schemes are denoted by their topological space \( X \), and morphisms by the continuous function \( f : X \to Y \).

Schemes of interest in algebraic geometry usually satisfy additional requirements. For example, a scheme is said to be integral if \( O(U) \) is an integral domain for every open \( U \).

An important restriction is that of a separated morphism, and a separated scheme. However, as this requires defining the product of schemes over a scheme, which is of some length, we leave off here, and refer the reader to the literature for further study of schemes.

11. Groebner bases. The fact that any ideal in \( F[x] = F[x_1, \ldots, x_n] \) is finitely generated is known as the Hilbert basis theorem. Originally it was proved for the polynomial rings; later the machinery of Noetherian rings was introduced, providing a more general setting for facts about the ring of multi-variable polynomials over a field, resulting in a various simplifications. More recently, mathematics has come full circle and considered the polynomial rings again, due to the fact that certain generating sets for an ideal (Groebner bases) have additional properties, of interest to computational commutative algebra and algebraic geometry. These generating sets had been considered previously, with interest in them intensifying starting around 1970.

Again, we give only a brief introduction. Standard references for further study include [CLO] and [Eisenbud]. We prove some facts about monomials. Recall that a monomial \( m \) may be written as \( x^v \) where \( v \in \mathbb{N}^n \); \( \deg(m) \) may be written for \( v \). Also recall that there is a natural partial order on \( \mathbb{N}^n \), the product order, which we denote in this section as \( \leq_p \). The following lemma is called Dickson’s lemma. This in fact follows from the Hilbert basis theorem, but it is instructive to give a direct proof.

**Lemma 32.** For \( S \subseteq \mathbb{N}^n \), \( S \) has only finitely many minimal elements in the product order.

**Proof:** Since the minimal elements of \( S \) and those of \( S^\circ \) are the same, we may assume that \( S \) is \( \geq \)-closed. If \( n = 1 \) the claim follows because \( \mathbb{N} \) is well-ordered. For \( n > 1 \), let \( S_i = \{ \langle u_1, \ldots, u_{n-1} \rangle : \langle u_1, \ldots, u_{n-1}, i \rangle \in S \} \). Inductively, each \( S_i \) has only finitely many minimal elements. The same holds for \( \cup_i S_i \), whence \( \cup_i S_i = \cup_{i \leq t} S_i \) for some \( t \). A minimal element of \( S \) must be of the form \( \langle u_1, \ldots, u_{n-1}, i \rangle \) where \( i \leq t \) and \( \langle u_1, \ldots, u_{n-1}, i \rangle \) is a minimal element of \( S_i \).

An ideal \( I \subseteq F[x] \) is called a monomial ideal iff it is generated by a set of monomials. By theorem 18.6 applied to \( \mathbb{N}^n \)-graded modules, this is so iff for a polynomial \( p \), \( p \in I \) iff each \( x^p \in I \). Clearly, a monomial is in a monomial ideal \( I \) iff it is a multiple of some generator. It follows easily from Dickson’s lemma that a monomial ideal is finitely generated.

Next we consider a class of total orders on the monomials, which have been seen to be of great interest. A total order \( \leq \) on \( \mathbb{N}^n \) is said to be a monomial order if \( u + w \leq v + w \) whenever \( u \leq v \) and \( 0 \) is a least element. Such induces an order on the monomials or the terms via the degrees. Examples of monomial orders on \( \mathbb{N}^n \) include the following.

- Lexicographic: \( u < v \) iff in the first \( i \) position where they differ, \( u_i < v_i \).
- Graded lexicographic: \( u < v \) iff either \( d_u < d_v \), or in the first position \( i \) where they differ, \( u_i < v_i \), where \( d_u \) and \( d_v \) are the total degrees.

- Graded reverse lexicographic: \( u < v \) iff either \( d_u < d_v \), or in the last position \( i \) where they differ, \( u_i > v_i \).

Verification that these are monomial orders is left to exercise 13.

If \( u \preceq p \) \( v \) then \( u \preceq v \) in any monomial order, since \( 0 \preceq v - u \). A monomial order is a well-order; by Dickson’s lemma a set \( S \) of monomials has a finite set of minimal elements in the product order, and the least of these is least in \( S \).

Suppose for the remainder of the section that \( \leq \) is a fixed monomial order on \( F[x] = F[x_1, \ldots, x_n] \). If \( p \in F[x] \), \( p \) may be written as \( \sum u c_u x^u \), where the monomials are in order, similarly to writing a single variable polynomial with the terms in order from highest to lowest degree. The term \( c_u x^u \) with \( u \) largest is called the leading term of \( p \); we denote it as \( \text{Lt}(p) \). As usual, for a set \( S \) of polynomials \( \text{Lt}(S) = \{ \text{Lt}(p) : p \in S \} \).

Finally, we write \( cx^u \) as \( dx^v \) if \( \text{lt}(u) \leq \text{lt}(v) \).

**Lemma 33.** If \( I \) is a nontrivial ideal then there are \( p_1, \ldots, p_t \in I \) such that \( \{ \text{Lt}(p_1), \ldots, \text{Lt}(p_t) \} \) is a generating set for \( [\text{Lt}(I)] \). Further any such \( p_1, \ldots, p_t \) generate \( I \).

**Proof:** \([\text{Lt}(I)] \) is a monomial ideal, and every monomial it contains is the leading monomial of some \( p \in I \). By Dickson’s lemma there are \( p_1, \ldots, p_t \in I \), whose leading terms generate \( [\text{Lt}(I)] \) as required. If \( q \in I \) then there is a \( q' \in [p_1, \ldots, p_t] \) with the same leading term. Proceeding inductively with \( q - q' \) proves the second claim.

A generating set for \( I \) with the property of the lemma is called a Groebner basis for \( I \). Note that the Hilbert basis theorem is an immediate consequence of the lemma.

Given an ordered sequence \( d_1, \ldots, d_t \) of polynomials in \( F[x] \), a remainder for a polynomial \( p \) can be computed by the following recursion.

\[
\text{Rem}(p) = \begin{cases} 
\text{Rem}(p - (\text{Lt}(p)/\text{Lt}(d_i))d_i) & \text{if } i \text{ is least such that } \text{Lt}(d_i) | \text{Lt}(p) \\
\text{Lt}(p) + \text{Rem}(p - \text{Lt}(p)) & \text{if no such } i \text{ exists.}
\end{cases}
\]

It is readily verified that \( p = \sum q_id_i + \text{Rem}(p) \), where none of the monomials in \( \text{Rem}(p) \) is divisible by any \( \text{Lt}(d_i) \), and \( \text{Lt}(q_id_i) \leq \text{Lt}(p) \) (because this is true at each step and the successive \( \text{Lt}(p) \) are decreasing in the monomial order). In general, \( \text{Rem}(p) \) depends on the order of the \( d_i \) (exercise 14).

**Theorem 34.** Suppose \( I \subseteq F[x] \) and \( p \in F[x] \). Then there is a unique \( r \in F[x] \) such that \( p - r \in I \), and no monomial of \( r \) is in \( [\text{Lt}(I)] \). \( \text{Rem}(p) \) equals this \( r \), with any \( d_1, \ldots, d_t \) such that \( \{d_1, \ldots, d_t\} \) is a Groebner basis for \( I \).

**Proof:** \( \text{Rem}(r) \) with respect to any ordered Groebner basis satisfies the requirements on \( r \), because \( \{\text{Lt}(d_1), \ldots, \text{Lt}(d_t)\} \) is a generating set for \( [\text{Lt}(I)] \). If \( r_1, r_2 \) satisfy the requirements then \( r_1 - r_2 \in I \), so \( \text{Lt}(r_1 - r_2) \) is divisible by some \( \text{Lt}(d_i) \), where \( \{d_1, \ldots, d_t\} \) is any Groebner basis for \( I \), which is a contradiction unless \( r_1 = r_2 \).

As a corollary, if \( \{d_1, \ldots, d_t\} \) is a Groebner basis for \( I \) then \( p \in I \) iff \( \text{Rem}(p) = 0 \), since \( p - \text{Rem}(p) \in I \).

This gives an algorithm for deciding membership in \( I \); simple compute \( \text{Rem}(p) \) by the procedure given above. The algorithm requires a Groebner basis; we will shortly show how to compute one from a generating set.

Suppose \( f, g \in F[x] \). Let \( w \) be the maximum under \( \leq_p \) of the degrees of \( \text{Lt}(f) \) and \( \text{Lt}(g) \). Define \( S(f, g) \) to be \((x^w/\text{Lt}(f))f - (x^w/\text{Lt}(g))g\); that is, \( S(f, g) \) is the polynomial obtained by taking the obvious \( F[x] \)-linear combination of \( f \) and \( g \) in which the leading terms cancel. The following theorem is known as Buchberger’s criterion.

**Theorem 35.** \( d_1, \ldots, d_t \) is a Groebner basis for \( \{d_1, \ldots, d_t\} \) iff \( \text{Rem}(S(d_i, d_j)) = 0 \) for all \( i, j \).

374
Proof: \( S(d_i, d_j) \subseteq [d_1, \ldots, d_i] \), so if \( \{d_1, \ldots, d_i\} \) is a Groebner basis then \( \text{Rem}(S(d_i, d_j)) = 0 \) for all \( i, j \). Suppose \( \text{Rem}(S(d_i, d_j)) = 0 \) for all \( i, j \). Suppose there is some \( p \in [d_1, \ldots, d_i] \) such that \( \text{Lt}(p) \notin [\text{Lt}(d_1), \ldots, \text{Lt}(d_i)] \). Choose \( p = \sum a_i d_i \) with the following properties.

1. Among such \( p, \) \( u \) is minimal, where \( u \) is the maximum under \( \leq \) of the exponents of the \( \text{Lt}(a_i d_i) \).

2. Among \( p \) having property 1, \( p \) has the smallest number of terms of degree \( u \).

It may be assumed that the \( a_i d_i \) with \( \text{Lt}(a_i d_i) \) of degree \( u \) are \( a_1 d_1, \ldots, a_m d_m \). By the hypothesis that \( \text{Lt}(p) \notin [\text{Lt}(d_1), \ldots, \text{Lt}(d_i)] \), \( m \geq 2 \). Since \( \text{Rem}(S(d_1, d_2)) = 0 \), there are \( b_i \) such that \( S(d_1, d_2) = \sum b_i d_i \) and \( \text{Lt}(S(d_1, d_2)) \geq \text{Lt}(b_id_i) \) for each \( i \). By definition, \( S(d_1, d_2) = (x^w/\text{Lt}(d_1))d_1 - (x^w/\text{Lt}(d_2))d_2 \) where \( x^w > \text{Lt}(b_id_i) \) for each \( i \). Since \( \text{Lt}(a_1 d_1) \) and \( \text{Lt}(a_2 d_2) \) have degree \( u \), \( w \leq p \); and there is some term \( e \) with \( \text{Lt}(a_1 d_1) = e x^w \). Then \( p = \sum a_i d_i - e((x^w/\text{Lt}(d_1))d_1 - (x^w/\text{Lt}(d_2))d_2 + \sum b_i d_i) \). Writing this as a sum \( a'_i d_i \), one verifies that \( \text{deg}(\text{Lt}(a'_i d_i)) \leq \text{deg}(\text{Lt}(a_i d_i)) \) for each \( i \), and \( \text{deg}(\text{Lt}(a'_i d_i)) < \text{deg}(\text{Lt}(a_i d_i)) \). This is a contradiction, so \( \{d_1, \ldots, d_i\} \) is a Groebner basis.

Buchberger’s algorithm for computing a Groebner basis for an ideal \( I \) from a generating set is to repeatedly compute the \( \text{Rem}(S(d_i, d_j)) \); if these are all 0, terminate. Otherwise add the nonzero \( \text{Rem}(S(d_i, d_j)) \) to the generating set. Since \( S(d_i, d_j) \in I \), \( \text{Rem}(S(d_i, d_j)) \in I \). Further, if termination does not occur, the ideal generated by the leading terms becomes strictly larger, since the polynomials added are remainders, so their leading terms are not in the current leading term ideal.

If \( \{d_1, \ldots, d_i\} \) is a set of polynomials, the kernel of the map \( \langle a_1, \ldots, a_i \rangle \mapsto \sum a_i d_i \) from the \( F[x] \)-module \( F[x]^i \) to \( F[x] \) is called the syzygy module of \( \{d_1, \ldots, d_i\} \). Among the syzygies are those arising from the expansions \( S(d_i, d_j) \) as \( \sum_k b_{ijk} d_k \). These in fact generate the syzygy module. The Hilbert syzygy theorem states that \( F[x_1, \ldots, x_n] \) has a free resolution of length at most \( n \). This may be proved by carrying the above methods further. See [Eisenbud] for a discussion of these facts.

Monomial orders have other uses.

Lemma 36. Suppose \( I \subseteq F[x] \) is an ideal. Let \( N \) be the monomials not in \( \text{Lt}(I) \). Under the canonical epimorphism \( \eta : F[x] \mapsto F[x]/I \), \( \eta[N] \) is a basis for \( F[x]/I \) as a vector space over \( F \).

Proof: \( \text{Rem}(p) - p \in I \), and \( \text{Rem}(p) + I \in \text{Span}(\eta[B]) \), so \( \eta[B] \) generates \( F[x]/I \). If \( \sum a_i n_i \in I \) for \( n_i \in N, a_i \in F \), \( \text{Lt}(\sum a_i n_i) \in \text{Lt}(I) \); but \( \text{Lt}(\sum a_i n_i) \) equals some \( a_i n_i \), a contradiction.

Theorem 37. Suppose \( I \subseteq F[x] \) is a homogeneous ideal. Then the Hilbert function of \( F[x]/I \) equals the Hilbert function of \( F[x]/\text{Lt}(I) \).

Proof: Let \( N \) be the monomials not in \( \text{Lt}(I) \). The monomials in \( N \) of degree \( l \) form a basis for \( (F[x]/I)_l \), and also for \( (F[x]/\text{Lt}(I))_l \).

Exercises.

1. a. Show that if \( V \subseteq F^n \) and \( W \subseteq F^m \) are algebraic sets then the Cartesian product \( V \times W \) is an algebraic set in \( F^{n+m} \). Hint: consider the union of defining sets of polynomials, with disjoint variables.

b. Show that \( V \times W \) is a product in the category of affine algebraic sets. Hint: the projections are polynomial maps, and the map induced in Set by polynomial maps is polynomial.

c. Show that \( F[V \times W] \) is isomorphic to \( F[V] \otimes F[W] \). Hint: This follows because \( V \mapsto F[V] \) is an equivalence of categories, and the tensor product of \( F \)-algebras is the coproduct.

d. Show that if \( V \) and \( W \) are irreducible then \( V \times W \) is irreducible. Hint: For \( y_0 \in W \), \( V \times \{y_0\} \) is isomorphic to \( V \) \( (x, y) \mapsto x \) yields an inverse to \( x \mapsto (x, y_0) \). Suppose \( Z \) is closed in \( V \times W \) and \( U = \{y : V \times y \nsubseteq Z\} \). If \( y_0 \in U \), say \( p(x_0, y_0) \neq 0 \) where \( p \in I(Z) \), then \( \{y : p(x_0, y) \neq 0\} \) is open. This
shows that $U$ is open. If $V \times W = Z_1 \cup Z_2$ where $Z_1$ is closed, let $U_i = \{ y : V \times y \not\subseteq Z_i \}$. One of the $U_i$ must be empty.

e. Show that the Zariski topology on $F^2$ is not the product of those on the factors.

2. Show that a $N^n$-graded ring is Noetherian iff $R_0$ is Noetherian and $R$ is finitely generated as an $R_0$-algebra. Hint: If $R$ is finitely generated and $R_0$ is Noetherian then $R$ is Noetherian, by the Hilbert basis theorem and facts from section 8.5. Suppose $R$ is Noetherian. The elements not in $R_0$ are an ideal $R^+$, and $R_0$ is isomorphic to $R/R^+$, $R^+$ is a homogeneous ideal, so has a finite homogeneous generating set $g_1, \ldots, g_t$. It suffices to show that $A_v \subseteq A_0 [g_1, \ldots, g_t]$ for any $v \in N^n$. Suppose otherwise, and let $v$ be minimal in the product order $\leq$ on $N^n$. Clearly $v \neq 0$, so any $x \in R_v$ equals $\sum_i y_i g_i$. Each $y_i$ is in $R_w$ where $w \leq v$ and $w \neq v$.

3. Show that a finitely generated $R$-module over an Artinian commutative ring $R$ has finite length. Hint: $R$ is Noetherian (Theorem 16.27). Hence $M$ is Noetherian and Artinian (Remarks at the start of sections 8.5 and 16.4). Hence $M$ has finite length (Theorem 16.11).

4. Show that if $M$ is a $\mathbb{Z}$-graded module finitely generated over a finitely generated $N^n$-graded $F$-algebra $R$, then each homogeneous submodule $M_\ell$ is finitely generated. Hint: Let $m_1, \ldots, m_t$ be a finite set of homogeneous generators for $M$ over $R$. There are only finitely many elements $rm_i$ in $M_v$, where $r$ is a monomial in the generators of $R$, and these generate $M_\ell$ over $R_\ell$.

5. Show that if $I$ is a homogeneous ideal in an $N$-graded ring then $\text{Rad}(I)$ is homogeneous. Hint: For $x \in \text{Rad}(I)$ suppose $x = x_0 + \ldots + x_t$ where $x_i$ is homogeneous of degree $i$, and $x_n \in I$. Then $x_n^i$ is the highest degree part of $x^n$, so $x_n^i \in I$, so $x_0 + \ldots + x_{t-1} \in \text{Rad}(I)$.

6. Verify that the points at infinity of $P^n$ form a copy of $P^{n-1}$.

7. Show that if $R$ is an integral domain, and $s \in R$, then $R_s$ is isomorphic to $R[x]/[sx - 1]$. Hint: Since $x + [sx - 1]$ is the multiplicative inverse of $s + [sx - 1]$ in $R' = R[x]/[sx - 1]$, by Theorem 6.7.c the map $r/s^i \mapsto rx^i + [sx - 1]$ is a homomorphism from $R_s$ to $R'$. It is injective, because if $rx^i = q(sx - 1)$ for some $q$ then $r = 0$, by induction on $i$. It is surjective, because $R$ and $x + [sx - 1]$ are in the image.

8. Show that the number of monomials in $n + 1$ variables, of degree $l$, equals $\binom{n+l}{n}$. Hint: induction on $m_n$. Write a monomial $\mu$ as $\mu_1 + x_{n+1}\mu_2$, and use $\binom{n+l}{n} = \binom{n}{n} + \binom{n+l-1}{n-1}$.

9. Suppose $R$ is a commutative ring, $M$ is an $R$-module, $I \subseteq R$ is an ideal, and $S = R/I$. Show that $S \otimes M$ is isomorphic to $M/[IM]$, where $[IM]$ is the submodule of $M$ generated by $IM$. Hint: The map $\langle r + I, m \rangle \mapsto rm + [IM]$ is bilinear, so induces a map $r + I \otimes m \mapsto rm + [IM]$ from $S \otimes M$ to $M/[IM]$. If $rm \in [IM]$, say $rm = \sum_j i_j m_j$, then $(r + I) \otimes m = (1 + I) \otimes m = (1 + I) \otimes \sum_j i_j m_j = 0$. The map is clearly surjective.

10. Show that under the hypotheses of Lemma 28, the dimension of $K^m$ over the residue class field $K$ of $R/P$ equals the length of $M_P$ as an $R/P$ module. Hint: Certainly the length is at least $m$, since an ascending chain of subspaces of dimension $0, 1, \ldots, m$ is an ascending chain of submodules. On the other hand, suppose $0 = M_0 \subset \cdots \subset M_l = K^m$ is an ascending chain of submodules. It suffices to show that for $i > 0$, $K \otimes M_{i-1} \subset K \otimes M_i$, since then $l \leq m$ follows. For this it suffices to show that if $R$ is a local ring, with maximal ideal $P$ and residue class field $K = R/P$, and $M \subset N$ are $R$-modules, then $K \otimes M \subset K \otimes N$. Suppose to the contrary that $K \otimes M = K \otimes N$, and $M$ is generated by $m_1, \ldots, m_t$; then $K \otimes M$ is generated by $1 \otimes m_1, \ldots, 1 \otimes m_t$. Letting $L \subseteq N$ be the submodule generated by $m_1, \ldots, m_t$, by Exercise 9 $M = L + [PM]$.

By Lemma 16.23.b with $M = M/L$, $M/L = 0$, i.e., $M = L$.

11. Show that, as stated at the end of section 8, $i(C_1, C_2; a) = O(a)/[p_1^i, p_2^j]$. Hint: Let $P$ be the ideal $[a_3x_1 - a_1x_2, a_2x_2 - a_2x_1]$ in $F[x_1, x_2, x_3]$. Since $F[x_1, x_2, x_3]_P$ equals $F$, both definitions are the dimension of a vector space over $F$. For $i(C_1, C_2; a)$ the space is $F[x_1, x_2, x_3]_P/[p_1, p_2]_P$. Note that $[p_1^i, p_2^j]$ is that generated in $O(a)$.

12. Show that under the Galois adjunction between $R$ and $\text{Spec}(R)$, the prime ideals and irreducible
subsets correspond. Hint: Suppose $V(I_1) \cup V(I_2) = V(P)$ where $P$ is prime and $I_j$ is radical. Certainly $I_j \supseteq P$; and $I_1 \cap I_2 = P$, so $I_1 \subseteq P$ or $I_2 \subseteq P$, and $V(P)$ is not irreducible. Since $V(r_1) \cup V(r_2) = V(r_1 r_2)$, if $I$ is not prime then $V(I)$ is reducible.

13. Verify that the examples given at the start of section 11 are monomial orders. Hint: In the lexicographic order, if two elements are not equal then there is a first position where they differ, so they are comparable. Adding a vector to to each of two vectors does not change their relationship in the lexicographic order. 0 is clearly the least element. In the graded lexicographic order, either one monomial has degree less than the other, or the monomials are related in the lexicographic order. Adding a vector does not change the relationship. The same arguments apply to the graded reverse lexicographic order.

14. Show that, with the lexicographic order, $d_1 = x_1^2$, $d_2 = x_1 x_2 - x_2^2$, and $p = x_1^2 x_2$, Rem($p$) = 0. Reversing $d_1$ and $d_2$ yields Rem($p$) = $x_2^3$. 
26. Algebraic number theory.

1. Basic facts. Algebraic number theory is concerned with finite extension fields of $\mathbb{Q}$, which are called algebraic number fields. It is fair to say that it has occupied a central place in the history of mathematics since its founders, who include Dedekind, Frobenius, Hilbert, and Kronecker. Tools from algebra of use in studying algebraic number fields include Dedekind domains, valuations, integral extensions, and Galois theory. Often this application has been an impetus in their development.

Other branches of number theory benefit from the use of algebra. The Hasse-Minkowski theorem, for example, states that a nondegenerate quadratic form represents 0 over $\mathbb{Q}$ if and only if it does over $\mathbb{R}$, and over $\mathbb{Q}_{\text{pad}}$ for each prime $p$. For a proof, see [Serre].

We begin by recalling some topics already covered. In this chapter, we will abbreviate “algebraic number field” as “a.n.f.” The a.n.f.’s are subfields of the algebraic closure of $\mathbb{Q}$, which in turn is a subfield of $\mathbb{C}$. Define the dimension of an a.n.f. to be its dimension over $\mathbb{Q}$.

- If $K$ is an a.n.f., of dimension $n$, and if $L$ is a finite extension of $K$, of dimension $m$ over $K$, then $L$ is an a.n.f., of dimension $mn$.
- If $K$ is an a.n.f., of dimension $n$, $L$ is an extension of $K$, and $L$ is an a.n.f. of dimension $m$, then $n/m$ and $L$ is a finite extension of $K$, of dimension $m/n$ over $K$.
- If $K$ and $L$ are a.n.f.’s then their join is an a.n.f.
- Any extension of a.n.f.’s is perfect (exercises 9.3, 9.4); thus, normal extensions of a.n.f.’s are Galois.
- Suppose $K$ is an a.n.f. and $L$ is a finite extension of $K$, of dimension $m$ over $K$. There is a multiplicative homomorphism $N_{L/K}$, called the norm; and a $K$-linear transformation $\text{Tr}_{L/K}$, called the trace. For $a \in K$, $N_{L/K}(a)$ ($\text{Tr}_{L/K}(a)$) equals the determinant (trace) of the matrix representing multiplication by $a$ in any basis for $L$ over $K$. If the minimal polynomial $p$ for $a$ has degree $d$, $N_{L/K}(a)$ ($\text{Tr}_{L/K}(a)$) equals the product (sum) of the (distinct) roots of $p$, each raised to the power (multiplied by) $n/d$.

Lemma 1. Suppose $L \supseteq K$ is a finite separable extension of dimension $n$, $\sigma_1, \ldots, \sigma_n$ are the embeddings of $L$ over $K$ in a normal closure of $K$, and $l \in L$ is of degree $d$. Then in the sequence $\sigma_1(l), \ldots, \sigma_n(l)$, each conjugate of $l$ occurs $n/d$ times. In particular, $N(l) = \prod \sigma_i(l)$, and $\text{Tr}(l) = \sum \sigma_i(l)$.

Proof: Each embedding takes $l$ to one of its conjugates; and for each conjugate, there are $n/d$ extensions of the embedding of $K(l)$.

The algebraic integers are defined to be the algebraic numbers, which are integral over $\mathbb{Z}$. Thus, they are those algebraic numbers whose irreducible polynomial over $\mathbb{Q}$ may be written as a monic polynomial with coefficients in $\mathbb{Z}$. By corollary 20.18 the algebraic integers are a subring of the algebraic numbers. It follows that the algebraic integers contained in an a.n.f. $K$ are a subring of $K$, called the ring of integers of $K$. We will use $O_K$ to denote it. Note that $O_K$ is the integral closure of $\mathbb{Z}$ in $K$.

We will call a ring which is $O_K$ for some $K$ an algebraic integer ring, and abbreviate this as a.i.r. Not every subring of the algebraic integers is an a.i.r.; as observed in section 20.6, $\mathbb{Z}[\sqrt{-3}]$ is not integrally closed.

Lemma 2. Suppose $A$ is an integrally closed integral domain, $K$ is its field of fractions, $L$ is a finite separable extension of $K$, and $B$ is the integral closure of $A$ in $L$.

a. For $b \in B$, the coefficients of the irreducible polynomial of $b$ over $K$, and the norm and the trace, are in $A$.

b. $B$ is finitely generated as an $A$-module.

Proof: For part a, since the quantities in question are in the ring generated over $A$ by the conjugates of $b$ they are all integral over $A$, and in $K$. Since $A$ is integrally closed, they are in $A$. For part b, choose a basis $e_1, \ldots, e_n$ for $L$ over $K$. By multiplying by an element of $A$ we may assume that each $e_i$ is integral over $A$, 378
and hence is in $B$. By theorem 15.8 and remarks following it there is a basis $e'_1, \ldots, e'_n$ for $L$ over $K$ with $\text{Tr}_{L:K}(e'_ie'_j) = \delta(i,j)$. Any $b \in B$ equals $\sum_i k_ie'_i$ where $k_i \in K$. $\text{Tr}(e_jb)$ is readily seen to equal $k_j$, and so $k_j \in A$ by part a.

**Theorem 3.** Suppose $A$ is a Dedekind domain, $K$ its field of fractions, $L$ is a finite separable extension of $K$, and $B$ is the integral closure of $A$ in $L$. Then $B$ is a Dedekind domain.

**Proof:** $B$ is integrally closed by theorem 20.20.c (it suffices that $A$ be an integral domain). $B$ is Noetherian by lemma 2 (it suffices that $A$ be an integrally closed Noetherian domain). Any prime ideal $P$ of $B$ lies above a prime ideal of $A$, which is maximal, and so by lemma 20.22 $P$ is maximal.

The separability restriction can be removed; see [Jacobson]. Since $Z$ is clearly a Dedekind domain, any a.i.r. is a Dedekind domain. There are direct proofs that unique factorization of ideals holds in an a.i.r.; see [IreRos]. Not every Dedekind domain is an a.i.r; examples are mentioned in [Eisenbud], following Corollary 11.9.

The maps $K \mapsto O_K$ and $A \mapsto A_{A\neq}$ are inverse maps between the a.n.f.’s and the a.i.r.’s, as is readily verified using facts from section 20.6.

By theorem 20.41 and remarks preceding it, an a.i.r. is a factorial domain iff it is a principal ideal domain. This is not always the case, a fact of great relevance in the history of number theory. For example the ring $Z[\sqrt{-5}]$ is an a.i.r., and 6 factors into primes as $(1 + \sqrt{-5})(1 - \sqrt{-5})$, and $2 \cdot 3$ (see [HardWrt]).

Under the hypotheses of lemma 2, a basis of $L$ over $K$, which is a basis for $B$ as a free $A$-module, is called an integral basis. When $A$ is a principal ideal domain, an integral basis exists. Indeed, it follows from lemma 2 and theorem 8.9 that $B$ is a free $A$-module. Further for each element of a basis for $L$ over $K$, some $A$-multiple must be in $B$. In particular any a.i.r. has an integral basis, i.e., over $Z$.

Suppose $R$ is a subring of $C$, which is finitely generated as a $Z$-module. Then for any $r \in R$, $R$ is a faithful $Z[r]$ module, so by theorem 20.16 any element in $R$ is an algebraic integer. An order in an a.i.r. $K$ is defined to be a subring of $K$ which is finitely generated as a $Z$-module, and has the same dimension over $Q$ as $K$. The orders in $K$ are partially ordered by inclusion, and $O_K$ is the greatest order.

**2. Discriminant.** If $L \supseteq K$ is a finite extension of dimension $n$, and $L = (l_1, \ldots, l_n)$ is a sequence of elements of $L$, the discriminant $\text{Disc}_{L:K}(l_1, \ldots, l_n)$ is defined to be $\det(M)$ where $M_{ij} = \text{Tr}_{L:K}(l_il_j)$. This use of the term is to be distinguished from the discriminant of a polynomial, as defined in section 7.3, although as will be seen there are relations between the two notions.

**Lemma 4.** Suppose $L \supseteq K$ is a finite extension of dimension $n$. If $l' = Tl$ for an $n \times n$ matrix $T$ over $K$ and column vectors $l, l'$ then $\text{Disc}(l'_1, \ldots, l'_n) = \det(T)^2 \text{Disc}(l_1, \ldots, l_n)$. In particular, $\text{Disc}(l_1, \ldots, l_n)$ does not depend on the ordering.

**Proof:** $\text{Disc}(l'_1, \ldots, l'_n) = \det(N)$ where

$$N_{ij} = \text{Tr}(l'_il'_j') = \text{Tr}(\sum_{rs} T_{ir}T_{js}l_rl_s) = \sum_{rs} T_{ir}T_{js} \text{Tr}(l_rl_s).$$

Thus, $\det(N) = \det(TMT')$, and the lemma follows.

**Lemma 5.** Suppose $L \supseteq K$ is a finite separable extension of dimension $n$, and $l_1, \ldots, l_n \in L$.

a. $l_1, \ldots, l_n$ is a basis for $L$ over $K$ iff $\text{Disc}(l_1, \ldots, l_n) \neq 0$.

b. If $\sigma_1, \ldots, \sigma_n$ are the embeddings of $L$ over $K$ into a normal closure of $K$ then $\text{Disc}(l_1, \ldots, l_n) = \det(N)^2$ where $N_{ij} = \sigma_i(l_j)$.

c. If $L = K(\lambda)$ for $\lambda \in L$ then $\text{Disc}(1, \lambda, \ldots, \lambda^{n-1})$ equals the discriminant of the monic irreducible polynomial for $\lambda$. 379
Proof: If $\sum_i k_i l_i = 0$ where the $k_i$ are elements of $K$, not all 0, then $\sum_i k_i \text{Tr}(l_i l_j) = 0$. That is, there is a linear dependence in the columns of $M$, so $\det(M) = 0$ (this direction holds without separability). Conversely suppose $l_1, \ldots, l_n$ is a basis and $\sum_i k_i \text{Tr}(l_i l_j) = 0$. Let $\lambda = \sum_i k_i l_i$; then $\text{Tr}(\lambda l_j) = 0$ for all $j$, so $\text{Tr}(\lambda l) = 0$ for all $l \in L$. By theorem 9.8, $\lambda = 0$; and so $k_i = 0$ for all $i$. This proves part a. By lemma 1 $\text{Tr}(l_i l_j) = \sum_i \sigma_i(l_i l_j)$. It follows that $M = N^N$, and part b follows. With $\sigma_1, \ldots, \sigma_n$ and $N$ as in part b, $\sigma_i(x^j) = \lambda_i^j$ where $\lambda_1, \ldots, \lambda_n$ are the conjugates of $\lambda$. By theorem 7.11, det($N$) equals $\prod_{i<j} (\lambda_i - \lambda_j)$, and the square of this is the discriminant of the monic irreducible polynomial. This proves part c.

The discriminant can be used to give an alternative proof of the existence of an integral basis. Suppose the hypotheses of lemma 2 hold, and $A$ is a principal ideal domain. Let $\{b_1, \ldots, b_n\} \subseteq B$ be a basis for $L$, and let $d = \text{Disc}(b_1, \ldots, b_n)$. Given $\beta \in B$, let $\beta = \sum_i k_i b_i$ where $k_i \in K$. Then $\text{Tr}(\beta b_j) = M k$ where $M_{ij} = \text{Tr}(b_i b_j)$ and $k$ is the column vector of the $k_i$. Since $M$ has entries in $A$ and $M^{-1} = M^{adj} / \det(M)$, it follows that $O_L \subseteq C$ where $C$ is the span over $A$ of $\{b_1/d, \ldots, b_n/d\}$. It follows that $O_L$ is a free $A$-module by theorem 8.6.

In the case of $A = \mathbb{Q}$, this reduces the problem of finding an integral basis from a basis to examining finitely many cases. Algorithms for computing both the discriminant and an integral basis can be found in [Cohen].

Suppose the hypotheses of lemma 2 hold, $l_1, \ldots, l_n$ and $l'_1, \ldots, l'_n$ are two bases for $L$ over $K$, and they generate the same $A$-module $M$. Then $l'_i$ is in $M$, so it is an $A$-linear combination of the $l_i$. Thus, the matrix $T$ where $T' = T$ has entries in $A$. Further, $T$ is invertible, and its inverse has entries in $A$. By lemma 4 $\text{Disc}(l'_1, \ldots, l'_n) = \alpha \text{Disc}(l_1, \ldots, l_n)$ where $\alpha$ is the square of a unit of $A$.

In particular, the discriminant of an integral basis for an a.n.f. $K$ is independent of the basis, and is called the discriminant of $K$. The bases for $K$ consisting of algebraic integers correspond to the nonsingular integer matrices. Such a basis has a discriminant dividing the discriminant of $K$, and the basis is an integral basis iff its discriminant equals the discriminant of $K$.

3. Extensions of real valuations. An important topic in algebraic number theory is how prime ideals in an a.i.r. factor in an extension a.i.r. This can be treated by considering how extensions of real valuations behave for finite dimensional field extensions, and then how primes in a Dedekind ring correspond to discrete valuations. We will give this treatment, following [Jacobson] and [Lang]. For more direct treatments see [Ash] or [Marcus].

Throughout this section, $F$ will denote a field, $v$ a valuation on $F$, $O$ a valuation ring in $F$, $M$ the maximal ideal, $U$ the group of units, and $G$ the value group (excluding 0, which is isomorphic to $F^{\#}/U$). $\hat{F}$ will denote an extension field of $F$, equipped with $\hat{v}$, $\hat{O}$, $\hat{M}$, $\hat{U}$, and $\hat{G}$. The valuation $\hat{v}$ extends $v$ if $\hat{v} \supseteq v$ as sets of ordered pairs. $\hat{M} = M \cap O$ since $M$ is maximal; and $U = U \cap O$. This gives a canonical embedding of $F^{\#}/U$ in $\hat{F}^{\#}/\hat{U}$, which is clearly order-preserving. Letting $\hat{v}$ be the valuation obtained from $\hat{O}$ as in theorem 20.15, $\hat{v}$ is an extension of $v$ via this embedding. There is also a canonical embedding of $O/M$ in $\hat{O}/\hat{M}$.

Lemma 6. Let $y_1, \ldots, y_r \in \hat{F}^{\#}$ be such that the $\hat{v}(y_i)$ are in distinct cosets of $G$ in $\hat{G}$. Let $z_1, \ldots, z_s \in \hat{O}$ be such that the $z_i + \hat{M}$ are linearly independent over $O/M$. Then the $y_iz_j$ are linearly independent.

Proof: Suppose $\sum_{ij} a_{ij} y_i z_j = 0$ for $a_{ij} \in F$ not all 0. Let $t_i = \sum_{ij} a_{ij} z_j$. Discarding those where all $a_{ij} = 0$, we may divide the coefficients by the $a_{ij}$ with the largest value. Then $v(a_{ij}) \leq 1$ and $v(a_{ij}) = 1$ for some $j$. By the assumption on the $z_j$, $\hat{v}(t_i)$ must equal 1. By lemma 20.13 $\hat{v}(t_i y_j) = \hat{v}(t_j y_i)$ for some $i, j$, whence $\hat{v}(y_k) = \hat{v}(y_j)$, a contradiction.

If $\hat{F} \supseteq F$ is a finite extension, of degree $n$, then both the index $e$ of $G$ in $\hat{G}$, and the dimension $f$ of $\hat{O}/\hat{M}$ over $O/M$, are finite, and $ef \leq n$. The value $e$ is called the ramification index of the extension, and the $f$ the residue class degree.
**Theorem 7.** Given \( v \), there is a valuation \( \tilde{v} \supseteq v \) on \( \tilde{F} \). If the extension \( \tilde{F} \supseteq F \) is algebraic (finite) and \( v \) is a real (discrete) valuation then there is a real (discrete) \( \tilde{v} \).

**Proof:** Let \( O \) be the valuation ring of \( v \), \( M \) the maximal ideal, \( \phi : O \to O/M \) the canonical homomorphism, and \( K \) the algebraic closure of \( O/M \). We thus have a pair \( \langle O, \phi \rangle \) where \( O \) is a subring of \( \tilde{F} \) and \( \phi : O \to K \) is a ring homomorphism. By a standard application of Zorn’s lemma (cf. theorem 15.4) there is a maximal such pair \( \langle \tilde{O}, \tilde{\phi} \rangle \) with \( \tilde{\phi} \) an extension of \( \phi \). We claim that given any such pair \( \langle A, \psi \rangle \), and \( x \in \tilde{F} \), there is an extension of \( \psi \) to either \( A[x] \) or \( A[1/x] \). It follows that for a maximal pair, \( \tilde{O} \) is a valuation ring. Let \( P = \ker \psi \); then \( \psi \) extends to \( A_P \) in the usual way, namely, \( \psi(r/s) = \psi(r)/\psi(s) \).

Suppose the ideal generated by \( P_P \) in \( A_P[1/x] \) is all of \( A_P[1/x] \), say \( 1 = a_0 + \cdots + a_n x^{-n} \) where \( a_i \in P_P \). Then since \( P_P \) is maximal and \( A_P \) is local, \( 1 - a_0 \) is a unit in \( A_P \). It follows that \( x \) is integral over \( A_P \). By theorem 20.23 there is a maximal ideal \( Q \) of \( A_P[x] \) lying over \( P_P \). Since \( A_P[x]/Q \) is an algebraic extension of \( A_P/P_P \), the canonical epimorphism \( A_P[x] \to A_P[x]/Q \) may be considered as having codomain \( K \), and the restriction to \( A[x] \) yields the desired extension. If the ideal generated by \( P_P \) in \( A_P[1/x] \) is not all of \( A_P[1/x] \) then there is a maximal ideal \( Q \) of \( A_P[1/x] \) containing it. Since \( P_P \subseteq A_P \cap Q \) and \( P_P \) is maximal in \( A_P \), \( P_P = A_P \cap Q \), so as noted in section 20.7, \( A_P/P_P \) maps injectively into \( A_P[1/x]/Q \). Pushing \( A_P/P_P \to K \) forward, and mapping \( 1/x \) arbitrarily, yields a homomorphism from \( A_P[1/x]/Q \) to \( K \), and thus a homomorphism from \( A[1/x] \) to \( K \). This completes the proof that \( \tilde{O} \) is a valuation ring. If \( \tilde{F} \supseteq F \) is finite, of degree \( n \), let \( e \) be the ramification index. The map \( \gamma \mapsto \gamma^e \) from \( \tilde{G} \) to \( G \) is injective. It follows that if \( v \) is discrete then \( \tilde{G} \) is.

If \( \tilde{F} \supseteq F \) is algebraic and \( v \) is real, for any \( x \in \tilde{F} \) there is an \( e \in \mathbb{N}^\circ \) and \( r \in \mathbb{R} \) with \( \tilde{v}(x)^e = r \); and we may let \( \tilde{v}(x) = r^{1/e} \).

**Theorem 8.** Suppose \( F \) is complete with respect to a real valuation \( v \), and \( \tilde{F} \supseteq F \) is an algebraic extension. Then there is only one \( \tilde{v} \). Indeed, \( \tilde{v}(a) = v(N_{\tilde{F}(a)/F}(a))^{1/n} \) where \( n \) is the degree of \( a \). If \( \tilde{F} \supseteq F \) is finite then \( \tilde{F} \) is complete.

**Proof:** There is some \( \tilde{v} \) by theorem 7. For the remaining claims we may assume \( \tilde{F} \supseteq F \) is finite. By theorem 24.14 any two extensions to \( \tilde{F} \) are equivalent as norms, Applying lemma 20.10, and noting that \( \lambda \) must be 1 since the valuations agree on \( F \), the valuations are identical. That \( \tilde{F} \) is complete follows by remarks in section 17.8. Suppose \( a \) is algebraic over \( F \), \( \tilde{F} \) is the normal closure of \( F[a] \), and \( \tilde{v} \) is the extension of \( v \). For any automorphism \( \sigma \) of \( \tilde{F} \) over \( F \), \( x \mapsto \tilde{v}(\sigma(x)) \) is a real valuation on \( \tilde{F} \). Thus, \( \tilde{v}(b) = \tilde{v}(a) \) for all conjugates \( b \) of \( a \), and so \( v(N(a)) = \tilde{v}(a)^n \).

Theorem 8 is true in the Archimedean case as well. Existence follows by Ostrowski’s theorem, and the proof of uniqueness is unchanged. Theorems 9 and 10 below are also true in the Archimedean case, with proofs unchanged.

As observed in section 20.12, if \( F \) is a field equipped with an absolute value \( v \) then there is a unique complete metric space \( F_v \) containing \( F \) as a dense subspace, with a unique field structure defined on it making \( F_v \) a field extension. Further, as is readily verified, the uniquely determined norm on \( F_v \) is the unique absolute value on \( F_v \) extending \( v \). By theorem 8 (and the corresponding fact for in the Archimedean case), this in turn has a unique extension to the algebraic closure of \( F_v \).

**Theorem 9.** Suppose \( K \) is the algebraic closure of \( F_v \).

a. An embedding \( \sigma \tilde{F} \to K \) gives rise to a real valuation \( \tilde{v} \supseteq v \) on \( \tilde{F} \).

b. Every real valuation on \( \tilde{F} \) with \( \tilde{v} \supseteq v \) arises this way.

c. Two embeddings \( \sigma_1, \sigma_2 \) give rise to the same real valuation if and only if there is an automorphism \( \lambda \) of \( K \) over \( F_v \) such that \( \sigma_2 = \lambda \sigma_1 \).

**Proof:** For part a, let \( w \) be the unique real valuation on \( K \) extending \( v \). Restricting \( w \) to an embedding of \( \tilde{F} \) yields \( \tilde{v} \). Given \( \tilde{v} \), ignoring canonical embeddings, \( F \subseteq \tilde{F} \subseteq \tilde{F}_0 \). Let \( F_\tilde{v} \) be the closure of \( F \) in \( \tilde{F}_0 \).
$F_{\bar{v}}$ is the completion of $F$, and the compositum $\bar{F} \sqcup F_{\bar{v}}$ is a finite extension of $F_{\bar{v}}$ (corollary 20.17), so is complete (theorem 7). Since it contains $\bar{F}$ and is contained in $F_{\bar{v}}$, it equals $F_{\bar{v}}$. There is an isomorphism $F_{\bar{v}} \to F_v$, which extends to an embedding of $F_{\bar{v}}$ in $K$, determined by the embedding of $\bar{F}$ given by the restriction. This proves part b. If $\sigma_2 = \sigma_1\lambda$ then by the uniqueness of the extension of $v$, $\bar{v}\lambda = \bar{v}$, whence $\bar{v}\sigma_2 = \bar{v}\sigma_1 = \bar{v}\sigma_1$. If $\bar{v}\sigma_2 = \bar{v}\sigma_1$ let $\lambda_0 : \sigma_1[\bar{F}] \to \sigma_2[\bar{F}]$ be an isomorphism; then by the assumption $\lambda_0$ is an isometry. By arguments above, $\sigma_i[\bar{F}]$ is dense in $\sigma_i[\bar{F}] \sqcup F_v$. Thus $\lambda_0$ extends to an isomorphism $\lambda_1 : \sigma_1[\bar{F}] \sqcup F_v \to \sigma_2[\bar{F}] \sqcup F_v$. This extends to the required $\lambda$.

As seen from the proof, $F_{\bar{v}}$ may be considered as an extension of $F_v$.

**Theorem 10.** Suppose $\bar{F} \supseteq F$ is finite and separable, let $n$ be the degree of $\bar{F}$ over $F$, and let $n_{\bar{v}}$ be the degree of $F_{\bar{v}}$ over $F_v$. Then $n = \sum_{\bar{v} \supseteq v} n_{\bar{v}}$.

**Proof:** By lemma 9.2 $\bar{F} = F[c]$ for an element $c \in \bar{F}$. Let $p$ be the irreducible polynomial for $c$. In $F_v$, $p$ factors as $p_1 \cdots p_e$ for distinct $p_i$. An embedding of $\bar{F}$ into the algebraic closure of $F_v$ is determined by the root of $p$ that $c$ maps to. There is an automorphism mapping one embedding to another iff the image of $c$ belongs to the same $p_i$. For a given $i$, with embedding $\sigma$, the degree of $\bar{F}_{\bar{v}} = \sigma[\bar{F}] \sqcup F_v$ over $F_v$ is the degree of $p_i$. Indeed, $\bar{F}_{\bar{v}} = F_v[\sigma(c)]$, and $p_i$ is the irreducible polynomial.

**Theorem 11.** Suppose $F$ is complete, $\bar{F} \supseteq F$ is finite, of degree $n$, and $v$ is discrete. Then $ef = n$ where $e$ is the ramification index and $f$ the residue class degree.

**Proof:** By results already shown, $\bar{F}$ is complete and $\bar{v}$ is a discrete valuation. Let $\pi$ be a prime in $O$ (as observed in the proof of theorem 20.42 there is up to units a unique such, and $v(\pi)$ generates the value group). Let $\pi$ be a prime in $\bar{O}$. Then $\Pi^e = u\pi$ where $u$ is a unit of $U$. By remarks in section 20.12, every element of $\bar{O}$ can be written as $\sum_i b_i\Pi^i$, where $b_i$ is from a system of representatives of the cosets of $M$ in $\bar{O}$. If $z_1, \ldots, z_f$ are such that the $z_i + M$ are a basis of $\bar{O}/M$ over $O/M$, and $R$ is a system of representatives of the cosets of $M$ in $O$, then the linear combinations $a_1z_1 + \cdots + a_fz_f$ with $a_i \in R$, form a system of representatives for the cosets of $M$. Indeed, given a coset $w + M$ there are unique cosets $w_i + M$ giving $w + M$ in terms of the $z_i$, and there are unique $a_i$ representing them. Thus, every element of $\bar{O}$ is of the form $\sum_{i=1}^f \sum_{j=1}^{c-1} \sum_k a_{ijk} \pi^{i-k}$, and thus $\{z_i\Pi^i\}$ generates $\bar{F}$ over $F$. By lemma 6 it is linearly independent.

By remarks in section 20.12, the residue class field and the value group do not change when the completion of a field with a valuation $v$ is taken. It follows that for a finite extension $\bar{F} \supseteq F$ with real valuations $\bar{v} \supseteq v$, the residue class degree and the ramification index of $\bar{v} \supseteq v$ do not change when the completions of $F$ and $\bar{F}$ are taken. The degree of the extension does not change either (exercise 2).

**Theorem 12.** Suppose $\bar{F} \supseteq F$ is finite and separable, let $n$ be the degree of $\bar{F}$ over $F$, suppose $v$ is a discrete valuation, and let $e_{\bar{v}}$ and $f_{\bar{v}}$ be the ramification index and residue class degree of the extension $\bar{v} \supseteq v$. Then $n = \sum_{\bar{v} \supseteq v} e_{\bar{v}} f_{\bar{v}}$.

**Proof:** This follows by theorems 10 and 11, and remarks in the preceding paragraph.

It may also be seen that for an arbitrary real valuation $v$, and without the separability assumption, $n \leq \sum_{\bar{v} \supseteq v} e_{\bar{v}} f_{\bar{v}}$.

**4. Valuations on Dedekind Domains.** A discrete valuation ring $R$ was defined in section 20.12 to be a local principal ideal domain. In particular it is a local Dedekind domain (see the remarks preceding theorem 20.41). Every element can be written uniquely as $u\pi^n$ where $\pi$ generates the unique maximal principal ideal. This gives rise to a discrete valuation on the field of fractions, which is the essentially unique discrete valuation with $R$ as its valuation ring.

**Lemma 13.** A local Dedekind domain is a discrete valuation ring.
PROOF: Let $R$ denote the ring. Since every nontrivial prime ideal is maximal, and there is only one maximal ideal, there is a unique nontrivial prime ideal $P$. By unique factorization of ideals, every ideal is a power of $P$ (i.e., $[P^k]$ for some $k$), and the powers are distinct. Choose $\pi \in P - [P^2]$. Then $\pi R = [P^k]$ for some $k$, and $k$ must be 1 since $\pi \notin [P^2]$. Thus, $P$ is principal, and hence every ideal, being a power of $P$, is principal.

LEMMA 14.

a. If $R$ is a Noetherian commutative ring and $S$ is a multiplicative subset than $RS$ is Noetherian.

b. If $R$ is an integrally closed integral domain and $S$ is a multiplicative subset than $R_S$ is integrally closed.

c. If $R$ is a Dedekind ring and $P$ is a prime ideal then $R_P$ is a local Dedekind ring.

PROOF: For part a, by remarks in section 20.8, if $J \subseteq R_S$ is a proper ideal then $J = IS$ for some ideal $I \subseteq R$. A generating set for $I$ is a generating set for $J$. For part b, let $K$ be the field of fractions of $R$; $K$ is also the field of fractions of $R_S$. If $r/s \in K$ satisfies a monic polynomial of degree $d$, with coefficients $a_i/t_i$ $(a_d = t_d = 1)$ in $R_S$, let $t$ be the product of the $t_i$. Then $tr/s$ satisfies the monic polynomial with coefficients $a_it_i^{d-1}$, hence is integral over $R$, hence is in $R$, and hence $r/s \in R_S$. For part c, it suffices to observe that for $P$ nonzero, by the correspondence of prime ideals under localization (remains in section 20.7), $P_P$ is the unique maximal ideal of $R_P$.

LEMMA 15. Suppose $R$ is a Dedekind domain, $F$ is the field of fractions, $O$ is a valuation ring in $F$ containing $R$, $M$ is the maximal ideal of $O$, and $P = M \cap R$. Then $O = R_P$.

PROOF: Let $U = O - M$. If $s \in R - P$ then $s \in O - M = U$; this shows that $R_P \subseteq O$. If $P = \{0\}$ then $F = O$ follows. Otherwise, $P_P \subseteq [MO] = M$. That is, we have valuation rings $O_2 \subseteq O_1$, with maximal ideals $M_2 \subseteq M_1$. Then $O_1 = F^\# - M_1^{-1} \subseteq F^\# - M_2^{-1} = O_2$ also, and $O_2 = O_1$; that is, $R_P = O$.

Suppose $P$ is a prime ideal in a Dedekind domain $R$. If $r \in R$ then the principal ideal $Rr$ has a unique factorization $[P_1^{e_1} \cdots P_g^{e_g}]$; choosing $\gamma < 1$, we let $v_P(r) = \gamma^e$ where $e$ is the power of $P$ in the factorization. Recall from section 20.11 that for ideals $I$ and $J$, $I$ divides $J$ iff $I \supseteq J$.

LEMMA 16. With notation as in the preceding paragraph, $v_P$ is the discrete valuation with valuation ring $R_P$. Every non-Archimedean valuation on $R$ is, up to equivalence, $v_P$ for some $P$.

PROOF: That $v_P$ is a non-Archimedean valuation is shown as in the case of the rationals (section 20.4), and again left to the reader. Since $r \in R_P$ implies $v_P(r) \leq 1$, and $r \in P_P$ implies $v_P(r) < 1$, by facts noted in the proof of lemma 15 $R_P$ is the valuation ring. Suppose $v$ has $R_P$ as its valuation ring. Let $\pi$ be a prime element of $R_P$; then any element $q \in R_P$ can be written uniquely as $u\pi^k$ where $u$ is a unit of $R_P$ and $k \geq 0$, and $v(q) = \gamma^k$ where $v(\pi) = \gamma$ and $0 < \gamma < 1$. Suppose $a \in [P^k] - [P^{k+1}]$. Then $a \in [P^k R_P] - [P^{k+1} R_P]$, else $a \in [P]^{k+1}$ for some $s \notin P$, a contradiction. Thus, $v(a) = \gamma^k$. Also, since $P^k \supseteq Da$ and $P^{k+1} \supseteq Da$, $P$ appears to the $k$th power in the factorization of $Ra$ into prime ideals. The lemma follows.

Note that the valuations $v_P$ for distinct nonzero $P$ are inequivalent, since the $R_P$ are distinct.

THEOREM 17. Suppose $R$ is a Dedekind domain, $F$ is the field of fractions, $F \supseteq F$ is a finite extension of dimension $n$, $P \subseteq R$ is a nonzero prime ideal, and $[\tilde{P}_1^{e_1} \cdots \tilde{P}_g^{e_g}]$ is the factorization of $[P\tilde{R}]$ into prime ideals in $\tilde{R}$.

a. $\tilde{P}$ occurs among the $\tilde{P}_i$ iff it lies above $P$.

b. The $\tilde{P}_i$ are different for different $P$.

c. The extensions of $v_P$ are (up to equivalence) the $v_{\tilde{P}_i}$.

d. The ramification index of $v_{\tilde{P}_i}$ over $v_P$ equals $e_i$.

e. The residue class degree equals the dimension of $\tilde{R}/\tilde{P}_i$ over $R/P$.

f. $n = \sum_i e_i f_i$.
Proof: For part a, if $\bar{P} \supseteq P$ then $\bar{P} \supseteq [P\bar{R}]$; and if $\bar{P} \supseteq [P\bar{R}]$ then $\bar{P} \cap R = P$ since $P$ is maximal. Part b follows immediately from part a. For $r \in R$, $v_P(r) = \gamma o$ where $o$ is the order of $Rr$ at $P$, and $v_{\bar{P}}(r) = \bar{\gamma} o$ where $\bar{o}$ is the order of $\bar{R}r$ at $\bar{P}$. Clearly $\bar{o} = e_o o$. Replacing $\gamma$ by $\bar{\gamma} o$, $v_{\bar{P}}$ is an extension of $v_P$. Also, part d follows. If $v_{\bar{P}}$ extends $v_P$ then by lemma 15 $P = R \cap M$ where $M$ is the maximal ideal of the valuation ring $\bar{R}r$; likewise $\bar{P} = \bar{R} \cap \bar{M}$ since likewise $\bar{R}\bar{r}$ is isomorphic to $\bar{R}_{\bar{P}} / \bar{P}$. The map is an embedding because $P$ is prime (if $r \in R \cap \bar{P}$ then $r \in P$). It is surjective because $P$ is maximal. Indeed, given $r/s$ with $r \in R$ and $s \in R - P$, since $R/P$ is a field there is a $t \in R$ such that $st - 1 \in P$. It follows that $rt - r/s \in P$. Part f is immediate from theorem 12.

5. Decomposition and inertia groups.

In this section, $A$ denotes an integrally closed integral domain, $K$ its field of fractions, $L$ a finite Galois extension of $K$, $G$ the Galois group, and $B$ the integral closure of $A$ in $L$.

Lemma 18. $G[B] = B$, and $G \mid B$ is a ring automorphism fixing $A$. In particular $G$ acts on the ideals, the prime ideals, and the maximal ideals. Suppose $P \subseteq A$ is a prime ideal. Then $G$ acts on the prime ideals of $B$ lying over $A$; further it acts transitively on them.

Proof: If $b \in B$ and $\sigma \in G$ then $b$ satisfies some monic polynomial $p$ with coefficients in $A$; $\sigma(b)$ also satisfies $p$, so $\sigma(b) \in B$. Thus, $\sigma[B] \subseteq B$, and $\sigma[B] = B$ follows by properties of group action. All remaining claims other than transitivity are clear. Localizing at $P$, $P_P$ is maximal, and if $Q \subseteq B_P$ lies above $Q_P$ then $Q_P$ is maximal by lemma 20.22. If $Q_1, Q_2 \subseteq B_P$ are such that $\sigma[Q_1] \neq Q_2$ for any $\sigma \in G$ then the $\sigma[Q_1]$ for $\sigma \in G$ and the $\sigma[Q_2]$ for $\sigma \in G$ are all distinct, hence comaximal. As in the proof of theorem 6.8, there is a $t \in B_P$ such that $t \in \sigma[Q_1]$ for $\sigma \in G$, and $t - 1 \in \sigma[Q_2]$ for $\sigma \in G$. $N_{L,K}(t) = \prod_{\sigma \in G} \sigma(t)$ is in $B_P \cap K$, and since $B_P$ is integrally closed (lemma 14), $B_P \cap K = A_P$, and $Q_1 \cap A_P = P_P$, so $N_{L,K}(t) \in B_P$. But $\sigma(t) \notin Q_2$ for any $\sigma$, whence $N_{L,K}(t) \notin Q_2$, whence $N_{L,K}(t) \notin P_P$, a contradiction.

Suppose $Q \subseteq B$ is a prime ideal lying over $P \subseteq A$. The stabilizer of $Q$ is called the decomposition group. We use $G^D$ to denote this subgroup of $G$ (for given $P$ and $Q$). The fixed field will be denoted $L^D$, and is called the decomposition field.

If $\sigma \in G^D$ then there is an induced homomorphism $\bar{\sigma} : B/Q \to B/Q$, which in fact is readily verified to be an automorphism of $B/Q$ over $A/P$. The map $\sigma \to \bar{\sigma}$ is a homomorphism from $G^D$ to the group of such automorphisms. The kernel of this map, a normal subgroup of $G^D$, is called the inertia group, and denoted $G^I$. The fixed field will be denoted $L^I$, and is called the inertia field.

If $L \supseteq E \supseteq K$ then as noted in section 9.4, $L$ is Galois over $E$. The integral closure of $A$ in $E$ clearly equals $B \cap E$. Also, $Q \cap E$ is a prime ideal of $B \cap E$, with $Q$ lying over it. In the case $E = L^D$, $\text{Aut}_{L^D}(L) = G^D$, so by lemma 18 $Q$ is the only prime of $L$ lying over $Q \cap L^D$. On the other hand suppose $Q$ is the only prime lying over $Q \cap E$, and let $H$ be the Galois group of $L$ over $E$. Then by lemma 18 $H$ leaves $Q$ invariant, so $H \subseteq G^D$, so $L^D \subseteq E$. Thus, $L^D$ is the smallest $E$ such that $Q$ is the only prime lying over $Q \cap E$.

Lemma 19. The canonical embedding $A/P \to (B \cap L^D)/(Q \cap L^D)$ is an isomorphism.

Proof: Let $B^D$ denote $B \cap L^D$ and let $Q^D$ denote $Q \cap L^D$. Suppose $x \in B^D$. For $\sigma \in G - G^D$, $\sigma^{-1}[Q] \neq Q$, whence $\sigma^{-1}[Q] \cap B^D \neq Q^D$. As in the proof of lemma 18 let $y \in B^D$ be such that $y - x \in Q^D$ and $y - 1 \in \sigma^{-1}[Q] \cap B^D$ for $\sigma \in G - G^D$. Then $y \in x + Q$, and $\sigma(y) \in 1 + Q$ for $\sigma \in G - G^D$. It follows that $N_{L^D,K}(y) - x$ is in $Q$. Now, $N_{L^D,K}(y) \in A$ since $y$ is integral over $K$ and $A$ is integrally closed in $K$. Also $N_{L^D,K}(y) - x$ is in $Q^D$ since $N_{L^D,K}(y)$ and $x$ both are. The lemma follows.
As observed in section 20.6, the separable extensions form a distinguished class. Hence if $F_2 \supseteq F_0$ is a finite (indeed algebraic) extension, there is a largest $F_1$ with $F_2 \supseteq F_1 \supseteq F_0$, such that $F_1 \supseteq F_0$ is separable; indeed, it consists of the elements of $F_2$ which are separable over $F_0$. By exercise 9.3, for every element $x \in F_2$ there is a $t$ such that $x^p^t \in F_1$, where $p$ is the characteristic. Given any embedding of $F_2$ over $F_1$ into a normal closure of $F_2$ each element $x \in F_2 - F_1$ must map to a root of its irreducible polynomial, which divides $x^p^t - a$ for some $a \in F_1$, and so $x$ can only map to the unique root of $x^p^t - a$. Thus, $[F_2 : F_1]_s = 1$.

**Lemma 20.** The map $\sigma \mapsto \bar{\sigma}$ defined above, from $G_D$ to the automorphisms of $B/Q$ over $A/P$, is surjective.

**Proof:** By lemma 19 we may replace $K$ by $L^D$, in which case $G^D$ is the entire automorphism group of $L$ over $K$. With $F_0 = A/P$ and $F_2 = B/Q$, let $F_1$ be the maximal separable subextension as in the remarks above. Let $x \in B$ be such that $x + Q$ generates $F_1$ over $A/P$. $G$ acts transitively on the roots of the irreducible polynomial of $x$ over $K$, so every automorphism $\alpha$ if $F_1$ over $A/P$ is the restriction of the image of some $\sigma$, and so the unique extension of $\alpha$ is the image of some $\sigma$.

Now let us assume that $A$ is a Dedekind domain; by theorem 3 $B$ is. By lemma 18, if $Q_1, \ldots, Q_g$ are the prime ideals of $B$ lying over $P$ then the ramification indexes $e_i$ have a common value $e$, and the residue class degrees $f_i$ have a common value $f$. By theorem 17, $n = efg$ in the case where $L \supseteq K$ is Galois.

Suppose $Q \subseteq B$ is a prime ideal lying over $P \subseteq A$, with ramification index $e$ and residue class degree $f$. Then the order of $G^D$ clearly equals $ef$. By lemma 20, the order of $G^f$ equals $e$. By remarks in section 9.4, and the multiplicativity of vector space dimension, we can conclude that $L^D$ is an extension of $K$ of degree $g$, $L^f$ is an extension of $L^D$ of degree $f$, and $L$ is an extension of $L^f$ of degree $e$.

Finally let us assume that $A$ is an a.i.r., whence $B$ is also. $P \cap \mathbb{Z}$ is a prime ideal of $\mathbb{Z}$, so is the principal ideal of some prime number $p \in \mathbb{Z}$. $\mathbb{Z}/p\mathbb{Z}$ is the finite field $\mathcal{F}_p$. It follows that $A/P$ and $B/Q$ are finite fields of characteristic $p$. By remarks in section 9.6, the automorphism group of $B/Q$ over $A/P$ is cyclic, and generated by the map $x \mapsto x^q$ where $q$ is the cardinality of $A/P$. Any $\sigma \in G$ such that $\bar{\sigma}$ equals this map is called a Frobenius automorphism. If $e = 1$ (the extension is “unramified”) then the Frobenius automorphism is unique.

6. Ideal norm. As we just saw, if $A$ is an a.i.r. and $P$ is a prime ideal, then $A/P$ is a finite field of order $p^f$ where $p$ is the integer prime contained in $P$ and $f$ is the residue class degree of $P$ over $p\mathbb{Z}$.

**Lemma 21.** Suppose $A$ is a Dedekind domain, $P$ is a prime ideal, and $e > 1$. Then the $A$-module $P^{e-1}/P^e$ is isomorphic to $A/P$.

**Proof:** Since $P^{e-1} \supseteq P^e$, there is an element $a \in P^{e-1} - P^e$. The map $x \mapsto ax + P^e$ from $A$ to $P^{e-1}/P^e$ is an $A$-module homomorphism. It is surjective, because $Aa + P^e$ contains $P^e$ and is contained in $P^{e-1}$, so equals $P^{e-1}$. $x$ is in the kernel iff $ax \in P^e$, and this holds iff $x \in P$ by unique factorization of ideals.

Returning to the case of an a.i.r., we claim that $A/P^e$ has order $p^{ef}$. This follows by induction on $e$, the basis $e = 1$ being immediate. $P^{e-1}/P^e$ is an $A$-submodule of $A/P^e$, and by theorem 4.11 $A/P^{e-1}$ is isomorphic as an $A$-module to $(A/P^e)/(P^{e-1}/P^e)$. $A/P^{e-1}$ has order $p^{(e-1)f}$ by induction, and $P^{e-1}/P^e$ has order $p^f$ by lemma 21, proving the claim.

If $A$ is a Dedekind domain, and $P_1, P_2$ are distinct prime ideals, then there can be no prime ideal $Q$ with $Q \supseteq P_1^e$ and $Q \supseteq P_2^e$. Thus, $P_1^e$ and $P_2^e$ are comaximal. Suppose $I$ is an ideal in $A$, and $[P_1^e \cdots P_g^e]$ is the factorization into prime ideals. By theorem 6.8, the map $x \mapsto (\eta_1(x), \ldots, \eta_g(x))$, where $\eta_i : A \to A/P_i^e$ is the canonical homomorphism, is a surjection with kernel $I$.

Thus, for an a.i.r. $A$, using obvious notation, the order of the ring $A/I$ equals $\prod P_i^{e \cdot f_i}$. In particular it is finite. This value is called the norm of the ideal $I$; we will use $\text{Inrm}(I)$ to denote it. That $\text{Inrm}(IJ) = \text{Inrm}(I)\text{Inrm}(J)$ follows by unique factorization of ideals.

385
Fixing an integral basis for an a.i.r. A establishes a bijective correspondence between $A$ and $\mathbb{Z}^n$ where $n$ is the degree of $A$ over $\mathbb{Z}$. An ideal $I$ corresponds to an integer lattice $L$; $L$ is full rank, because $Ax \subseteq I$ for any $x \in I$. Further, a system of coset representatives for $L$ corresponds to a system of coset representatives for $I$. By remarks in section 23.4, the number of these, i.e. $\text{Inrm}(I)$, equals $\det(L)$.

It was observed in section 23.4 that there are only finitely many integer lattices with a given determinant. Thus, there are only finitely many ideals in an a.i.r. with a given ideal norm.

$x$ for $I$ the lattice $L$ of $A$ is the degree of $A$, known as Minkowski’s bound, can be found in [Ash].

Suppose $z_1, \ldots, z_n$ is an integral basis for $A$. If $Aa$ is a principal ideal then $az_1, \ldots, az_n$ is a basis for the lattice $L$ of $Aa$. $\text{Disc}(az_1, \ldots, az_n)$ equals the determinant of the matrix with entries $\sigma_i(az_j)$ where the $\sigma_i$ are as in lemma 5. This in turn equals $N(a)^2 \text{Disc}(z_1, \ldots, z_n)$, where the norm is from $A/A^\#$ to $\mathbb{Q}$. On the other hand, by lemma 4, $\text{Disc}(az_1, \ldots, az_n)$ equals $\det(L)^2 \text{Disc}(z_1, \ldots, z_n)$. Thus, $\text{Inrm}(Aa) = |N(a)|$.

7. Ideal class group. If $R$ is an integral domain, define a relation $\equiv$ in the ideals of $R$, where $I \equiv J$ iff $[aI] = [bJ]$ for some $a, b \in R$. We leave it to exercise 1 to show that $\equiv$ is an equivalence relation, ideal multiplication respects $\equiv$, the principal ideals form an equivalence class, and the class of the principal ideals is a multiplicative identity. That is, the equivalence classes from a monoid under ideal multiplication.

Recall from section 20.11 that if $R$ is a Dedekind domain, the fractional ideals form a group $\text{FrIdl}(R)$ under multiplication of fractional ideals, which generalizes multiplication of ideals. The principal fractional ideals form a subgroup $\text{PFrIdl}(R)$ isomorphic to $R/R^\#$.

**Lemma 22.** If $R$ is a Dedekind domain then $\text{FrIdl}(R)/\text{PFrIdl}(R)$ is the monoid of ideal classes of $R$, which is therefore a group, called the ideal class group of $R$.

**Proof:** Let $K$ be the field of fractions. For ideals $I_1, I_2 \subseteq R$ let $I_1 \equiv_R I_2$ iff for some $r_1, r_2 \in R$, $[r_1 I_1] = [r_2 I_2]$. For fractional ideals $F_1, F_2 \subseteq K$ let $F_1 \equiv_K F_2$ iff for some $k \in K$, $[F_1 F_2'] = kR$ where $F_2'$ is the inverse ideal, iff for some $k \in K$, $F_1 = [kF_2]$. Then $I_1 \equiv_R I_2$ iff $I_1 \equiv_F I_2$, so the map taking an $\equiv_R$ equivalence class to the $\equiv_F$ equivalence class is a well defined, injective monoid homomorphism from the ideal classes to $\text{FrIdl}(R)/\text{PFrIdl}(R)$. It is surjective, because every class in the latter contains ordinary ideals.

For an alternative proof that the ideal classes in a Dedekind domain form a group, observe that for some multiple $J$ of $I$, the inverse fractional ideal of $I$, $[IJ]$ is a principal ideal,

**Theorem 23.** Given an a.i.r. $A$, there is a constant $\beta_A$, such that for every ideal $I$ of $A$ there is an $x \in I$ such that $|\text{Inrm}(x)| \leq \beta_A \text{Inrm}(I)$.

**Proof:** Let $z_1, \ldots, z_n$ be an integral basis for $A$. Given $I$, let $m$ be such that $m^n \leq \text{Inrm}(I) < m^{n+1}$. Then two of the elements $\sum_{i=1}^n l_i z_i$ where $l_i \in \mathbb{Z}$ and $0 \leq l_i \leq m$ must be equal, and so there is an element $x = \sum_{i=1}^n l_i z_i$ in $I$, where $|l_i| \leq m$. Then

$$|\text{Inrm}(x)| = \prod_{j=1}^n |\sigma_j(x)| \leq m^n \beta_A$$

where $\beta_A = \prod_{j=1}^n \sum_{i=1}^n |\sigma_j(z_i)|$.

Finally $|\text{Inrm}(x)| \leq \beta_A \text{Inrm}(I)$.

A better value for $\beta_A$, known as Minkowski’s bound, can be found in [Ash].

**Corollary 24.** Given an a.i.r. $A$, let $\beta_A$ be as in theorem 23. Then every ideal class of $A$ contains an ideal $I$ with $\text{Inrm}(I) \leq \beta_A$.

**Proof:** Let $C$ be the class, let $J$ be an ideal in the inverse class, and choose $x \in J$ such that $N(x) \leq \beta_A \text{Inrm}(J)$. Then $J \supseteq Ax$, so $[IJ] = Ax$ for some $I$, and $I \in C$. Also $|\text{Inrm}(x)| = \text{Inrm}([IJ]) = \text{Inrm}(I)\text{Inrm}(J)$, and the claim follows.

386
Theorem 25. The ideal class group of an a.i.r. $A$ is finite.

Proof: If $\beta_A$ is as above, then since there are only finitely many ideals $I$ with $\text{Inrm}(I) \leq \beta_A$, there can be only finitely many ideal classes.

8. Gauss’ lemma. Recall from section 7.1 Gauss’ lemma for factorial domains. If $R$ is a factorial domain, $F$ is its field of fractions, $p(x)$ is primitive, and $p(x) = q(x)q'(x)$ in $F[x]$, then $p(x) = q'(x)q''(x)$ in $R[x]$ where $q'(x), q''(x) \in R[x]$ are primitive. In this section an analogous fact for a.i.r.’s will be proved. For an interesting discussion see [MagMcK].

Lemma 26. Suppose $p$ is a polynomial whose coefficients are algebraic integers, and $K$ and $L$ are a.n.f.’s containing the coefficients. Then the ideal in $O_K$ generated by the coefficients equals $O_K$ iff the ideal in $O_L$ generated by the coefficients equals $O_L$.

Proof: By taking the compositum it suffices to consider the case $K \subseteq L$. Let $I (J)$ be the ideal generated in $O_K (O_L)$. $I (J)$ is proper iff there is some prime ideal $P \subseteq O_K (Q \subseteq O_L)$ with $P \supseteq I (Q \supseteq J)$. If $P$ exists let $Q$ be a prime factor of the ideal it generates in $O_L$, and if $Q$ exists let $P = O_K \cap Q$.

A polynomial $p$ with algebraic integer coefficients is said to be primitive if its coefficients generate $O_K$ in any a.n.f. $K$ containing them.

Lemma 27. The product of primitive polynomials with algebraic integer coefficients is primitive.

Proof: The proof is the same as the proof of theorem 7.4.a, with the prime element replaced by a prime ideal (recalling that a prime ideal “divides” an element $x$ iff $x \in P$).

Since the ideal class group of an a.i.r. $O_K$ is finite, for any ideal $I$ there is an integer $k$ such that $[I^k]$ is principal, say $[I^k] = aO_K$. Let $\xi$ be a $k$th root of $a$, and let $L = K[\xi]$. Then $[IO_L]^k = [I^kO_L] = [aO_L] = [\xi^kO_L] = [\xi O_L]$. By unique factorization into prime ideals, $[IO_L] = [\xi O_L]$. That is, for any ideal $I \subseteq O_K$ there is a finite extension $L \supseteq K$ such that $[IO_L]$ is a principal ideal.

Lemma 28. If $p$ is a polynomial whose coefficients are algebraic numbers then there is an a.n.f. $K$ containing the coefficients, such that $p = cp'$ where $c \in K$ and $p' \in O_K$ is primitive. In any such $O_K$, $c$ and $p'$ are unique up to units.

Proof: Suppose $p = \sum_{i=0}^n (a_i/b_i)x^i$, and let $F$ be an a.n.f. containing the coefficients. Let $e = b_1 \cdots b_n$; then $ep \in O_F[x]$. Let $I \subseteq O_F$ be the ideal generated by the coefficients of $ep$. Let $K \supseteq F$ be a finite extension where $[IO_K]$ is a principal ideal, say $fO_K$ where $f \in O_K$; clearly $fO_K$ is the ideal generated in $O_K$ by the coefficients of $ep$. It follows that $p' = ep/f$ is primitive, and $p = cp'$ where $c = f/e$. For uniqueness, if $c_1p_1' = c_2p_2'$ then $c_1O_K = c_2O_K$; that $c_1/c_2$ is a unit follows (exercise 3).

Theorem 29. Suppose $K$ is an a.n.f., $p \in O_K[x]$ is primitive, $p$ is primitive, and $p = qr$ in $K[x]$. Then there are a finite extension $L$ of $K$, primitive polynomials $q', r' \in O_L[x]$, and $c, d \in L$, such that $p = qr'$, $q = cq''$, and $r = dr'$. Then $p = cdq''r'$, and by uniqueness $cd$ is a unit, so we may let $p' = cdq''$.

Exercises.

1. Verify the properties of the relation $\equiv$ stated at the beginning of section 3. Hint: $[I] = [I]$, if $[aI] = [bJ]$ then $[bJ] = [aI]$, and if $[aI] = [bJ]$ and $[bJ] = [cK]$ then $[aI] = [cK]$. Thus, $\equiv$ is an equivalence relation. Since $[[aI][a'I']] = [aa'I']$, if $[aI] = [bJ]$ and $[a'I'] = [b'J']$ then $[aa'I'] = [bb'J']$; thus, ideal multiplication respects $\equiv$. Since $ar = ar, aR \equiv R$. Trivially, $I \equiv RI$. 387
2. Suppose $E$ is an $n$-dimensional vector space over a field $F$, with basis $b_1, \ldots, b_n$, equipped with the max norm. Show that (ignoring canonical embeddings) $b_1, \ldots, b_n$ is a basis for $E^{\text{compl}}$ over $F^{\text{compl}}$. Hint: $(F^n)^{\text{compl}}$ is a completion, and the isomorphism is a vector space isomorphism.

3. Suppose $R$ is an integral domain and $F$ its field of fractions. Show that, for $c_1, c_2 \in R$ where $c_1, c_2 \in F$ then $c_1/c_2$ is a unit in $R$. Hint: $c_1 = c_2a$ and $c_2 = c_1b$ where $a, b \in R$. 


27. Lie groups.

1. Definition. In section 24.7 we gave a definition of an analytic function from \( F^n \) to \( X \), where \( F \) is a field complete with respect to an absolute value \( |x| \), and \( X \) is a complete normed linear space over \( F \). In this chapter \( X \) will be \( F^m \) for some \( m \). A function \( f \) is analytic at a point \( x \) if there is a power series such that \( f(x + \xi) = \sum a_\mu \xi^\mu \) in some polydisc around \( x \). One verifies that a function \( f : F^n \rightarrow F^m \) is analytic at \( x \) iff each \( f_i \) is analytic, where \( f = (f_1, \ldots, f_m) \) and \( f_i : F^n \rightarrow F \). We leave it to the exercises to verify that the analytic functions are closed under composition.

\( C^n \) is a \( 2n \)-dimensional vector space over \( \mathbb{R} \). If \( f : C^n \rightarrow C \) is a complex analytic function its real and imaginary parts are real analytic functions on \( \mathbb{R}^{2n} \) (exercise 5). The function \( 1/(x^2 + 1) \) is readily seen to be real analytic on all of \( \mathbb{R} \). However, a function \( f : C \rightarrow C \) which is bounded in the entire complex plane is a constant (Liouville’s theorem; see [Rudin2]).

In section 24.11 manifolds were defined, where the “model space” was an arbitrary topological vector space. In particular, the model space might be \( F^n \) for some \( n \). In this case, the transition maps may be required to be analytic. Lie groups are considered in this generality (see [Serre2] for example), but for the remainder of the chapter we will consider only the cases \( F = \mathbb{R} \) or \( F = \mathbb{C} \).

The category of real (complex) analytic manifolds has as morphisms those continuous functions which, when expressed in terms of coordinates as described in section 24.11, are analytic. There is a forgetful functor from the complex analytic manifolds to the real analytic manifolds; an \( n \)-dimensional complex manifold becomes a \( 2n \)-dimensional real manifold. As seen above, the complex case must be distinguished.

A real (complex) Lie group is defined to be a group in the category of real (complex) analytic manifolds. Thus, it is an analytic manifold \( M \), with analytic morphisms \( m : M \times M \rightarrow M \) and \( i : M \rightarrow M \), satisfying the group axioms. The requirement that \( i \) be analytic is redundant; see [Serre2]. Relaxing the smoothness requirement is not really more general; see [Vara] for some discussion. By the forgetful functor, a complex Lie group “has an underlying” real Lie group.

The real \( n \times n \) matrices are an \( n^2 \)-dimensional real analytic manifold. The invertible \( n \times n \) real matrices are an open subset, and as noted in section 24.11 comprise an analytic manifold with “inherited” charts. Matrix multiplication is clearly analytic. The function \( x \mapsto x^{-1} \) is also; by exercise 6 \( 1/\det(x) \) is, and the claim follows by exercise 4 and the fact that \( x \mapsto x^{adj} \) is analytic. Similar remarks apply to the complex matrices.

In chapter 25 we defined the algebraic sets only over an algebraically closed field, but they may be defined the same way over any field \( F \), and form a category with the polynomial maps (note, though, that the Nullstellensatz might not hold). We define a linear algebraic group over \( F \) to be a group object in this category. Thus, it is an algebraic set \( G \), with morphisms (polynomial maps) \( m : M \times M \rightarrow M \) and \( i : M \rightarrow M \), satisfying the group axioms. The general linear group, which we denote as \( GL_F(n) \), or simply \( GL(n) \), can be considered such, by adding an extra variable \( \delta \) for \( 1/\det(x) \). \( GL(n) \) is defined by the polynomial \( \det(x)^{\delta} = 1 \) where \( \det(x) \) is a polynomial in the \( x_{ij} \). Multiplication and inverse are easily seen to be polynomial maps.

A subgroup \( H \) of a linear algebraic group \( G \) which is closed in the Zariski topology is a clearly linear algebraic group. In particular, such subgroups of \( GL(n) \) are. Such groups with \( F = \mathbb{R} \) or \( F = \mathbb{C} \) include the main examples of Lie groups for applications.

It is readily seen that the Euclidean topology is stronger than the Zariski topology, whence a subgroup closed in the Zariski topology is closed in the Euclidean topology. A theorem of Cartan (see [Serre2]) implies that it is a Lie group; the proof requires a fair amount of machinery. Alternatively, this can be shown by particular arguments for specific groups; see [Vara] for some such. We give a general theorem for Zariski closed subgroups of \( GL(n) \) in section 3, which can also be found in [Vara].
Commonly encountered Zariski closed subgroups of $GL(n)$ include the following. Unless otherwise noted, the group is defined in both the real and complex cases. Let $I$ denote the $n \times n$ identity matrix.

- The special linear group $SL(n)$, $\{ x \in GL(n) : \det(x) = 1 \}$.
- The orthogonal group $O(n)$, $\{ x \in GL(n) : x^t x = I \}$.
- The special orthogonal group $SO(n)$, $O(n) \cap SL(n)$.
- The unitary group $U(n)$, $\{ x \in GL(n) : x^t x = I \}$, in the complex case only (but $U(n)$ is not a complex Lie group; see below).
- The special unitary group $SU(n)$, $U(n) \cap SL(n)$, in the complex case only.
- The symplectic group $Sp(2n)$, $\{ x \in GL(n) : x^t J x = J \}$, where $J$ is the matrix with $n$ copies of $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ down the diagonal. Authors vary on whether the notation $Sp(2n)$ or $Sp(n)$ is used. In the complex case, $Sp(2n) \cap U(2n)$ is called the unitary symplectic group; although the notation is not standard, we use $Spu(2n)$ to denote it.

The generalized real orthogonal group $O(n, r)$, $\{ x \in GL(n) : x^t J x = J \}$ where $J$ is the matrix with $r + 1$’s and $n - r$ -1’s down the diagonal, is also commonly encountered; $O(4, 3)$ for example is the Lorentz group of relativity theory. Other examples of Zariski closed subgroups include the upper triangular matrices, the upper triangular matrices with 1’s on the diagonal, and the diagonal matrices.

Various redundancies exist in the above list. Various relations exist between the real and complex groups also. For example $GL_C(n)$ can be embedded in $GL_R(2n)$. Each entry $z = z_1 + iz_2$ in a column vector is replaced by $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$. Each entry $a = a_1 + ia_2$ in a matrix is replaced by $\begin{bmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{bmatrix}$.

The unitary group $U(1)$ has real dimension 1, so cannot be a complex Lie group. By results of section 3, it is not a Zariski closed subset; this can be seen directly. The Zariski closed subsets of $C$ are the finite sets and $\mathbb{C}$, so $\{ z : z\bar{z} = 1 \}$ is not a closed set, and it follows that $U(n)$ is not a closed set in the $n \times n$ complex matrices. In fact it is not a closed set in $GL(n)$; $\{ (z, w) : z\bar{z} = 1 \text{ and } zw = 1 \}$ is not a closed subset of $C^2$, because its projection would be a constructible subset of $C$ ([CLO], Corollary 5.6.9), and the claim follows. It is also worth noting that the map $z \mapsto z^*$ does not satisfy the Cauchy-Riemann equations (see [Ahlfors]), and so is not analytic.

The image of $U(n)$ (indeed any of the above complex groups) under the above embedding of $GL_C(n)$ in $GL_R(2n)$ is readily seen to be a Zariski closed subgroup of $GL_R(2n)$, though. The image is thus a real Lie group; the same applies for $SU(n)$ and $Spu(2n)$.

Generalizations of the groups defined above are of great interest in group theory, when $F$ is a finite field (see [Carter]). There are various other theorems concerning relationships between linear algebraic groups and Lie groups; we omit further discussion, and refer the reader to the literature. We mention that general algebraic groups are defined to be the group objects in the appropriate category of algebraic sets, which is a category of schemes (see [Springer]). Projective algebraic groups are also called Abelian varieties, and are the other main type considered. There are theorems that under suitable hypotheses, a general algebraic group is an extension of a linear algebraic group by an Abelian variety; see [Conrad].

Another type of Lie subgroup $H \subseteq G$ of a Lie group $G$ is obtained, when $H$ is a subgroup and an open subspace. Indeed, $H$ is an analytic manifold with the inherited charts (restriction of $\phi$ to $U \cap H$), and the restrictions of the group operations are clearly analytic. By remarks in section 24.2, the component of the identity is a clopen normal subgroup. For example, $O(n)$ has two components, the matrices of determinant $+1$ and $-1$ (these are clearly both closed subspaces); $SO(n)$ is the component of the identity.

2. Differentiation on manifolds. Suppose $f : F^n \to F$, where $F = R$ or $F = C$, is analytic on an open subset $U \subseteq F^n$. In exercise 7 it is shown that $\partial f / \partial x_i$ is analytic in $U$. It follows by results in section 24.12 that $f$ is differentiable in $U$. Further, it follows that the derivative is analytic. By induction,
the derivative of any order exists; such a function is said to be $C^\infty$.

For multinomial exponents $\mu, \nu \in \mathbb{N}^n$ let

$$\Phi_{\mu, \nu} \text{ denote } \frac{(\mu_1 + \nu_1)! \cdots (\mu_n + \nu_n)!}{\nu_1! \cdots \nu_n!}.$$  

From exercise 7 and induction we may conclude that if $f$ is given by $\sum_\mu a_\mu \xi^\mu$ at $x \in U$, then $\partial f/\partial x^\nu$ is given by $\sum_\mu a_{\mu + \nu} \Phi_{\mu, \nu} \xi^\mu$. Note that the result is independent of the order in which the partial derivatives is taken; as mentioned in section 24.6 this holds more generally.

Although we will continue to assume that manifolds are analytic, in many contexts it suffices that they be $C^\infty$, and many facts hold with even weaker hypotheses. By observations in section 24.12, the tangent bundle of a $C^\infty$ manifold is a $C^\infty$ manifold. Indeed, the argument showing this shows that the tangent bundle of an analytic manifold is an analytic manifold. Similar arguments show that the bundle of mixed tensors of any type is an analytic manifold.

The analytic functions $f : U \rightarrow F$, where $U \subseteq M$ is an open subset of the analytic manifold $M$, comprise an $F$-algebra $A$ with the pointwise operations. A linear operator $D : A \rightarrow A$ is said to be a derivation at $x$ if $D(fg) = f(x)D(g) + g(x)D(f)$. These are readily verified to comprise a vector space over $F$ with operations $(D_1 + D_2)(f) = D_1(f) + D_2(f)$ and $(aD)(f) = aD(f)$.

When $M$ is $F^n$, the map $f \mapsto \partial f/\partial x_i(x)$ is readily verified to be a derivation at $x$ on $A$, using facts from sections 24.7 and 24.12; we denote it as $D^x_i$. As mentioned in section 24.12, the space generated by these is in fact all of the derivations at $x$; we omit a proof, and simply consider only this space.

For arbitrary $M$, let $\phi$ be a chart with domain $U$, and let $x$ be an element of $U$. The map $D \mapsto \tilde{D}$, where $\tilde{D}(f) = D(f \circ \phi^{-1})$, is readily verified to be an isomorphism of the derivations at $x$, to the derivations at $\phi(x)$ on the functions on $\phi[U]$. We use $D^\phi_i$ to denote the derivation at $x$ whose value at $f$ equals $(\partial(f \circ \phi^{-1})/\partial x_i)(\phi(p))$.

The notation $\partial f/\partial \phi_i$ is often used for $D^\phi_i(f)$, with $x$ understood in the former. Indeed, we have used this in the tensor transformation law in section 24.12. Note that derivations at $x$ obey the covariant vector transformation law $D^\phi_i(f) = \sum_{i'} D^\psi_{i'}(\phi_i) D^\psi_{i'}(f)$; this follows readily by writing $f \circ \psi^{-1}$ as $(f \circ \phi^{-1}) \circ (\phi \circ \psi^{-1})$ and using the chain rule, noting that $(\phi \circ \psi)^{-1}_i = \phi_i \circ \psi^{-1}$. Letting $f$ be $\psi_k$, it follows that the matrix whose $ij$ entry is $D^\phi_i(\phi_k)$ is invertible, the $ki$ entry of the inverse being $D^\phi_i(\psi_k)$.

If $f : M \rightarrow F$ is analytic let $df$ be the analytic section of $T^0_1(M)$ where $df(x) = \{w \mid w^iD^\phi_i(f)\}$. This is called the differential of $f$. With respect to a chart $\phi$ at a point $x$, if the tangent vectors are considered as column vectors then $df(x)$ may be considered as the row vector $\langle D^\phi_1(f), \ldots, D^\phi_n(f) \rangle$.

3. Zariski closed subgroups of $GL(n)$. In this section $GL(n)$ will be considered to be a Zariski open subset of the $n \times n$ matrices over $F$, where $F = \mathbb{R}$ or $F = \mathbb{C}$. Specializing a definition of section 24.11, a $k$-dimensional submanifold $Y$ of an $n$-dimensional analytic manifold $X$ is a subspace, such that at any point $y \in Y$ there is a chart $\phi$ with domain $U$ containing $y$ with $\phi[Y \cap U] = \phi[U] \cap F^k$. By $F^k$ we mean a fixed copy of $F^k$ as a subspace of $F^n$.

Terminology varies, and some authors call this a regular submanifold; a figure 8 in the real plane is an example which is not regular.

**Lemma 1.** If $G$ is a Lie group, and $H \subseteq G$ is a subgroup and a submanifold, then $H$ is a Lie group with the inherited charts.

**Proof:** It must be shown that multiplication $m$ and inverse $i$ are analytic. Given $h_1, h_2 \in H$, and charts around them, consider charts of $G$ from which they are inherited. The group operations are analytic on $F^n \times F^n$, so they are on $F^k \times F^k$.

**Lemma 2.** Suppose $U$ is an open connected subset of $F^n$, $f$ is analytic on $U$, $V \subseteq U$ is open, and $f(x) = 0$ for $x \in V$; then $f(x) = 0$ for $x \in U$.
Proof: First we prove the claim when $U$ is convex. Suppose $u \in U$ and $v \in V$. Let $g(t) = f((1-t)v + tu)$; then $g$ is analytic at every point of the closed interval $[0,1] \subseteq \mathbb{R}$. Let $K = \{ s \in [0,1] : g(t) = 0 \text{ for } 0 \leq t \leq s \}$. Let $t_m = \sup(K)$. By hypothesis $t_m > 0$. By continuity $g(t_m) = 0$. There is a power series such that $g(t_m + \tau) = \sum c_i \tau^i$ in some interval. If some $c_i$ is nonzero let $i$ be least such that $c_i$ is nonzero; then $g(t) = z^i h(\tau)$ where $h(0) \neq 0$; but this is a contradiction, so every $c_i$ is 0. Thus, $t_m$ must equal 1, else it is not the supremum. If $U$ is any open set let $K = \{ x \in U : f(x) = 0 \}$. Let $x$ be a limit point of $K^{\text{int}}$. Let $P$ be an open polydisc containing $x$ on which there is a power series for $f$ at $x$. Then $P$ is convex, and $f(x) = 0$ on some open subset of $P$, so $f(x) = 0$ on $P$. This shows that $K^{\text{int}}$ is closed, and since $U$ is connected $K = U$.

An open subset $U \subset F^n$ may be equipped with the Zariski topology. Given a Zariski closed subset $W \subseteq U$, and a point $x \in W$, let $r_x$ be the rank of the subspace generated by the cotangent vectors $\langle D_1^2(p), \ldots, D_n^2(p) \rangle$ for $p \in I(S)$ for $r = \max\{r_x : x \in W\}$. Let $W_m = \{ x \in W : r_x = r \}$.

Lemma 3. With notation as above, if $W$ is nonempty then $W_m$ is a nonempty open subset of $W$, and a submanifold of $F^n$ of dimension $n - r$.

Proof: Clearly $W_m$ is nonempty. Let $\{ p_1, \ldots, p_s \}$ generate $I(W)$. From the fact that $D_i^2(pq) = D_i^2(p)D_i^2(q) + q(x)D_i^2(p)$ it follows that the cotangent vectors $\langle D_1^2(p_j), \ldots, D_n^2(p_j) \rangle$ for $1 \leq j \leq s$ generate the entire space of cotangent vectors. Then $x \in W$ iff there is a $r \times r$ minor of the matrix with these $s$ cotangent vectors as rows, with nonzero determinant. It follows that $W_m$ is open. By reordering the variables, and shifting $x$ to the origin, there is an $a \in F$ such that the map $\phi$ where $\phi(t) = \langle t_1, \ldots, t_{n-r}, p_1(t), \ldots, p_r(t) \rangle$ maps $(-a,a)^n$ bijectively to an open subset $V \subseteq U$. Denote then $\phi^{-1}$ is a chart on the analytic manifold $F^n$, which maps $V \cap W_m$ to $F^{n-r}$. Let $V_0 = \{ x \in V : p_i = 0 \text{ for } 1 \leq i \leq r \}$. It suffices to show that $V_0 \subseteq W$, since this implies $V_0 = V \cap W_m$ and maps under $\phi$ to $F^{n-r} \cap (-a,a)^n$. For this it suffices to show that if $p \in I(W)$ and $x \in V_0$ then $p(x) = 0$. Let $I = I(W)$, let $A$ be the $F$-algebra of analytic functions on $V$, and let $\bar{I}$ be the ideal in $A$ generated by $I$. Given $q \in I$, and $k$ with $1 \leq k \leq n-r$, let $q_k = q$, $q_i = t_i$ if $1 \leq i \leq n-r$ and $i \neq k$, and $q_i = p_i-(n-r)$ if $i > n-r$. Let $A_1, A_2, A_3$ be the matrices whose $ij$ entry is $\partial q_i/\partial \phi_j$, $\partial q_i/\partial t_j$, and $\partial t_i/\partial \phi_j$ respectively; then $A_1 = A_2A_3$. Further, $\det(A_1) = \partial q/\partial \phi_k$, $\det(A_2)$ is a polynomial which must vanish on $W$ because of the rank condition on the matrix of partials $\partial p_i/\partial t_j$, and $\det(A_3)$ is in $A$. Thus, $\partial q/\partial \phi_k \in \bar{I}$ for any $q \in I$, whence by the derivation property $\partial f/\partial \phi_k \in \bar{I}$ for any $f \in \bar{I}$. It follows that $\partial f/\partial \phi_k \in \bar{I}$ for any $f \in \bar{I}$ and $\mu \in F^{n-r}$. It follows from this and the fact that $f(x) = 0$ that in the series for $f$ in terms of the $\phi_i$ at 0, all coefficients $c_{\mu}$ with $\mu \in F^{n-r}$ are 0. It follows using lemma 2 that $f$ is identically 0 on $V_0$.

Theorem 4. A Zariski closed subgroup $G \subseteq GL(n)$ is a Lie group.

Proof: By lemmas 2 and 1 it suffices to show that $G_m = G$. Let $x_{ij}$ be indeterminates for the entries of a matrix of $GL(n)$, and let $I = I(G)$ in $F[x]$. For $p \in F[x]$ and $m \in G$ let $p^m$ be defined by $p^m(x) = p(mx)$. The map $p \mapsto p^m$ is an automorphism of $F[x]$ which preserves $I$. Letting $Df$ denote the array of partial derivatives $\partial f/\partial x_{ij}$, $p \mapsto p^m$ induces a $F$-linear isomorphism of the vector space generated by $\{ Di : p \in I \}$ and $\{ Dp^m : p \in I \}$. It follows that the spaces generated by $\{ Di(n) : p \in I \}$ and $\{ Dp^m(n) : p \in I \}$ are isomorphic.

4. Lie algebras. Two main applications of Lie algebras occur in the theories of Lie groups, and linear algebraic groups. In the former, the field $F$ is complete; in the latter it is arbitrary. Over an arbitrary field $F$, a Lie algebra is a generalized algebra (which recall from section 20.13 is a vector space with a bilinear form), with bilinear form $[x,y]$, which satisfies the following axioms.

1. $[x,x] = 0$
2. \([x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0.\)

The first axiom states that the bilinear form is alternating; it follows as usual that \([x, y] + [y, x] = 0\), and if the characteristic of \(F\) is not 2 then the latter identity implies \([x, x] = 0.\) The second axiom is called the Jacobi identity.

The Lie algebras over \(F\) are the objects of a category of the type \(\text{Mdl}_F\) for a system of equations \(T.\) If \(L\) is a Lie algebra and \(U, V\) are subspaces the notation \([U, V]\) is used to denote the subspace generated by \(\{u, v : u \in U, v \in V\}.\) Note that \([U, V] = [V, U].\) A subspace \(J\) such that \([L, J] \subseteq J\) is called an ideal. The following facts about ideals in Lie algebras may be shown to hold by similar arguments to those for ideals in rings.

- The ideals are in bijective correspondence with the congruence relations; the kernel of a homomorphism is defined to be the ideal.
- The ideals form an algebraic closure system; the join is the sum.
- An ideal is a subalgebra.
- If \(I\) and \(J\) are ideals then the canonical isomorphism of \((I + J)/I\) and \(I/(I \cap J)\) as vector spaces is a Lie algebra isomorphism.

If \(A\) is an associative \(F\)-algebra, the commutator \([x, y]\) of two elements is defined to be \(xy -yx.\) One readily verifies that the vector space \(A,\) with the operation \([x, y],\) is a Lie algebra. We state without proof that any Lie algebra \(L\) may be embedded in an associative algebra \(A,\) so that the Lie product becomes the commutator, in a universal manner; \(A\) is called the universal enveloping algebra. A proof may be found in [Serre2] or [Vara].

If \(A\) is a generalized \(F\)-algebra, a derivation on \(A\) is defined to be a linear operator \(\delta : A \mapsto A\) such that \(\delta(x \cdot y) = x \cdot \delta(y) + \delta(x) \cdot y.\) Straightforward computation verifies that with pointwise operations, the derivations form a vector space over \(F;\) call this \(D.\) Note that it is a subspace of the associative \(F\)-algebra, where the multiplication is composition. Given \(\delta_1, \delta_2 \in D\) let \([\delta_1, \delta_2]\) be \(\delta_1 \circ \delta_2 - \delta_2 \circ \delta_1.\) Straightforward computation shows that for \(\delta_1, \delta_2 \in D, \ [\delta_1, \delta_2] \in D.\) Thus, \(D\) is a Lie algebra with this bilinear form. \(D\) is called the derivation algebra of \(A,\) and denoted \(\text{Der}(A).\)

If \(L\) is a Lie algebra and \(x \in L\) let \(\text{Ad}_x\) be the map from \(L\) to \(L,\) where \(\text{Ad}_x(y) = [x, y].\) Straightforward computation shows that \(\text{Ad}_x\) is a derivation on \(L.\) Additional straightforward computation shows that the map \(x \mapsto \text{Ad}_x\) is a Lie algebra homomorphism from \(L\) to \(\text{Der}(L).\)

A representation of a Lie algebra \(L\) is defined to be a map \(\rho : L \mapsto E\) where \(E = \text{End}_F(V)\) for an \(F\)-linear space \(V,\) such that when \(E\) is considered as a Lie algebra with the commutator as the bilinear form, \(\rho\) is a Lie algebra homomorphism. Representations are of most interest when \(V\) is finite dimensional. As for groups, in a finite dimensional representation the elements are represented by matrices; in a group representation the group product is represented by matrix multiplication, whereas in a Lie algebra representation the Lie product is represented by the commutator.

The map \(x \mapsto \text{Ad}_x\) is a representation, with \(V\) being \(L\) itself, considered as a vector space. This representation is called the adjoint representation of \(L.\) If \(L\) is finite dimensional, the trace of the matrix \(\text{Ad}_x \text{Ad}_y,\) as a function of \(x\) and \(y,\) may be verified to be a bilinear form on \(L,\) called the Killing form.

A Lie algebra is said to be Abelian if \([x, y] = 0\) for all \(x, y.\) In this case, \([x, y] = [y, x];\) if the characteristic is not 2 then the latter identity implies \([x, y] = 0.\)

Define \(C_0(L) = L,\) and \(C_{k+1}(L) = [L, C_k(L)].\) Clearly, \(C_1(L) \subseteq C_0(L,\) and if \(C_{k+1}(L) \subseteq C_k(L)\) then \(C_{k+1}(L) \subseteq C_k(L),\) so by induction the series \(C_0(L), \ldots\) is descending. This series is called the lower central series. If it eventually becomes 0 then \(L\) is said to be nilpotent. In this case, the series must be strictly descending. If \(L\) is finite dimensional then the series eventually becomes constant.

Define \(D_0(L) = L,\) and \(D_{k+1}(L) = [D_k(L), D_k(L)].\) By a similar induction to the preceding the series \(D_0(L), \ldots\) is descending. This series is called the derived series. If it eventually becomes 0 then \(L\) is
said to be solvable. In this case, the series must be strictly descending. If \( L \) is finite dimensional then the series eventually becomes constant. It follows from \([z, [x, y]] = -[x, [y, z]] = [y, [z, x]]\) that if \( I \) is an ideal then \([I, I]\) is an ideal. Thus, \( Ds_k(L) \) is an ideal.

An Abelian Lie algebra is nilpotent, and a nilpotent Lie algebra is solvable. The first implication is a triviality, and for the second, it follows from \( Ds_k(L) \subseteq Cs_k(L) \) that \([Ds_k(L), Ds_k(L)] \subseteq [L, Cs_k(L)]\), and \( Ds_k(L) \subseteq Cs_k(L) \) follows by induction.

As ideals are subalgebras, we may speak of nilpotent or solvable ideals. A subalgebra of a nilpotent subalgebra is nilpotent, and likewise for solvable subalgebras. As is readily verified, for a homomorphism \( \phi \) for the partials in \( V \) said to be solvable. In this case, the series must be strictly descending. If \( L \) is a Lie algebra, \( I \subseteq L \) is a solvable ideal, and \( L/I \) is solvable, then \( L \) is solvable (exercise 8).

The sum \( I + J \) of two solvable ideals is solvable; this follows because \((I + J)/I \) is solvable, since it is isomorphic to \( J/(I \cap J) \). A Lie algebra \( L \) thus has only one maximal solvable ideal. It is called the radical of \( L \). A Lie algebra is said to be semisimple if its radical is 0. It is readily seen that this is the case iff it contains no nontrivial solvable ideal, iff it contains no nontrivial Abelian ideal. A Lie algebra which is not Abelian, and contains no nontrivial proper ideal, is said to be simple.

The normalizer \( N(V) \) of a subspace \( V \) of a Lie algebra \( L \) is \( \{x : Ad_x[V] \subseteq V\} \). That \( N(V) \) is a subspace follows by bilinearity. In fact, \( N(V) \) is a subalgebra, because \([[[x, y], v] = [x, [y, v]] - [y, [x, v]]]\). If \( V \) is a subalgebra then \( V \) is an ideal in \( N(V) \), as is readily verified.

For the rest of the section, \( F \) is assumed to have characteristic 0 and Lie algebras are assumed to have finite dimension. No proofs are given; proofs may be found in [Vara]. A Lie algebra is semisimple iff its Killing form is nondegenerate. The canonical epimorphism from a Lie algebra \( L \) with radical \( R \) to \( L/R \) splits, and \( L \) is a semidirect product of a semisimple Lie algebra and a solvable Lie algebra. A semisimple Lie algebra is an internal direct sum of simple Lie algebras.

A Cartan subalgebra \( K \subseteq L \) of a Lie algebra \( L \) is defined to be a nilpotent subalgebra such that \( N(K) = K \). Any real or complex Lie algebra has a Cartan subalgebra. A root system may be defined, using a Cartan subalgebra. This determines the algebra up to isomorphism, and is one of those having a connected Dynkin diagram, as listed in section 23.6.

5. Lie algebra of a Lie group. In this section \( F \) again equals \( R \) or \( C \). \( T \) denotes the tangent bundle functor.

**Theorem 5.** Suppose \( M \) is an \( n \)-dimensional analytic manifold, and \( V \) and \( W \) are analytic vector fields on \( M \). In a chart \( \phi \), let

\[
u_i = \sum_{j=1}^{n} \left( v_j \frac{\partial u_i}{\partial x_j} - u_j \frac{\partial v_i}{\partial x_j} \right).
\]

a. The \( u_i \) transform as a contravariant vector.

b. The \( u_i \) are the components in \( \phi \) of an analytic vector field \( U \) on \( M \), called the Lie bracket of \( V \) and \( W \), and denoted \([V, W]\).

c. The analytic vector fields on \( M \), with the Lie bracket, form a Lie algebra.

**Proof:** For part a, we write \( v^i \) (\( \bar{v}^i \)) for the components of a contravariant vector in the chart \( \phi \) (\( \psi \)); and

\[
\frac{\partial}{\partial \bar{x}^i}, \frac{\partial}{\partial x^i}, \quad \text{and} \quad \frac{\partial \bar{v}^i}{\partial x^j}
\]

for the partials in \( \phi \), the partials in \( \psi \), and the components of the transformation matrix \((\psi \circ \phi^{-1})'\). Also, we use the Einstein summation convention. Then

\[
\bar{v}^k \frac{\partial}{\partial \bar{x}^i} = \frac{\partial \bar{x}^i}{\partial x^k} v^k \frac{\partial}{\partial \bar{x}^i} = \delta^i_j \bar{v}^k \frac{\partial}{\partial \bar{x}^i} = v^j \frac{\partial}{\partial x^j}.
\]

394
Also,
\[ \frac{\partial \bar{u}^i}{\partial x^j} = \frac{\partial}{\partial x^j} \left( \frac{\partial \bar{x}^i}{\partial x^k} w^k \right) = \frac{\partial^2 \bar{x}^i}{\partial x^j \partial x^k} w^k + \frac{\partial \bar{x}^i}{\partial x^j} \frac{\partial w^k}{\partial x^j}. \]
Applying the first identity, then the second, to
\[ \bar{u}^i = \bar{v}^j \frac{\partial \bar{x}^i}{\partial x^j} - \bar{w}^j \frac{\partial \bar{v}^i}{\partial x^j}, \]
and canceling the second partials terms, yields the claim. For part b, part a shows that a map \( M \mapsto T(M) \) may be defined, since the bundle equivalence relation is respected. It is clearly analytic in the charts \( \phi \) on \( M \) and \( \phi \times \iota \) on \( T(M) \). For part c, consider a contravariant vector as a derivation \( \sum v_i (\partial/\partial x_i) \). These may be considered as an algebra of operators on the analytic functions on an open set \( U \), with composition as multiplication. The Lie bracket is just the commutator in this algebra, as is readily verified. As was noted in section 4, with this operation, the vector fields form a Lie algebra.

The notation used in the proof of part a is widely used. The Lie bracket may also be described using the “Lie derivative”; see [Spivak].

If \( G \) is a Lie group let \( \lambda_g : G \mapsto G \) denote the left translation \( h \mapsto gh \). By definition of a Lie group this map is a morphism in the category of analytic manifolds; it is in fact an isomorphism, with inverse \( \lambda_{g^{-1}} \). \( T(\lambda_g) : T(G) \mapsto T(G) \) is also an isomorphism. The tangent space \( T_\lambda(G) \) is mapped isomorphically to the tangent space \( T_{gh}(G) \). Letting \( I \) denote the identity element of \( G \), if \( \tau = ([\alpha, 1, w]) \) is element of \( T_1(G) \) then \( T(\lambda_\tau)(\alpha) = [(\beta, g, (\phi_\beta (\lambda_g^{-1})(\phi_\alpha(1))(w))] \). A vector field \( V \) on \( G \) is said to be left invariant if \( T(\lambda_g)(V(h)) = V(gh) \) for all \( g, h \in G \).

If \( f : M \mapsto N \) is a morphism of analytic manifolds, \( V : M \mapsto T(M) \) is a vector field, and \( \bar{V} : N \mapsto T(N) \) is a vector field, say that \( V \) and \( \bar{V} \) are \( f \)-related, \( W \) and \( \bar{W} \) are \( f \)-related, \( U = [V, W] \), and \( \bar{U} = [\bar{V}, \bar{W}] \). Then \( U \) and \( \bar{U} \) are \( f \)-related.

**Proof:** Using the notation preceding the theorem, the lemma follows just as part a of theorem 5.

**Theorem 7.** The Lie bracket of two left invariant vector fields on a Lie group \( G \) is left invariant.

**Proof:** Letting \( f = T(\lambda_g) \), let \( \bar{V} = T(f) \circ V \circ f^{-1} \). \( T(\lambda_g)(V(h)) = T(\lambda_g)(V(\lambda_g^{-1}(gh))) = \bar{V}(gh) \) for any \( V \). So
\[ T(\lambda_g)([V, W](h)) = [V, W]f(gh) = [\bar{V}, \bar{W}](gh) \]
by lemma 6. But by hypothesis \( \bar{V}(gh) = V(gh) \) and \( \bar{W}(gh) = W(gh) \). It follows that \( [\bar{V}, \bar{W}](gh) = [V, W](gh) \).

For \( \tau \) an element of \( T_1(G) \), let \( \bar{\tau} \) denote the map \( g \mapsto T(\lambda_g)(\tau) \). We claim that it is analytic, and hence is an analytic vector field on \( G \). To see this, note that if \( \phi_\beta(g) = x \) then the image of \( \bar{\tau}(g) \), under the chart given in section 24.12 for the tangent bundle, is \( \langle x, (\phi_\beta \lambda_g^{-1})(\phi_\alpha(1))(w) \rangle \), so it suffices to show that \( \phi_\beta \lambda_g^{-1} \), as a function of \( x \) and \( y = \phi_\alpha(h) \), is analytic around \( x \) and \( \phi_\alpha(1) \). This follows by the analyticity of multiplication.
We have \(T(\lambda_g)(\tau(h)) = T(\lambda_g)(T(\lambda_h)(\tau)) = T(\lambda_g \lambda_h)(\tau) = T(\lambda_{gh})(\tau) = \tau(gh)\); thus, \(\tau\) is left invariant. On the other hand, if \(V\) is a left invariant vector field then \(V = \tilde{\tau}\) where \(\tau = V(1)\), as direct computation also shows. The map \(\tau \mapsto \tilde{\tau}\) is readily verified to be linear, and so is a vector space isomorphism. As a result, the Lie bracket on a Lie group may be “transferred” to the fiber \(T_1(G)\).

The Lie algebra is thus determined by the group operation on any open neighborhood of the identity. In particular it is determined by the subgroup on the component of the identity. The Lie algebra does not completely determine the Lie group, but does determine its “infinitesimal” behavior. For further discussion see the literature.

Let \(L(G)\) denote the Lie algebra of the Lie group \(G\). \(L\) is the object map of a functor from the Lie groups to the Lie algebras. If \(f : G \to H\), \(T(f)\) induces a map from \(T_1(G)\) to \(T_1(H)\), which may be verified to be a Lie algebra homomorphism.

6. Covering projections. Covering projections are a topic in topology which are useful in the study of Lie groups, and other areas of topology and analysis, such as Riemann surfaces. Suppose \(p : Y \to X\) is a continuous function between topological spaces. A subset \(U \subseteq X\) is said to be evenly covered by \(p\) if \(p^{-1}[U]\) is a disjoint union of open subsets of \(Y\), called sheets, each of which is mapped homeomorphically to \(U\) by \(p\). If for each \(x \in X\) there is an open set \(U\) with \(x \in U\) which is evenly covered by \(p, p\) is called a covering projection. A basic example is \(e^{it}\) from the real line to the unit circle in the complex plane.

Recall that the components of a locally path connected space are its path components. The following are readily verified, for a covering projection \(p : Y \to X\):

- If \(U \subseteq X\) is open and connected, and evenly covered by \(p\), then the sheets of \(p^{-1}[U]\) are its components.
- If \(X\) is locally (path) connected then \(Y\) is.
- If \(X\) is locally (path) connected then \(p\) is a covering projection iff \(p\) is a covering projection for each component \(C\) of \(X\).
- If \(X\) is locally (path) connected and \(D\) is a component of \(Y\) then \(p[D]\) is a component of \(X\), and \(p[D]\) is a covering projection.

If \(p : Y \to X\) and \(f : Z \to X\), \(f\) is said to lift if it factors through \(p\), that is, there is some \(\tilde{f} : Z \to Y\) such that \(f = p\tilde{f}\); \(\tilde{f}\) is called a lift of \(f\).

**Theorem 8.** Suppose \(p : Y \to X\) is a covering projection, \(f : Z \to X\), \(Z\) is connected, \(\tilde{f}_1\) and \(\tilde{f}_2\) are lifts of \(f\), and \(\tilde{f}_1(z) = \tilde{f}_2(z)\) for some \(z\). Then \(\tilde{f}_1(z) = \tilde{f}_2(z)\) for every \(z\).

**Proof:** Let \(Z_e = \{z : \tilde{f}_1(z) = \tilde{f}_2(z)\}\). For \(z \in Z\) let \(U\) be an open subset of \(X\) containing \(f(z)\) which is evenly covered by \(p\). For \(i = 1, 2\) let \(V_i\) be the sheet of \(p^{-1}[f(z)]\) containing \(\tilde{f}_i(z)\). Then \(W = \tilde{f}_1^{-1}[V_1] \cap \tilde{f}_2^{-1}[V_2]\) is an open set containing \(z\). If \(z \in Z_e\) then \(W \subseteq Z_e\), and if \(z \in Z^e\) then \(W \subseteq Z^e\), as is readily verified. Thus, \(Z_e\) is clopen; by hypothesis it is nonempty, and since \(Z\) is connected \(Z_e = Z\).

A map \(p : Y \to X\) is called a fibration if whenever \(f : Z \times [0, 1] \to X\) and \(\tilde{f}_0 : Z \to Y\) are continuous functions such that \(p(\tilde{f}_0(x)) = f(z, 0)\) for \(z \in Z\), there is a continuous function \(\bar{f} : Z \times [0, 1] \to Y\) such that \(p \circ \bar{f} = f\).

In particular, if \(f_0\) lifts to \(\tilde{f}_0\), and \(f_1\) is homotopic to \(f_0\) in \(X\), then \(f_1\) lifts to a function \(\tilde{f}_1\) which is homotopic to \(\tilde{f}_0\) in \(Y\). One says that the lifting problem is a problem in the homotopy category.

**Theorem 9.** A covering projection is a fibration.

**Proof:** Suppose \(W \subseteq Z\) is open, \(0 \leq s < t \leq 1\), \(U \subseteq X\) is open and evenly covered by \(p, f[U \times [s, t]] \subseteq U\), and \(\tilde{f}_s\) is a lift of \(f[U \times [s, t]]\). To construct a lift of \(f[U \times [s, t]]\), define \(\tilde{f}(z, r) = \phi_i(\tilde{f}(z, r))\) where \(\phi_i\) is the homeomorphism from \(U\) to the sheet over \(U\) containing \(\tilde{f}_s(z)\); \(\tilde{f}\) is continuous because its corestriction to each sheet is. Given \(z \in Z\) there is an open set \(W_z \subseteq Z\), and real numbers \(t_0, \ldots, t_n\), such that for \(0 \leq i < n\)
Suppose $p : Y \hookrightarrow X$ is a covering projection, $\alpha$ is a path in $X$ from $x_0$ to $x$, and $p(y_0) = x_0$. It follows using theorem 9 that there is a lift $\bar{\alpha}$ of $\alpha$ which starts at $y_0$. By theorem 8 such a lift is unique. More generally, we have the following. For this section, recall the notation $\alpha \beta$ from chapter 24 for the concatenation of two paths, and let $\alpha^{-1}$ denote the reversed path.

**Theorem 10.** Suppose $p : Y \hookrightarrow X$ is a covering projection, $f : Z \hookrightarrow X$, $Z$ is path connected, and $f(z_0) = p(y_0)$. Then $f$ lifts to $Y$ iff $f_L[\pi(Z,z_0)] \subseteq p_L[\pi(Y,y_0)]$. Further if so then there is a lift $\bar{f}$ with $\bar{f}(z_0) = y_0$.

**Proof:** If a lift $\bar{f}$ exists then

$$f_L[\pi(Z,z_0)] = p_L\bar{f}_L[\pi(Z,z_0)] \subseteq p_L[\pi(Y,y_0)].$$

Suppose $f_L[\pi(Z,z_0)] \subseteq p_L[\pi(Y,y_0)]$. Given $z \in Z$, let $\gamma$ be a path from $z_0$ to $z$. The lift $\beta$ starting at $y_0$ of $\alpha = f \circ \gamma$ is a path to some $y$. If $\gamma'$ is another path from $z_0$ to $z$, let $\beta'$ be the lift of $\alpha' = f \circ \gamma'$. The path $\gamma\gamma^{-1}$ is a closed loop $\Gamma$ in $Z$. This maps under $f$ to a closed loop $A$ in $X$, namely $\alpha\alpha^{-1}$. By hypothesis $A$ is the image of a closed loop $B$ in $Y$, which by uniqueness must be $\beta\beta^{-1}$. Thus, $y$ does not depend on $\gamma$ and so a map $\bar{f}(z) = y$ from $Z$ to $Y$ is determined. Clearly $p \circ \bar{f} = f$. Suppose $z \in Z$, $\bar{f}(z) = y$, and $V$ is an open set containing $y$. We may suppose $V$ is a sheet of $p^{-1}[U]$ where $U$ is an open subset of $X$ containing $f(z)$ which is evenly covered by $p$. Then $W = f^{-1}[U]$ is an open subset of $Z$ containing $z$, with $\bar{f}[W] \subseteq V$. Thus, $\bar{f}$ is continuous. Finally, $\bar{f}(z_0) = y_0$.

Suppose $X$ is connected and locally path connected. A covering projection $p : Y \hookrightarrow X$ is said to be universal if $Y$ is connected, and given any other covering projection $q : Z \hookrightarrow X$ where $Z$ is connected there is a continuous function $f : Y \rightarrow Z$ such that $q \circ f = p$.

A topological space $X$ is said to be semilocally 1-connected if for every $x \in X$ there is an open set $U$ with $x \in U$, such that every loop with base point $x$ that lies in $U$ is homotopic to the constant path at $x$. Note that a loop in such a $U$, based at some other point $x' \in U$, is homotopic to the constant path at $x'$.

**Theorem 11.** Suppose $X$ is connected and locally path connected.

- **a.** $X$ has a covering projection $p : Y \hookrightarrow X$ with $Y$ simply connected iff $X$ is semilocally 1-connected.
- **b.** If $X$ is semilocally 1-connected a covering projection $p : Y \hookrightarrow X$ is universal iff $Y$ is simply connected.

**Proof:** Suppose $p : Y \hookrightarrow X$ is a covering projection with $Y$ simply connected. For $x \in X$ let $U$ be an open subset of $X$ which is evenly covered by $p$. Considering the homeomorphism from $U$ to any sheet $V$ it follows that any loop in $U$ based at $x$ is homotopic to the constant path at $x$. Suppose $X$ is semilocally 1-connected. Let $x_0$ be any point of $X$. Let $P$ be the set of paths in $X$ starting at $x_0$. Say that paths $\alpha_1, \alpha_2 \in P$ are equivalent if their ends are the same, and $\alpha_1\alpha_2^{-1}$ is homotopic to the constant path at $x_0$. Let $Y$ be the quotient of $P$ by this equivalence relation. Let $p : Y \hookrightarrow X$ map $[\alpha]$ to $\alpha(1)$; clearly this is well defined. If $U$ is an open subset of $X$ and $\alpha$ is a path from $x_0$ with $\alpha(1) \in U$ let $B_{\alpha U}$ be the set of $[\alpha \xi]$ where $\xi$ starts at $\alpha(1)$ and lies in $U$. It is readily verified that if $\alpha_1$ and $\alpha_2$ are equivalent then $B_{\alpha_1 U} = B_{\alpha_2 U}$. Also, if $\alpha_3 \in B_{\alpha_1 U_1} \cap B_{\alpha_2 U_2}$ then $B_{\alpha_3, U_1 \cap U_2} \subseteq B_{\alpha_1 U_1} \cap B_{\alpha_2 U_2}$. Hence $\{B_{\alpha U}\}$ is a base for a topology on $Y$; we
suppose $Y$ is equipped with this topology. If $U \subseteq X$ is open and $p([\alpha]) \in U$ then $p(B_{\alpha,U}) \subseteq U$, showing that $p$ is continuous. Also, $p(B_{\alpha,U})$ is the path component of $U$ containing $\alpha(1)$, which is open by the hypothesis that $X$ is locally path connected, showing that $p$ is open, and also that $p$ is surjective on $B_{\alpha,U}$ when $U$ is path-connected. The union of the sets $B_{\alpha,U}$ such that $[\alpha] \in p^{-1}[U]$ equals $p^{-1}[U]$. It is readily verified that if $B_{\alpha,U} \cap B_{\alpha',U} \neq \emptyset$ then $B_{\alpha,U} = B_{\alpha',U}$. Now suppose $U$ is such that every loop that lies in $U$ is homotopic to the constant path at its base point (by hypothesis such $U$ cover $X$). We claim that on each $B_{\alpha,U}$ with $[\alpha] \in p^{-1}[U]$, $p$ is injective, the only remaining claim needed to show that $p$ is a covering projection. Suppose $p([\alpha_1]) = p([\alpha_2])$; then $\xi_1\xi_2^{-1}$ is a closed path lying in $U$, and it follows that $\alpha_1$ and $\alpha_2$ are equivalent. It remains to show that $Y$ is simply connected. If $x_0 \in X$ and $t \in [0,1]$ let $t$ be the path where $\alpha_{x_0}(u) = \alpha(tu)$. Let $\beta(t) = [\alpha_t]$; this is a map from $[0,1]$ to $Y$, and is in fact continuous. Indeed, suppose for some $t, \alpha' \in B_{\alpha,U}$; then $\alpha(t) = p(\beta(t))$, so $\alpha(t) \in U$, so $B_{\alpha,U} = B_{\alpha',U}$. Let $N$ be an open subset of $[0,1]$ such that $t \in N$ and $\alpha([N]) \subseteq U$. For $s \in N$ $\beta(s) = [\alpha_s] = [\alpha_{t\xi}]$ where $\xi(r) = \alpha(t + r(s-t))$. This shows that $\beta(s) \in B_{\alpha,U}$. Now, $p\beta = \alpha$, so $\beta$ is the lift of $\alpha$ starting at $y_0 = [\alpha_0]$. Given any loop based at $y_0$, it is the lift of its projection. Thus, $B(U,t) = [\alpha_{tu}]$ is a homotopy showing that the loop is homotopic to the constant path at $y_0$. Suppose $\alpha : Y \rightarrow X$ is a covering projection with $Y$ simply connected. Suppose $y_0 \in Y$ and $q(z_0) = p(y_0)$; then $\pi(Y,y_0) = 0$, so $p_L[p(Y,y_0)] \subseteq q_L[p(Z,z_0)]$, so by theorem 10 $p$ factors through $q$. Thus, $Y$ is universal. Suppose $X$ is semilocally 1-connected and $p : Y \rightarrow X$ is universal. By part 1 there is a covering projection $q : Z \rightarrow X$ with $Z$ simply connected. Since $Y$ is universal there is a continuous function $f : Z \rightarrow Y$ with $qf = p$. As just shown there is a continuous function $g : Y \rightarrow Y$ with $pg = q$; further by theorem 10 we may assume $g(f(x)) = x$. Since $qf$ is the identity, $gf$ is. Let $\beta$ be a loop in $Y$ at base $y$. Applying $f$ yields a loop $\gamma$ in $Z$ with origin $z$; this is homotopic to the constant map at $z$. Applying $g$ yields a homotopy of $\alpha$ to the constant map at $y$.

The induced function $f : Y \rightarrow Z$ from a universal covering space is a covering projection. Indeed, suppose $X$ is locally connected, $p : Y \rightarrow X$ and $q : Z \rightarrow X$ are covering projections, $f : Y \rightarrow Z$, and $q \circ f = p$. Suppose $U \subseteq X$ is connected and evenly covered by $p$ and $q$. It is easily seen that each component of $p^{-1}[U]$ is mapped homeomorphically to a component of $q^{-1}[U]$. In particular if $f$ is surjective it is a covering projection. If $X$ is locally path connected and $X$ and $Z$ are connected then using theorem 10 $f$ is surjective (given $z \in Z$ choose any $y \in Y$ and consider a path from $f(y)$ to $z$ in $Z$).

The proof of the theorem shows that the induced map from one universal covering space to another is bijective.

Although the definition will not be needed, a space is locally 1-connected if for every open set $V$ and $x \in V$ there is an open set $U \subseteq V$ with $x \in U \subseteq V$, such that every loop with base point $x$ that lies in $U$ is homotopic within $V$ to the constant path at $x$. The “Hawaiian earring” is the union of the circles $C_n$ in the Euclidean plane where $C_n$ has center $(1/n,0)$ and radius $1/n$. This is not semilocally 1-connected. A cone over this in Euclidean 3-space is semilocally 1-connected, but not locally 1-connected.

**Theorem 12.** Suppose $p : Y \rightarrow X$ is a covering projection where $Y$ is connected and locally path connected, and $X$ is a topological group. Suppose $p(f) = e$ where $e$ is the identity of $X$. Let $\mu : Y \times Y \rightarrow X$ be the map where $\mu(x,y) = p(x) \cdot p(y)$. Then $\mu$ lifts to a map $\tilde{\mu} : H \times H \rightarrow H$ where $\mu(f,f) = f$. Further this map makes $Y$ into a topological group, with $p$ a group homomorphism.

**Proof:** Suppose $\omega : t \mapsto (y_1(t), y_2(t))$ is a loop in $Y \times Y$ with base point $\langle f,f \rangle$. Let $t$ denote $p(y_1)$. Let $\xi(t, u) = x_1(t)y_1(ut)$ for $t, u \in [0,1]$; $\xi$ is a homotopy in $X$ from $x_1(t)$ to $x_2(t)$. This lifts to a homotopy $\Xi$ in $H$ from $y_1(t)$ to the path $y_2(t)$ which is the lift $x_3$. Since $\xi(1, t) = x_2(t) \Xi(1, t) = y_2(t)$, whence $y_{1}(1) = f$. This proves that $\mu_{*}[\pi(Y \times Y, \langle f,f \rangle)] \subseteq p_*[\pi(Y, f)]$, and since $Y \times Y$ is path connected theorem 10 applies and $\tilde{\mu}$ exists. That $p$ preserves multiplication is immediate. That $\tilde{\mu}$ is associative follows using theorem 8 on the two maps from $Y \times Y \times Y$ to $Y$, which are equal at $\langle f,f,f \rangle$. That $f$ is a multiplicative
identity follows similarly. Given \( y \in Y \), let \( \Theta \) be a path from \( f \) to \( y \) in \( Y \); let \( x = p(y) \), let \( \vartheta(t) = \theta(t)^{-1} \), and let \( \Theta' \) be the lift of \( \vartheta' \) starting at \( f \). As usual, \( \Theta'(t) \Theta(t) = f \), so \( y^{-1} = \Theta'(1) \).

In particular if \( X \) a topological group which is locally path connected, semilocally 1-connected, and connected then the universal covering space \( Y \) can be made into a group, called the universal covering group. For a connected covering group \( q : Z \rightarrow X \) (where \( p(e) = q(e) = e \)) the induced map \( f : Y \rightarrow Z \) (where \( f(e) = e \)) is a group homomorphism. This follows using theorem 12 because \( f \) is a covering projection.

If \( X \) is a connected Lie group then there is a natural analytic manifold structure induced on \( Y \) by the local homeomorphisms. For further properties of this simply connected covering Lie group see the literature.

**Exercises.**

1. For exercises 1 to 7, let \( F \) be a field complete with respect to an absolute value. in \( F^n \) let \( B_{wr} \) denote \( \{ x \in F^n : |x_i - w_i| < r_i \} \) for \( 1 \leq i \leq n \) where \( w \in F^n \) and \( r \in R^{>n} \); and similarly for let \( B_{gr}^{\leq} \) Show that if a multivariate power series \( \sum_{\nu} a_{\nu} x^\nu \) converges normally (see section 24.7) in \( B_{gr} \), then it converges uniformly (i.e., the partial sums do) in any \( B_{gr}^{s} \) with \( s < r \) (where \( s < r \) iff \( s_i < r_i \) for all \( i \)). Hint: Let \( M = \sum_{\nu} |a_{\nu}| s^\nu \); \( M \) bounds the sum throughout.

2. Show that if \( \sum_{\nu} a_{\nu} x^\nu \) converges in \( B_{wr} \), then the function it defines is continuous. Hint: Use exercise 1, and facts observed in section 24.6.

3. Suppose \( f : U \rightarrow V \) and \( g : V \rightarrow W \) where \( U \subseteq F^n \), \( V \subseteq F^m \), and \( W \subseteq F^l \) are open, and \( f, g \) are analytic. Show that \( g \circ f \) is analytic. Hint: Let \( F(\xi) \) be a power series for \( f(x + \xi) - y \) where \( y = f(x) \), and \( G(\eta) \) a power series for \( g(y + \eta) \), and w.l.g. let \( l = 1 \); then \( G \circ F \), provided it is well-behaved, is a power series for \( g(f(x + \xi)) \). Write \( f_\mu \) for the coefficients of \( F \), and \( g_\nu \) for the coefficients of \( G \). Write \( G \circ F \) as \( \sum_{\mu, \nu} g_\nu f_\mu t_\mu \), where \( \rho \) ranges over maps from \( J_\nu \) to \( \{ \mu \} \), \( J_\nu = \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq \nu_i \} \), and \( t_\mu \) is the product of the \( f_{\mu j} x^\mu \) selected by \( \rho \) (the selection for \( (i,j) \) is from the \( j \)th copy of \( (\sum_{\mu} f_{\mu j} x^\mu)^{\nu_i} \)). Suppose \( F \) (\( G \)) converges normally in \( B_{gr} \) (\( B_{gr} \)). Let \( \phi(\xi_1, \ldots, \xi_m) = \sum_{\mu} |f_\mu| \xi^\mu \), where \( \xi \) takes real values. Let \( \eta \in R^{>m} \) be such that \( \phi(\eta) < s \). For any finite set of terms of \( G \circ F \), replacing \( f \) by \( \phi \), and evaluating at \( \eta \), yields a bounded value. Thus, the sum is absolutely convergent, so \( G \circ F \) is, and converges to the desired value.

4. Show that the analytic functions \( f : U \rightarrow F^m \) where \( U \subseteq F^n \) form an \( F \)-algebra. Hint: Addition and scalar multiplication are straightforward. Multiplication follows by exercise 3; for a direct proof, define a product formal power series and show it is normally summable.

5. Show that if \( f : C^n \rightarrow C \) is a complex analytic function then its real and imaginary parts are real analytic functions in \( R^{2n} \). Hint: Show normal summability of the real series.

6. Show that if \( f : U \rightarrow F \) for an open \( U \subseteq F^n \) is analytic and nonzero then \( 1/f \) is analytic. Hint: Show that \( 1/y \) is analytic in a disc around \( y_0 \) for any \( y \neq y_0 \), and use exercise 3. Modify the power series \( 1/(1-x) = \sum_i x^i \).

7. Show that if \( f \) is given by \( \sum_{\mu} a_{\mu} \xi^\mu \) at \( x \in U \), then \( \partial f/\partial x_i \) is given by \( P = \sum_{\mu} a_{\mu+e_i}(\mu_i + 1)\xi^\mu \), where \( e_i \) is the unit vector which is 1 in the \( i \)th component. Hint: By exercise 3 and manipulation of formal power series, \( f(x + \xi + te_i) - f(x + \xi) = t(P + O(t)) \). The claim follows using exercise 2.

8. Show that if \( L \) is a Lie algebra, \( I \subseteq L \) is a solvable ideal, and \( L/I \) is solvable, then \( L \) is solvable. Hint: Suppose \( D_{s_k}(I) = 0 \) and \( D_{s_l}(L/I) = 0 \). If \( \eta \) is the canonical epimorphism then \( \eta[D_{s_l}(L)] = 0 \), so \( D_{s_l}(L) \subseteq I \), so \( D_{s_k+l}(L) = 0 \),
28. Some algorithms.

1. Gaussian elimination. It is well known that an entry of a matrix $M$ after some stages of Gaussian elimination is a ratio of determinants of minors of $M$. This may be proved quite simply as follows.

If $M$ is an $m \times n$ matrix, define a partial traversal to be a sequence $\tau = i_1,j_1, \ldots, i_k,j_k$ where $1 \leq i_1 \leq m$, $1 \leq j_1 \leq n$, the $i_a$ are distinct, and the $j_a$ are distinct. Such arises for example as the sequence of positions of the pivot points up to stage $k$ during Gaussian elimination. Let $M_\tau$ denote the minor $M_{ST}$ as in appendix 2, where $S = \{i_a\}$ and $T = \{j_a\}$.

Let $M_k$ denote $M$, after $k$ stages of Gaussian elimination, say where the pivot column is cleared in rows not yet containing a pivot position. Let $\tau, i, j$ denote an extension of $\tau$ by an additional position. Clearly, $\det(M_{k+1,\tau,i,j}) = \det(M_{k,\tau})_{ij} \det(M_{\tau,ij})$. Since $M_{k+1,\tau,i,j}$ and $M_{k,\tau}$ are obtained by elementary row operations, $(M_{k,\tau})_{ij} = \det(M_{\tau,ij})/\det(M_\tau)$.

To show that Gaussian elimination over $Q$ is a polynomial time algorithm, it remains to bound the size of the determinant in terms of the size of the matrix. It is easy to obtain a polynomial bound; with a bit more work a tighter bound can be obtained.

Suppose $v$ is a vector of nonnegative reals. A bound on $\log_2 |v|$ in terms of $\sum_i \log_2 v_i$ may be obtained using Lagrange multipliers. Indeed, at a stationary point of $\log_2 |v|^2$ subject to the constraint $\sum_i \log_2 v_i = c$, $\log_2 |v| \leq (\log n)/2 + c/n$.

If $M_{ij} = p_{ij}/q_{ij}$ where $p_{ij} \in Z$ and $q_{ij} \in Z^+$, let $q_i = \Pi_{j} q_{ij}$, $q = \Pi_q$, and $p = \Pi_{i} \min(|p_{ij}|,1)$. Clearing denominators and using Hadamard’s inequality (section 10.12) and the above fact, $\log_2 |\det(qM)| \leq \sum_i (\log_2 n/2 + (1/n) \sum_j (\log_2 q_i + \log_2 p_{ij}))$, whence $\log_2 |\det(qM)| \leq \log_2 q + \log_2 p/n + n \log n/2$.

2. Coset enumeration. This section and the next provide a brief introduction to two topics in computational group theory. A recent reference for this field, which has expanded greatly since its origins in the 1960’s (with some work in the 1950’s), is [Holt]. Extensive libraries of computational group theory routines are available on the World Wide Web.

Recall from section 12.9 that a finite presentation of a group $G$ consists if a finite set of a generators $A$, their inverses $A^{-1}$, and a finite set of equations $u = v$ between words in the alphabet $A \cup A^{-1}$. Even though the word problem for groups is unsolvable, it is of interest to try to determine facts about a group from a finite presentation. If the effort is successful then knowledge about the group has been gained. Such efforts were of value in the classification of the finite simple groups, for example.

Let $A^\pm$ denote $A \cup A^{-1}$. We will assume without loss of generality that the equations of the presentation are of the form $w = 1$, and let $R$ denote the set of $w$, which are called among other things the “relators” of the presentation. They generate a normal subgroup, and the group $G$ is the quotient of the free group by the normal subgroup.

Given a finite presentation, a subgroup $H \subseteq G$ may be specified by giving a set $S$ of words over $A^\pm$, which specify generators for $H$. Supposing the index of $H$ in $G$ to be finite, let $C$ be the set of cosets. The coset table is defined to be the function $C \times A^\pm \to C$ where $\lambda a = \mu$ (while either left or right cosets could be used, right cosets seem to be the standard convention).

Say that a partial coset table is a partial function $C \times A^\pm \to C$, where $C$ is a finite set, with $\iota \in C$ a distinguished element. We let $\lambda a$ denote the value, which is either some $\mu \in C$, or undefined. For $\lambda \in C$ and $w$ a word over $A^\pm$, let $\lambda w$ be obtained by evaluation of $\mu a$ for successive letters $a$ of $w$, with the result being undefined if any $\mu a$ is undefined. Also, for a word $w$ let $w^{-1}$ be the inverses of the letters of $w$, in reverse order.

The coset table of $H$ in a finitely presented group has the following properties, where $\iota$ denotes $H$.

1. It is correct, that is, $\lambda a$ is the correct coset.
2. It is connected, that is, for any $\lambda \in C$ there is a word $w$ with $\lambda = \iota w$. 
3. It is complete, that is, $\lambda a$ is always defined.
4. It is reduced, that is, if $\lambda \in C$ and $w \in R$ then $\lambda w = w$, and if $w \in S$ then $\lambda w = w$.

Given any partial coset table, say that it is correct if there is a map $\lambda \mapsto \lambda'$ from $C$ to the cosets such that $\lambda' = H$, and if $\lambda a = \mu$ then $\lambda' a = \mu'$. Note that if the table is connected the map is unique if it exists.

**Theorem 1.** If a partial coset table has properties 1-4 above, it is isomorphic to the coset table of $H$.

**Proof:** By connectedness $\lambda \mapsto \lambda'$ is surjective. For a word $w$ let $w_G$ denote its image in $G$. We claim that if $w_G = 1$ then $\lambda w = \lambda$ for all $\lambda \in C$. The claim is clear if $w \in R$, or inductively if $w = uw$ where $u_G = 1$ and $v_G = 1$. Suppose $w = xvx^{-1}$ where $v_G = 1$. Let $\lambda = x$. Then $\lambda w = \lambda vx^{-1}$, which inductively equals $\lambda x^{-1}$. Since in a reduced table $\lambda a^{-1} = \lambda$, $\lambda x^{-1} = \iota$. The claim is thus proved. Now suppose $\mu' = \nu'$; let $\mu = \nu w$ and $\nu = \nu w$. Then $(\nu w^{-1})_G H$, so $(\nu w^{-1})_G s_1 \ldots s_n G$ for some $s_1, \ldots, s_n \in S$, and by the claim, and the fact that $\nu s_1 \ldots s_n = \iota$, $\nu u = \nu w$, which shows that $\lambda \mapsto \lambda'$ is injective.

Coset enumeration, also called the Todd-Coxeter algorithm, is a nondeterministic strategy for attempting to determine the coset table. It is a fundamental algorithm of computational group theory. During coset enumeration a partial coset table is maintained, which represents the present state of knowledge of the coset table. Initially, $C = \{\iota\}$ where $\iota$ denotes $H$, and $\iota a$ is undefined for all $a$. Two possible operations may be performed.

1. Define. If $\lambda a$ is undefined, add a new coset $\mu$ to $C$, and set $\lambda a = \mu$; $\mu a$ is undefined for all $a$.
2. Reduce. Suppose that for some $\lambda \in C$ and $w \in R$, or $\lambda = \iota$ and $uw \in S$, $\lambda u = \mu$ and $\lambda v = \nu$ where $\mu \neq \nu$; then identify $\mu$ and $\nu$, and make any further identifications that follow.

In a reduction step, define $\equiv$ to be the least equivalence relation on $C$ such that $\mu \equiv \nu$; and such that if $\kappa \equiv \lambda$, $a \in A^\pm$, and $\kappa a$ and $\lambda a$ are both defined, then $\kappa a \equiv \lambda a$. The coset table on the equivalence classes is that where $\mu a = \nu a$ implies $[\mu]a = [\nu]$. Remaining details are left to the reader.

Note that any partial coset table produced by such operations is correct and connected. If we arrive at a complete and reduced table, by theorem 1 we have determined the coset table. On the other hand, if

1. the elements of $C$ are kept in order of the first created member of the equivalence class,
2. at each step the first $C$ with an undefined entry is chosen for an entry definition, and
3. all possible reductions are applied after a definition,

then if the index of $H$ in $G$ is finite a complete and reduced table will eventually be reached. Of course, there is no computable bound on how long this might take.

Some “shortcuts” may be taken in carrying out operations. First, when a new coset $\mu$ is defined with $\lambda a = \mu$, $\mu a^{-1}$ may be set to $\lambda$. Second, suppose $\mu a$ is undefined. Suppose that for some $\lambda \in C$ and $u a v \in R$, or $\lambda = \iota$ and $u a v \in S$, $\lambda u = \mu$ and $\lambda v^{-1} = \nu$; then set $\mu a = \nu$ and $\nu a^{-1} = \mu$.

For $\lambda \in C$ and $w \in R$, or $\lambda = \iota$ and $w \in S$, define the scan operation as follows.

1. Let $u$ be the longest prefix of $w$ such that $\lambda u$ is defined.
2. If $u = w$, if $\lambda u \neq \iota$ then condense $\lambda u$ and $\lambda$; otherwise do nothing.
3. Otherwise let $w'$ be $w$ with $u$ removed from the front, and let $v$ be the longest suffix of $w'$ such that $\lambda v^{-1}$ is defined.
4. If $v = w'$, if $\lambda u \neq \lambda v^{-1}$ then condense $\lambda u$ and $\lambda v^{-1}$; otherwise do nothing.
5. Otherwise if $w = u a v$ for some $a$, then define $\mu a = \nu$ and $\nu a^{-1} = \mu$, where $\lambda u = \mu$ and $\lambda v^{-1} = \nu$.
6. Otherwise do nothing.

Various strategies have been considered for the order in which to apply possible operations. New rows may be defined to ensure that scans have an effect (strategies of HLT type), or definitions based on current information may be made (strategies of Felsch type). Experiments suggest that mixed strategies can improve performance. See the literature for further discussion.
3. The Schreier-Sims algorithm. Algorithms for permutation groups are a branch of computational group theory. The Schreier-Sims algorithm is a basic one, developed in the late 1960’s. Recall from section 16.1 the definition of a series of subgroups \( G = G_0 \supset \cdots \supset G_n \). For \( G \) finite, such a series with \( G_n = 1 \), together with a system \( U_i \) of coset representatives of \( G_i \) in \( G_{i-1} \) for \( 1 \leq i \leq n \), yields considerable information about \( G \) (for example the order). The Schreier-Sims algorithm produces such a series for a finite permutation group \( G \), given by its generating permutations.

We first define Schreier generators for a subgroup of a group; these are used in the Schreier proof of the Schreier-Nelson theorem (see exercise 12.4). Suppose \( G \) is a group, \( X \) a set of generators, and \( H \subseteq G \) is a subgroup. Choose representatives \( U \) for the right cosets, and for \( g \in G \) let \( u_g \) be the element of \( U \) with \( g \in Hu_g \). Clearly, if \( u \in U \) and \( x \in X \) then \( uxu^{-1}_x \in H \); such an element is called a Schreier generator for \( H \).

**Theorem 2.** The Schreier generators for \( H \) generate \( H \).

**Proof:** First note that if \( v = u^{-1}_x \), then \( Hv = Hux^{-1} \), so \( Hux = Hu \), so \( u = u_v \); it follows that \( (ux^{-1}u^{-1}_v)^{-1} = vxu^{-1}_x \). Suppose \( h \in H \) equals \( y_1 \cdots y_l \) where \( y_i \) is either an element of \( X \) or the inverse thereof. For \( 0 \leq i \leq l \), let \( u_i = u_{y_1} \cdots u_{y_i} \). Then \( u_0 = 1 \) and \( u_1 = 1 \). Also, for \( i > 0 \), \( u_i = u_{u_{i-1}y_i} \), since \( y_1 \cdots y_{i-1} \) and \( u_{i-1} \) are in the same coset. Thus,

\[
h = u_0hu_i = \prod_{i \geq 1} (u_{i-1}y_iu^{-1}_i) = \prod_{i \geq 1} (u_{i-1}y_iu_{u_{i-1}y_i}^{-1}),
\]

proving the theorem.

From here on in this section let \( G \) be a finite permutation group acting on a finite set \( \Omega \), generated by the set \( S \) of permutations. We will write the group action on the right, so that \( \alpha g \) denotes the value when \( g \) is applied to \( \alpha \).

If \( \alpha \in \Omega \), and \( H = \text{Stab}_\alpha \) then by the correspondence of theorem 5.4, the Schreier generators may be written as \( u_\beta xu_\beta^{-1} \) where \( \beta \) runs over the orbit of \( \alpha \), and \( u_\beta \) is the representative of the corresponding coset.

Given \( \Omega \), \( S \), and \( \alpha \in \Omega \), a basic computational task is to compute the orbit \( \Delta = \alpha G \), and for each \( \beta \in \alpha G \), a \( u_\beta \in G \) such that \( \alpha u_\beta = \beta \). Note that \( U = \{ u_\beta \} \) is a system of coset representatives for \( \text{Stab}_\alpha \). This task may be accomplished by a straightforward “breadth-first search” as follows.

Set \( \Delta = \{ \alpha \} \), set \( u_\alpha = 1 \), and initialize a queue to contain \( \alpha \).

Until the queue is empty {

Remove \( \beta \) from the front of the queue.

For \( x \in S \)

if \( \beta x \) is not in \( \Delta \),

add \( \beta x \) to \( \Delta \) and the end of the queue, and set \( u_\beta x = u_\beta x \)

}

The Schreier-Sims algorithm makes use of a special type of series in a permutation group called a “stabilizer chain”. Suppose \( B = (\beta_1, \ldots, \beta_k) \) is a sequence of distinct elements of \( \Omega \). We introduce the commonly used notation \( G_{\beta_1, \ldots, \beta_k} \) for the pointwise stabilizer in \( G \) of \( \beta_1, \ldots, \beta_k \). For \( 1 \leq i \leq k + 1 \) let

\[
G^i = G_{\beta_1, \ldots, \beta_{i-1}}, \quad S^i = S \cap G^i, \quad H^i = [S^i], \quad \Delta^i = \beta_i H^i, \quad \text{and } U^i = \{ u^i_\beta : \beta \in \Delta^i \}
\]

where \( [S^i] \) denotes the subgroup generated by \( S^i \). Note that \( G^1 = G \); this is standard in the literature (rather than \( G^0 = G \)).

The groups \( G_i \) form a stabilizer chain. If \( G_{k+1} = \{ 1 \} \) is said to be a base. In general \( H_i \subseteq G_i \); however if \( S \) is “fat” enough, then equality will hold. In this case \( S \) is said to be a strong generating set. A
“base and strong generating set” (abbreviated BSGS) facilitates many computations concerning $G$; see the literature for further discussion. The Schreier-Sims algorithm is an algorithm for computing such.

The following procedure, called “stripping” the element $g$ with respect to a stabilizer chain and associated data as above, is used in the Schreier-Sims algorithm. It iteratively translates $g$ into the next stabilizer, as long as $\beta_i$ is in the orbit.

Set $h = g$.

For $i$ from 1 to $k$ {
  If $\beta_i h \notin \Delta^i$ return $h, i$.
  Let $u_i \in U^i$ be such that $\beta_i u = \beta h$, and set $h = hu_i^{-1}$.
}

return $h, k + 1$.

It is readily verified that a stabilizer chain is a BSGS iff $H_i^i = H_i^{i+1}$ for all $i \leq k$, and $H_k^{k+1} = \{1\}$ (use $G_{\beta_i}^i = G_i^{i+1}$, and induction for the converse). The Schreier-Sims algorithm ensures that the Schreier generators for $H_i^i$, as a subgroup of $H_i$, are in $H_i^{i+1}$.

To start the Schreier-Sims algorithm, find a sequence $B$ such that no element of $S$ fixes every element of $B$; this ensures $H_k^{k+1} = \{1\}$. This may be done by considering each element $x \in S$ in turn; if $x$ fixes every element added to $B$ so far, add an element $\beta \in \Omega$ to $B$ which $x$ moves. Next compute the stabilizer chain data as above. Finally, set $i = k$ and repeat the following loop body until $i = 0$.

For $\beta \in \Delta^i$ for $x \in S^i$ {
  Let $z = uxu^{-1}_{\beta}$ where $u = u_{\beta}^i$ be the Schreier generator in $H_i^i$ for $\beta$ and $x$.
  If $u_{\beta x}$ equals $ux$ by definition, continue.
  Strip $z$, with result $h, j$.
  If $h = 1$ continue.
  If $j = k + 1$ increment $k$, set $\beta_k$ to some element of $\Omega$ moved by $h$,
  and set $S^k$ to empty.
  For every $l$ from $i + 1$ to $j$, add $h$ to $S^l$ and update $\Delta^l$.
  Set $i = j$.
}

If all generators were in $H_i^{i+1}$ decrement $i$.

The algorithm runs in polynomial time, because if a change is made due to level $i$, an element of $\Omega$ is added to $\Delta^i$, so at most $|\Omega|^2$ changes are made. The loop body runs in polynomial time.

There are various implementation details and variations of the Schreier-Sims algorithm of interest, in particular for large $\Omega$, or for matrix groups. The storage and retrieval of the coset representatives is an important issue. Coset enumeration and / or randomization are used in variations. Algorithms for computing a BSGS which give a theoretical bound on the running time have been given. Again see the literature for further discussion.

4. Decision procedure for Presburger arithmetic. Consider $\mathcal{N}$ as a structure for the first order language $0, \text{Suc}, +, = \text{ (where Suc}(x) = x + 1)$, that is, without multiplication. Let $P$ be the theory of $\mathcal{N}$ in this language. This theory is called Presburger arithmetic, because Presburger proved in 1929 that it is decidable.

Presburger’s proof used the method of quantifier elimination, wherein an explicit method is given for translating an arbitrary formula to an equivalent open formula.

Theorem 3. Suppose $T$ is a first order theory in the language $L$, and for any open formula $F(x, y)$ of $L$, there is an open formula $F'(y)$ of $L$ such that $\models_T \exists x F \equiv F'$. Then for any formula $F$ of $L$ there is a open formula $F'$ of $L$ with the same free variables other than $x$, such that $\models_T \exists x F \equiv F'$. 

403
Proof: The proof is a straightforward induction on formulas.

If the removal of the existential quantifier is effective then the translation of any formula is. In particular, if it decidable whether a open sentence is true then it is decidable if any sentence is. In applications of this theorem \( L \) may be expanded from the original language; this occurs in quantifier elimination for Presburger arithmetic for example.

For the variation of Presburger’s procedure given in [Cooper], the domain is \( \mathbb{Z} \), and the language has the integer constants, +, and <. The language is expanded to contain \( \neg, n \cdot x \) for each integer \( n \), the divisibility predicate \( n|x \) for each positive integer \( n \), and its negation \( n \nmid x \). Clearly the open sentences are decidable.

Given a formula \( \exists x F \), the following steps may be applied to \( F \).
1. Move negations to atomic formulas, and eliminate them by replacing \( \neg t < u \) by \( u < t + 1 \), \( \neg n|u \) by \( n \nmid u \), and \( \neg n \nmid u \) by \( n|u \).
2. Transform each atomic formula involving \( x \) to one of \( mx < u \), \( u < mx \), \( n|mx + u \), or \( n \nmid mx + u \), where \( m \) is a positive constants and \( u \) does not involve \( x \).
3. Let \( M \) be the least common multiple of the constants \( m \) of step 2. Multiply both sides of all atomic formulas involving \( x \) by the necessary amount to replace \( m \) by \( M \).
4. Replace \( Mx \) by \( x \), and add the conjunct \( M|x \).

Assume \( F \) is as in the result of these steps, so that \( F \) is positive and all atomic formulas are \( x < u \), \( u < x \), \( n|x + u \), or \( n \nmid x + u \). Let \( N \) be the least common multiple of the \( n \). Let \( F_{-\infty} \) be \( F \), with atomic formulas \( x < u \) replaced by “true” and atomic formulas \( k < x \) replaced by “false”. Let \( S \) be the values \( k \) occurring in atomic formulas \( k < x \).

Theorem 4. With notation as above, \( \exists x F \) is equivalent to

\[
\bigvee_{j=1}^{N} F_{-\infty}(j) \lor \bigvee_{i=1}^{N} \bigvee_{k \in S} F(k + j).
\]

Proof: Exercise 1.

Let \( \text{PresA} \) denote the system of axioms obtained by deleting the axioms concerning \( x \cdot y \) from the system \( \text{PA} \) of section 12.6 (i.e., axioms 6 and 7 of \( Q \)). The elimination procedure just described is “\( \text{PresA} \) verifiable”, that is, every equivalence can be proved in \( \text{PresA} \), provided the formulas are interpreted in the language \( 0, \text{Suc}, +, = \). The proof is tedious, and omitted. As a result, \( \text{PresA} \) is complete, and its consequences are Presburger arithmetic.

There is another decision procedure for Presburger arithmetic, given in [FerRack]. We give a brief outline, and refer the reader to the reference for the complete proof. The language is \( 0, \leq, +_R \) where \( +_R \) is the relation \( x + y = z \).

Let \( \Phi_{n,k} \) denote the prenex normal form formulas with at most \( n \) quantifiers and free variables among \( x_1, \ldots, x_k \). Let \( \equiv_{n,k} \) denote the equivalence relation on \( \mathbb{Z}^k \) which holds at \( a \) and \( b \) iff \( F(a) \equiv F(b) \) for all \( F \in \Phi_{n,k} \).

For a set \( V \subseteq \mathbb{Z} \) let \( \text{lcm}(V) \) denote the positive least common multiple of the nonzero elements of \( V \). Let \( V_0 = \{-2, -1, 0, 1, 2\} \). For \( i \geq 0 \) let \( \delta = \text{lcm}(V_i), V'_i = \{\delta/v' : v', v' \in V_i, v \neq 0\} \), and \( V_{i+1} = V'_i + V'_i \). For each \( n \) and \( k \) let \( \delta = \text{lcm}(V_n) \), and define the equivalence relation \( E_{n,k} \) on \( \mathbb{Z}^k \) to hold at \( a \) and \( b \) iff for every \( v_1, \ldots, v_k \in V_n \) and every \( v \) with \( |v| \leq \delta^2 \),

\[
v + \sum_{i=1}^{k} v_i a_i \leq 0 \text{ iff } v + \sum_{i=1}^{k} v_i b_i \leq 0, \text{ and } \\
\delta^2 |a_i - b_i| \text{ for } 1 \leq i \leq k.
\]
It is shown that
1. if \( aE_{0,k}b \) then \( a \equiv_{0,k} b \); and
2. if \( aE_{n+1,k}b \) where \( |b_i| \leq m \) for all \( i \), and \( a_{k+1} \in \mathbb{Z} \), then there is a \( b_{k+1} \) with \( |b_{k+1}| \leq B(n,k,m) \) such that \( aE_{n+1,k+1}b' \).

Here \( a' \) is \( a \) with \( a_{k+1} \) appended, and similarly for \( b' \); and \( B(n,k,m) = (m+1)2^{2^c(n+1)} \) for a suitable constant \( c \). The proof is about four pages long.

It is also shown that given relations satisfying 1 and 2 above,
3. if \( aE_{n,k}b \) then \( a \equiv_{n,k} b \); and
4. if \( \exists x F(x,a) \) then \( \exists x (|x| \leq B(n,k,m) \land F(x,a)) \).

Finally, if requirement 4 holds, and \( m_0, \ldots, m_a \) is a nondecreasing sequence of positive integers with \( H(n+k-i, i-1, m_{i-1}) \leq m_i \) for \( 1 \leq i \leq m \), then
5. \( Q_1x_1 \ldots Q_kx_kF \) is equivalent to \( Q_1x_1 \ldots Q_kx_k(|x_1| \leq m_1 \land \cdots \land |x_k| \leq m_k \land F) \).

where \( F \in \Phi_{n,k} \) and \( Q_i \) is an existential or universal quantifier.

We have stated the above facts only for the structure \( \mathcal{Z} \). Some are proved in greater generality, involving pairs of structures from some class, reflecting the fact that they are related to the theory of Ehrenfeucht-Fraissé games.

The last fact may be applied for the above \( B \) with \( m_i = 2^{2^cn+i} \). It follows that the space complexity of deciding a formula of Presburger arithmetic is \( 2^{2^cn} \) for some \( c \). In [FischRab] it is shown that the time complexity of deciding a formula of Presburger arithmetic is bounded below by \( 2^{2^cn} \) for some \( c \). The methods of [Berman] may be used to reduce the gap between the bounds.

5. Basic algorithms for polynomials. Recall from section 7.1 that the division law for \( F[x] \) where \( F \) is a field states that given nonzero \( p, d \in F[x] \), there are unique polynomials \( q \) and \( r \) such that \( p = qd + r \), and \( \deg(r) < \deg(d) \). We note for later use that \( q \) and \( r \) can be computed in \( O(n)O(m' n) \) field operations, where \( n = \deg(d) \), \( m = \deg(p) \), and \( i' = j \) if \( i > j \), else 1.

Also, the product \( pq \) of two nonzero polynomials can be computed in \( O(\deg(p) \deg(q)) \) field operations. More sophisticated algorithms can be given to “asymptotically” reduce the number of field operations for multiplication and division of polynomials [AHU]. The bound for the “faster” algorithms eventually grows more slowly; however the constant factor may be larger, so that degree where the faster algorithm actually is faster may be too large for the algorithm to be practical. If one is merely concerned with demonstrating that a computation can be carried out in polynomially many steps, the faster algorithm is irrelevant.

Euclid’s algorithm (section 6.3) for computing the gcd of two nonzero polynomials \( p_1, p_2 \in F[x] \) may be given as follows. Let \( n_i \) denote \( \deg(p_i) \), and assume \( n_1 \geq n_2 \).

Set \( i = 2 \), and repeat the following.
1. If \( p_i = 0 \) the algorithm has completed; the gcd is \( p_i \).
2. Write \( p_{i-1} = qp_i + r \) where \( \deg(r) < \deg(p_i) \).
3. Set \( p_{i+1} = r \), \( i = i + 1 \).

Euclid’s algorithm can be performed in \( \sum_i O(n_i(n_{i-1} - n_i)) \) field operations, which is \( O(n_1n_2) \).

There is a division law for \( R[x] \) where \( R \) is any commutative ring, as follows. Suppose \( p, d \in R[x] \) are nonzero, \( n = \deg(d) \), \( m = \deg(p) \), \( m \geq n \), \( d \) has leading coefficient \( d_n \), and \( \mu = d_m^{m-n+1} \). Then there are polynomials \( q \) and \( r \) such that \( \mu p = qd + r \) and \( \deg(r) < n \). This follows by induction on \( m \). If \( R \) is an integral domain then \( q \) and \( r \) are unique, by the usual argument.

Unless otherwise specified, for the rest of this section \( R \) is a factorial domain. When a sequence of \( p_i \in R[x] \) is being considered let \( n_i \) denote \( \deg(p_i) \) and let \( \pi_i \) denote the leading coefficient of \( p_i \). Note that
\[
gcd(p_1, p_2) = \gcd(c_1, c_2) \gcd(\hat{p}_1, \hat{p}_2)
\]
where for \( i = 1, 2 \), \( c_i \) is the content of \( p_i \) and \( \hat{p}_i \) is the primitive part.

Euclid’s algorithm may given for \( R[x] \), as follows. For \( i \geq 2 \) let \( \mu_i = \pi_i^{n_i-1-n_i+1} \).

For \( i = 1, 2 \) compute \( c_i \) and \( \hat{p}_i \), and replace \( p_i \) by \( \hat{p}_i \). Set \( i = 2 \), and repeat the following.
1. If \( p_i = 0 \) the algorithm has completed; the gcd is \( \gcd(c_1, c_2) \cdot p_{i-1} \).
2. Write \( \mu_i p_{i-1} = qp_i + r \) where \( \deg(r) < \deg(p_i) \).
3. Set \( p_{i+1} = \hat{r} \), \( i = i + 1 \).

Replacing \( r \) by its primitive part \( \hat{r} \) in step 3 does not change the gcd, because \( p_i \) is primitive, so any common divisor of \( p_i \) and \( r \) is primitive, so is a divisor of \( \hat{r} \).

The case where \( R = \mathbb{Z}[w] \) has become of considerable interest due to computing in such rings. Variations of this algorithm have been devised (see [Knuth], [BrTr]), to reduce the amount of computation involved in using \( \hat{r} \). For example in Collins’ reduced method, rather than \( r/c \) where \( c \) is the content, \( p_{i+1} \) is set to \( r/\mu_{i-1} \), where \( \mu_1 = 1 \). In step 1 the gcd is \( \gcd(c_1, c_2) \cdot \hat{p}_{i-1} \). Authors report that this method yields considerably improved performance even for \( R = \mathbb{Z}[w] \), and is of interest even for \( R = \mathbb{Z} \).

It remains to be shown that Collins’ reduced method yields a sequence of polynomials in \( R[x] \). One proof of this is facilitated by introducing a definition which has found numerous applications in the theory of computing with polynomials, including other improved gcd algorithms (see [BrTr]).

Suppose \( p_1, p_2 \in R[x] \), and \( n_1 \geq n_2 > 0 \). Let \( M \) be the \((n_2 + n_1) \times (n_2 + n_1)\) Sylvester matrix, defined in section 7.3. For \( 0 \leq i < n_2 \) let \( M_i \) be the matrix whose rows are rows 1 to \( n_2 - i \), and rows \( n_2 + 1 \) to \( n_1 + n_2 - i \), of \( M \). For \( 0 \leq j \leq i \) let \( M_{ij} \) be the matrix whose columns are columns 1 to \( n_1 + n_2 - 2i - 1 \), and column \( n_1 + n_2 - i - j \), of \( M_i \). The polynomial \( \sum_j \det(M_{ij})x^j \) is called the \( i \)th subresultant of \( p_1 \) and \( p_2 \). We denote it as \( S_i(p_1, p_2) \); note that \( S_0 \) is just the resultant.

Suppose \( R \) is a field, and in \( R[x] \), \( p_1 = qp_2 + p_3 \). By row operations Sylvester’s matrix \( M \) may be transformed so that the top \( n_2 \) rows are \( p_3 \); call this \( M' \). By left multiplying \( M' \) by \[
\begin{bmatrix}
0 & I \\
I & 0
\end{bmatrix}
\]
where the top \( I \) is \( n_1 \times n_1 \) and the bottom \( I \) is \( n_2 \times n_2 \), the top \( n_2 \) rows are moved to the bottom; call the result \( M'' \). Removing the top \( n_1 - n_3 \) rows and the left \( n_1 - n_3 \) columns leaves Sylvester’s matrix for \( p_2 \) and \( p_3 \). It follows that \( S_0(p_1, p_2) = (-1)^{n_1 n_2} x_2^{n_1 - n_3} S_0(p_2, p_3) \), since \( M'' \) is upper triangular in the top left \((n_1 - n_3) \times (n_1 - n_3)\) corner.

**Lemma 5.** Suppose \( R \) is a field, and \( p_1 = qp_2 + p_3 \), where \( n_1 \geq n_2 > n_3 \) and \( p_3 \neq 0 \).

a. For \( 0 \leq i < n_3 \), \( S_i(p_1, p_2) = (-1)^{(n_1-i)(n_2-i)} x_2^{n_1 - n_3} S_i(p_2, p_3) \)
b. \( S_{n_3}(p_1, p_2) = (-1)^{(n_1-n_3)(n_2-n_3)} x_2^{n_1 - n_3} x_3^{n_2 - n_3 - 1} p_3 \)
c. For \( n_3 < i < n_2 - 1 \), \( S_i(p_1, p_2) = 0 \)
d. \( S_{n_2-1}(p_1, p_2) = (-1)^{n_1-n_2+1} x_2^{n_1 - n_2 + 1} p_3 \)

**Proof:** The argument is a generalization of the argument for \( i = 0 \) given above. Let \( N \) be the copy of Sylvester’s matrix for \( p_2 \) and \( p_3 \) in the lower right corner of \( M'' \). Then \( N_i \) for \( 0 \leq i < n_3 \) is obtained from \( N \) by deleting the rows of \( M'' \), which correspond to the rows deleted from \( M \) to obtain \( M_i \). Part a follows. For part b, \( N_i \) becomes \( n_2 - n_3 \) rows of \( p_3 \); this yields the last two factors. For part c, column \( n_1 - i + 1 \) is 0 in the last \( n_2 - i \) rows of \( M''_i \), is to the right of the first \( n_1 - i \) columns, and is in the first \( n_1 + n_2 - 2i - 1 \) columns. For part d, the first factor comes from moving the top row of \( M''_i \) to the bottom, the second factor comes from the upper triangular matrix above the bottom row of \( M''_i \) and the third factor comes from the bottom row.

In a generalized form of Euclid’s algorithm,

\[
\alpha_i p_{i-2} = q_i p_{i-1} + \beta_i p_i \quad \text{for } i \geq 3
\]
where $\alpha_i, \beta_i \in R$, $q_i \in R[x]$, $n_2 \leq n_1$, and $n_i < n_{i-1}$ for $i \geq 3$. Lemma 5 holds in modified form; if $\alpha p_1 = q p_2 + \beta p_1$, the left side is multiplied by by $\alpha^t$, and the right by $\beta^t$, where $t$ is $n_2 - j$, $n_2 - n_3$, and 1 in cases a, b, and d respectively.

**Theorem 6.** Suppose $p_i$ is a sequence of polynomials as as above, with $p_k$ the last nonzero $p_i$. Suppose $3 \leq i \leq k$. For $j < n_{i-1}$ let

$$
\eta_{ij} = \Pi_{l=3}^{n_{i-1} - j} \alpha_l^{n_i - 1 - j},
$$

$$
\zeta_{ij} = \Pi_{l=3}^{n_{i-1} - j} (1 - 1)(n_{i-2} - j)(n_{i-1} - j) \alpha_{n_{i-1} - n_i}.
$$

Suppose $0 \leq j < n_2$.

a. If $j = n_i$ for $3 \leq i \leq k$ then $\eta_{ji, n_j} S_j(p_1, p_2) = \pi_{i, n_i} \alpha_{n_i - 1} - 1 \zeta_{i, n_i} p_i$.

b. If $j = n_i - 1$ for $3 \leq i \leq k$ then $\eta_{ji, n_i - 1} S_j(p_1, p_2) = \zeta_{i, n_i - 1} - 1 p_i$.

c. $S_j(p_1, p_2) = 0$ otherwise.

**Proof:** By lemma 5a and remarks above, $\eta_{ij} S_j(p_1, p_2) = \zeta_{ij} S_j(p_i - 1, p_i)$. For part a, use this equality for $i - 1$ and lemma 5b. Part b follows similarly using lemma 5d. Part c follows for $j > n_k$ using lemma 5c. Part c follows for $j < n_k$ because if $\alpha p_1 = \beta p_2$ and $j < n_2$ then $S_j(p_1, p_2) = 0$. This follows because over the field of fractions, the top $n_2$ rows of the Sylvester matrix can be set to zero using row operations.

**Corollary 7.** In Collins’ reduced method, $p_i = \rho_i S_{n_i - 1}(p_1, p_2)$ for $3 \leq i \leq k$, where $\rho_i \in R$. In fact,

$$
\rho_i = (-1)^{n_i - 2 - n_i - 1 + 1} \Pi_{l=3}^{n_i - 1} (-1)^{(n_{i-1} - n_{i-1} + 1)}(n_{i-2} - n_{i-1} + 1) \pi_{i, n_i}^{n_i - 1} (n_{i-1} - n_i - 1)
$$

**Proof:** In this case, $\alpha_i = \alpha_i^{n_i - 2 - n_i - 1 + 1}$ for $i \geq 3$, $\beta_3 = 1$, and $\beta_i = \alpha_i - 1$ for $i \geq 4$. Substituting in to theorem 6b and writing $e$ for $n_{i-2} - n_{i-1} + 1$, $\pi_{i-1}$ occurs to the power $e(n_{i-1} - n_{i-1} + 1)$ in the left side, $e(1 + n_i - n_{i-1} + 1)$ in the right side when $l < i$, and $e$ in the right side when $l = i$. The corollary follows.

Either version of Euclid’s algorithm for factorial domains requires $O(n_1 n_2)$ ring operations. In the case $R = \mathbb{Z}$ corollary 7 can be used to bound the size of the integers occurring in Collins’ reduced method. Let $L$ denote the maximum length of a coefficient in $p_1$ or $p_2$. By Hadamard’s inequality, and the fact that $n_1 \geq n_2$, any coefficient of a subresultant has length $O(n_1 (L + \log n))$. A factor $\rho_i$ has length bounded by $O(n_i^2 L)$. Thus, the overall bound is $O(n_i^2 L)$.

The Collins’ reduced gcd algorithm can be modified to compute the resultant of two polynomials, as follows.

For $i = 1, 2$ compute $c_i$ and $\tilde{p}_i$, and replace $p_i$ by $\tilde{p}_i$. Set $i = 2$, and execute the following recursive procedure.

1. If $p_1 = 0$ the resultant is $0$.
2. If $n_1 = 0$ the resultant is $\pi_1^{n_1 - 1}$.
3. Set $i = i + 1$.
4. Write $\alpha_i p_i = q p_i - 1 + \beta_i p_1$ where $\deg(r) < \deg(p_1)$.
5. Compute $S_0(p_{i-1}, p_i)$.
6. Solve $\alpha_i^{n_{i-1}} S_0(p_{i-2}, p_{i-1}) = \beta_i^{n_{i-1}} (-1)^{n_{i-2} - n_{i-1} + 1} \pi_{i-1}^{n_{i-2} - n_{i-1}} S_0(p_{i-1}, p_i)$.
7. Set $i = i - 1$.

Multiply the final result by $c_1^{n_2} c_2^{n_1}$. The bounds for the number of ring operations and integer length are as for the gcd algorithm.

A polynomial $p \in R(x)$ has a unique factorization $\Pi_i r_i^{a_i}$. The derivative $p'$ equals $\sum_i a_i r_i' \Pi_{j \neq i} r_j^{a_j}$. $\Pi_i r_i^{a_i - 1}$ divides both $p$ and $p'$. If $R$ has characteristic 0 then the highest power of $r_i$ dividing $p'$ is $r_i^{a_i - 1}$,
may be computed using the extended Euclidean algorithm. Further, deg(\(PQ\)) may be required.

The following lemma is one of several related facts which are called Hensel’s lemma. As will be seen, it has a computational aspect. Other versions of Hensel’s lemma may be found in \[\text{[Eisenbud]}\] and \[\text{[Jacobson]}\].

**Lemma 8.** Suppose \(R\) is a commutative ring, and \(P,Q \in R\) with \(P|Q\). In \(R[x]\), if \(f \equiv gh \mod Q\) and \(ag + bh \equiv 1 \mod P\) then there are \(g', h'\) such that \(g' \equiv g \mod Q\), \(h' \equiv h \mod Q\), \(f \equiv g'h' \mod PQ\), and \(ag' + bh' \equiv 1 \mod P\).

**Proof:** Let \(e = f - gh\), \(g' = g + be\), and \(h' = h + ae\). Then \(f - gh' = e(1 - bh - ag) - abe^2\), which is divisible by \(PQ\). Also, \(ag' + bh' - 1 = ag + bh - 1 - 2abe\), which is divisible by \(P\).

If \(P|Q\) and \(d|g \mod Q\) then \(d|g \mod P\); thus \(ag + bh \equiv 1 \mod P\) implies \(g\) and \(h\) are relatively prime mod \(Q\) whenever \(P|Q\). So the lemma “lifts” a factorization of \(f\) into relatively prime factors mod \(Q\) to such mod \(PQ\). If \(R\) is a factorial domain and \(P\) is prime then when \(g\) and \(h\) are relatively prime \(a\) and \(b\) exist, indeed may be computed using the extended Euclidean algorithm. Further, \(\text{deg}(a) < \text{deg}(h)\) and \(\text{deg}(b) < \text{deg}(g)\) may be required.

Suppose \(g' = g + Qt\) and \(h' = h + Qs\) for some \(s,t \in R[x]\); then \(g'\) and \(h'\) satisfy the conclusions of the lemma iff \(e_2 \equiv sg + th \mod P\), where \(e_2 = (f - gh)/Q\). This holds if \(s \equiv a\bar{e}_2 + qh \mod P\) and \(t \equiv be_2 - qg \mod P\). If \(R/(PR)\) is a factorial domain then the latter must hold; indeed, writing \(f\) for \(mod P\), \((\bar{s} - a\bar{e}_2)g + (\bar{t} - b\bar{e}_2)h = 0\), so \(\bar{h}(\bar{s} - a\bar{e}_2)g\), so \(\bar{h}(\bar{s} - a\bar{e}_2)\), and similarly \(g(\bar{t} - b\bar{e}_2)\).

We return to requiring that \(R\) be a factorial domain. Assume (1) the hypotheses of the lemma, and (2) that the leading coefficient of \(g\) is a unit mod \(P\). Then \(q\) may be found using division of polynomials so that \(\text{deg}(\bar{b}\bar{e}_2 - \bar{qg}) < \text{deg}(\bar{g})\). It follows that \(t\) may be chosen so that \(\text{deg}(t) < \text{deg}(g)\), whence \(g'\) may be required to have the same degree and leading coefficient as \(g\). If in fact \(P\) is prime then \(t\) is unique, whence \(\bar{s}\) is unique, whence \(g'\) and \(h'\) are unique, mod \(PQ\).

Suppose that in addition to (1) and (2), (3) \(\text{deg}(f) = \text{deg}(gh)\) is required; as above, choose \(t\) with \(\text{deg}(t) < \text{deg}(g)\). Since \(\text{deg}(e_2) \leq \text{deg}(g) + \text{deg}(h)\), if \(\text{deg}(s) > \text{deg}(h)\) then \(\text{deg}(gs) > \text{deg}(gh) \geq \text{deg}(e_2)\).

It follows that the leading coefficient of \(gs\) is divisible by \(P\) (because \(e_2 \equiv sg + th \mod P\)), which is a contradiction if \(s\) is chosen to have the same degree as \(\bar{s}\). Thus, \(\text{deg}(s) \leq \text{deg}(h)\), so \(\text{deg}(h') \leq \text{deg}(h)\); using \(\text{deg}(g') = \text{deg}(g)\) it follows that \(\text{deg}(h') = \text{deg}(h)\).

Suppose \(aq + bh \equiv 1 \mod P\), the leading coefficient of \(g\) is a unit mod \(P\). Then \(aq + bh = 1 + Pl\) for some \(l \in R[x]\). Let \(q\) be such that \(\text{deg}(\bar{l}b - \bar{qg}) < \text{deg}(\bar{g})\). Let \(\bar{b}_2 = \bar{l}b - \bar{qg}\), where \(\bar{b}_2\) is the same degree as \(\bar{b}_2\). Let \(\bar{a}_2 = \bar{l}\bar{a} + \bar{q}\bar{h}\), where \(\bar{a}_2\) is the same degree as \(\bar{a}_2\). Then \(\bar{a}_2g + \bar{b}_2h = l + Pm\) for some \(m \in R[x]\). Let \(a' = a - \bar{P}\bar{a}_2\) and \(b' = b - \bar{P}\bar{b}_2\). Then \(a'g + b'h = (1 + Pl) - P(l + Pm)\), so \(a'g + b'h \equiv 1 \mod P^2\).

If in the foregoing, \(\text{deg}(a) < \text{deg}(h)\) and \(\text{deg}(b) < \text{deg}(g)\) is also required, \(\text{deg}(b') < \text{deg}(g)\) follows, whence, by an argument similar to one above, \(\text{deg}(a') < \text{deg}(h)\) also.

Letting \(Q = P\), polynomials \(g', h', a', b'\) may be obtained with \(g' \equiv g \mod P\), \(h' \equiv h \mod P\), \(f \equiv g'h'\mod P^2\), and \(a'g' + b'h' \equiv 1 \mod P^2\). These polynomials are commonly called a “quadratic Hensel lift” of \(g,h,a,b\). Starting with a prime \(P\), one may lift the factorization to \(P^2, P^4, \ldots\), rather than the slower \(P^2, P^4, \ldots; P^k\) for any \(k\) may be obtained in \(O(\log k)\) lift steps.

Next we prove some inequalities of use in analyzing algorithms involving polynomials. Suppose \(f \in C[x]\); let \(n = \text{deg}(f)\), and let \(r_i \in C\) for \(1 \leq i \leq n\) be the roots. Let \(|f| = \sqrt{f_0^2 + \cdots + f_n^2}\) denote the Euclidean norm of the vector of coefficients.
We claim that for any complex number \( r \), \(|(x - r)f| = |(\bar{r}x - 1)f|\), where \( \bar{r} \) is the complex conjugate. Indeed,

\[
|(x - r)f|^2 = |f|^2 - \sum (f_{i-1} \bar{r} \bar{f}_i + \bar{f}_i r f_i) + |rf|^2
\]

\[
= |\bar{r}f|^2 - \sum (\bar{r} f_{i-1} \bar{f}_i + r f_{i-1} f_i) + |f|^2 = |(\bar{r}x - 1)f|^2
\]

(as usual, if \( i \) is out of range \( f_i \) is taken to be 0).

Let \( S_\geq = \{ i : |r_i| > 1 \} \), and \( S_\leq = \{ i : |r_i| \leq 1 \} \). Let \( M = \prod_{i \in S_\geq} |r_i| \). We claim that \( M \leq |f|/|f_n| \).

Indeed, let \( \tilde{f} = f_n \prod_{i \in S_\geq} (\bar{r}_i x - 1) \prod_{i \in S_\leq} (x - r_i) \). By the previous observation, \(|\tilde{f}| = |f| \).

Thus,

\[
|f_n M| = |f_n \prod_{i \in S_\geq} r_i| = |f_n \prod_{i \in S_\geq} \bar{r}_i| = |\tilde{f}_n| \leq |\tilde{f}| = |f|.
\]

Suppose \( f, g \in \mathbb{Z}[x] \) and \( g/f \); let \( m = \deg(g) \), let \( s_i \in \mathbb{C} \) for \( 1 \leq i \leq n \) be the roots, and \( g_j \) be the coefficient of \( x^j \). From section 7.3, \( g_i = (-1)^{m-i} \sigma_{m-i}^{\mathbb{Z}} (s_1, \ldots , s_m) g_m \). Since the roots of \( g \) are roots of \( f \), it follows that \( |g_i|/|g_m| \leq (\binom{m}{i}) M \leq (\binom{m}{i}) |f|/|f_n| \). Also, \( g_m |f_n \), whence \( g_m \leq f_n \); hence \( |g_i| \leq (\binom{m}{i}) |f| \). In exercise 4 it is shown that \( \sum_{i=0}^{m} (\binom{m}{i})^2 = (2^m) \). It follows that \( |g| \leq (2^m)^{1/2} |f| \), a fact we will refer to as Mignotte’s theorem.

The upper bound \( (2^m)^{1/2} \leq 4^m / \sqrt{\pi m} \) may be proved using Stirling’s approximation for \( n! \), q.v. see [Apostol]. That \( (2^m)^{1/2} \leq 4^m \) follows by the binomial theorem.

6. Completeness of real closed fields. An ordered field is a field which is an ordered commutative ring; see section 6.5. In addition to the properties given there, one readily verifies that

- if \( x \) is positive then \( 1/x \) is positive, and
- if \( 0 < x < y \) then \( 0 < 1/x < 1/y \).

An ordered field \( F \) is said to be real closed if it satisfies the axioms

1. every positive \( x \) has a square root; and
2. for odd \( n \), every polynomial of degree \( n \) with coefficients in \( F \) has a root in \( F \).

The real numbers are real closed; other examples include the real numbers which are algebraic over \( \mathbb{Q} \).

Tarski proved in 1948 that the theory of real closed fields was complete, and hence equal to the theory of \( R \) in the language of ordered fields. The theorem may be proved by quantifier elimination, and a decision procedure for the sentences is provided as well. The theorem may alternatively be proved by model-theoretic methods; see [CK].

We begin with some basic facts about real closed fields. Note that \( x \) is positive iff \( x \) is a square, so the order is determined. The characteristic is 0, and the rationals as an ordered subfield have their usual order.

Lemma 9. If \( F \) is a real closed field then \( F[\sqrt{-1}] \) is algebraically closed.

Proof: Let \( E \) denote \( F[\sqrt{-1}] \). By the usual argument it follows that every element of \( E \) has a square root in \( E \) (exercise 2). Let \( p \in F[X] \) be a monic polynomial. Let \( L \) be the splitting field over \( F \) of \( p(x)(x^2 + 1) \), and assume w.l.o.g. that \( E \subseteq L \). Since the characteristic is 0 the extension is separable, and by theorem 9.1 it is normal; thus it is Galois. Let \( G \) be the Galois group, with \( G = 2^n m \). By Sylow’s theorem \( G \) has a subgroup of order \( 2^n \). The subgroup \( L \) corresponding to this subgroup under the correspondence of theorem 9.4 has degree \( m \) over \( R \); but \( R \) has no odd degree extensions, so \( m = 1 \). If \( e > 1 \) then by theorem 14.9 and theorem 9.4 there is an extension of \( E \) of degree 2. But this contradicts the existence of square roots. Thus, \( p \) splits in \( E \). Finally, if \( p \in E[x] \) then \( pp^* \in R[x] \), so \( pp^* \) has a root in \( E \), for \( p \) does.

Lemma 10. Suppose \( a, b \in F \), \( a < b \), \( p \in F[x] \), and \( p(a)p(b) < 0 \). Then \( p(c) = 0 \) for some \( c \) with \( a < c < b \).
Theorem 12. If $F$ is open and $a, b$ are free variables not occurring in $F$ then $a < b \Rightarrow \exists x (a < x < b \land F)$ is equivalent in the theory of real closed fields to an open formula.

Proof: Define the rank of a formula as in lemma 11.a to be the degree of $x$ in $p$, or one more than the degree of $x$ if $p$ is absent. The proof is by induction on the rank for such formulas $F$. The open formula will also have the properties that its rank is no greater than the rank of $F$, and no atomic formula involves both $a$ and $b$. We will use the notational device of writing $q(x)$ to denote that $x$ is a free variable, and write $q(u)$ for the result of replacing $x$ by $u$; and similarly for other formulas. Using lemma 10 and the hypothesis $a < b$, if $p = 0$ does not occur in $F$ then $F$ is equivalent to the disjunction of

\[ K(a, b) = \forall x (a < x < b \Rightarrow F) \]

\[ L_i = \exists u (a < u < b \land q_i(u) = 0 \land K(a, u)) \]

\[ M_i = \exists u (a < u < b \land q_i(u) = 0 \land K(u, b)) \]

\[ M_{ij} = \exists u \exists v (a < u < v < b \land q_i(u) = 0 \land q_j(v) = 0 \land K(u, v)) \]

$K(a, b)$ is equivalent to

\[ q_1(a) \geq 0 \land \cdots \land q_i(a) \geq 0 \land \neg \exists x (a < x < b \land q_i(x) = 0) \land \cdots \land \neg \exists x (a < x < b \land q_i(x) = 0) \]

Since the rank of $q_i = 0$ is lower than the rank of $F$, inductively $K(a, b)$ is equivalent to a disjunction of formulas $A(a) \land B(b) \land C$. $L_i$ is equivalent to a disjunction of formulas

\[ A(a) \land C \land \exists u (a < u < b \land q_i(u) = 0 \land B(u)); \]
again since the rank of $q_i = 0$ is lower, the induction hypothesis may be applied to each of these. The argument for $M_i$ is similar. $N_{ij}$ is equivalent to a disjunction of formulas

$$C \land \exists u(a < u < b \land q_i(u) = 0 \land A(u) \land \exists v(u < v < b \land q_j(v) = 0 \land B(v))).$$

This may be transformed to a disjunction of formulas

$$C \land C' \land B'(b) \land \exists u(a < u < b \land q_i(u) = 0 \land A(u) \land A'(u)),$$

and finally to the required form. If $p = 0$ does occur in $F$ then $F$ is equivalent to the disjunction of $F_\geq, F_\leq, F_\neq$, where $F_\leq$ is $p = 0 \land p' \leq 0 \land q_1 > 0 \land \cdots \land q_l > 0$ and $p'$ is the derivative with respect to $x$. For $F_\geq$, the formula obtained as in lemma 11 has lower rank than $F$ because $p'$ has lower degree than $p$. For $F$, let $J(a,b) = \forall x(a < x < b \implies p' > 0 \land q_1 \land \cdots \land q_l > 0)$. $F$ is equivalent to the disjunction of

$$K(a,b) = p(a) < 0 \land p(b) > 0 \land J(a,b)$$

where $q_1 = 0$, if any. Property 2. For a value $c < b$, $p' > 0 \land q_1 > 0$. $F$ is equivalent to the disjunction of

$$L_i = p(a) < 0 \land \exists u(a < u < b \land p(u) > 0 \land q_i(u) = 0 \land J(a,u))$$

$$M_i = p(b) > 0 \land \exists u(a < u < b \land p(u) < 0 \land q_i(u) = 0 \land J(u,b))$$

$$M_{ij} = \exists u \exists v(a < u < v < b \land p(u) < 0 \land p(v) > 0 \land q_i(u) = 0 \land q_j(v) = 0 \land J(u,v)),$$

$J$ is transformable as above; and the remaining formulas are since the degree of $x$ in $q_i$ is less than its degree in $p$.

Finally, $\exists x F$ iff $\exists v(0 < v < 1 \land \exists x(-1 < x < 1 \land F(x/v)))$. $F(x/v)$ can be written as $G(x,v)$ by clearing denominators, and two applications of theorem 12 transform $\exists x F$ into an equivalent open formula. Since a sentence in the first order language of ordered fields is true iff it is true in $\mathcal{R}$, it is no longer necessary to re-prove basic facts for real closed fields.

More recently, quantifier elimination methods have been given (for example [Collins] and [Basu]), together with theoretical bounds on their computational complexity, which are improvements over the simple method. Collins’ method has been implemented, and is available as open source code. Further discussion from the original paper may be found in [ACM].

We give some remarks on Collins’ method. First, we prove Sturm’s theorem.

Suppose $p$ is a polynomial in $\mathcal{R}(x)$. As observed in section 5, $p/gcd(p,p')$ has the same roots, and they occur with multiplicity $1$, so suppose $p$ has no multiple roots. Consider the sequence

$$p_0 = 0, \quad p_1 = p', \quad p_{i-2} = q \cdot p_{i-1} - p_i$$

where for $i \geq 2$ $\text{deg}(p_i) < \text{deg}(p_{i-1})$ ($p_i$ is the negative of the remainder). Since $p$ has no multiple roots, and thus $p$ and $p'$ are relatively prime, eventually some $p_i$ (say $p_s$) is a nonzero constant (as in Euclid’s algorithm). The sequence $p_0, \ldots, p_s$ is called the Sturm sequence for $p$. For a value $c$ let $c_i = p_i(c)$; let $\nu(c)$ be the number of sign changes in the sequence $c_0, \cdots, c_s$, where a sign change occurs at $i < s$ if $c_j$ has opposite sign to $c_i$, where $j$ is the least value with $j > i$ and $c_j \neq 0$, if any.

**Theorem 13 (Sturm’s Theorem).** If $p$ is a polynomial with no multiple roots, $a < b$, $p(a) \neq 0$, and $p(b) \neq 0$, then the number of roots of $p$ in the interval $(a,b)$ is $\nu(b) - \nu(a)$, where $\nu$ is as above.

**Proof:** First, the Sturm sequence has the following properties.

1. For $i < s$ if $p_i(x) = 0$ then $p_{i+1}(x) \neq 0$.
2. For $0 < i < s$ if $p_i(x) = 0$ then $p_{i-1}(x)$ and $p_{i+1}(x)$ have opposite signs.

Property 1 follows because otherwise $p_j(x) = 0$ for all $j > i$, contradicting the fact that $p_s(x) \neq 0$. Property 2 follows because $p_{i+1}(x) = -p_{i-1}(x)$ by definition. Clearly, $\nu(x)$ is constant on a closed interval unless it contains a root $c$ of some $p_i$, in which case we may choose $\epsilon$ such that if $x \in [c - \epsilon, c + \epsilon]$ and $x \neq c$ then $p_j(x) \neq 0$ for all $j$, $f_{i+1}$ has constant sign on $[c - \epsilon, c + \epsilon]$, and $f_{i-1}$ also if $i > 0$. It follows that if $i > 0$ then $\nu(x)$ does not change in the interval due to the zero of $p_i$; and if $i = 0$ then $\nu(x)$ decreases by 1 due to the zero of $p$ (with $p_0, p_1$ changing from $-+$ to $++$ or $+-$ to $--$).
Sturm’s theorem was proved by Sturm in 1829. For one example of its use, say that an extension $E \supseteq F$ of ordered fields is an order extension if the order of $F$ is the restriction of the order of $E$. The Artin-Schrier theorem states that for an ordered field $F$, there is a unique real closed algebraic order extension $E \supseteq F$; $E$ is called the real closure of $F$. Sturm’s theorem is used in a standard proof; see [Jacobson].

To describe Collins’ algorithm, let $\Pi_n$ denote $\mathcal{R}[x_1, \cdots, x_n]$. Let $F$ denote a prenex normal form formula, with no negations, and atomic formulas $p = 0$, $p \neq 0$, $p < 0$, $p \leq 0$, $p > 0$, and $p \geq 0$ where $p \in \Pi_n$. Let $\pi_F$ be the polynomials occurring in $F$. A list of “cells” $S \subseteq \mathcal{R}^n$ is constructed recursively, such that for each $p \in \pi_F$, and each cell $S$, either $\forall x \in S(p > 0)$, $\forall x \in S(p = 0)$, or $\forall x \in S(p < 0)$.

In the basis case $\pi_1 \subseteq \Pi_1$, the cell list is constructed using methods such as Sturm’s theorem to “isolate” the roots of the polynomials in the set $\pi_1$ of univariate polynomials. The cells are the roots and the open intervals bounded by consecutive roots. A “sample point” is given for each cell, as an exact representation of an algebraic number, by standard methods described in the literature, and commonly used in computational algebraic number theory.

When $n > 1$, methods involving subresultants are used to transform the set $\pi_n \subseteq \Pi_n$ to a set $\pi_{n-1} \subseteq \Pi_{n-1}$. The cell list for $\pi_{n-1}$ is obtained recursively. By the construction of $\pi_{n-1}$, the components of the union of the zero sets of the $p \in \pi_n$, restricted to any of the cells $S$ of the $\mathcal{R}^{n-1}$ list, are “delineable”, and may be used to decompose the “cylinder” $S \times \mathcal{R}$ into a “stack” of cells over $S$. Sample points are also computed.

By further computations, the above “cellular algebraic decomposition” algorithm may be used to construct an equivalent quantifier-free formula from a prenex normal form formula.

7. Factoring polynomials over $\mathbb{F}_p$. In chapter 12 we mentioned that for many purposes, for measuring the time required to perform a computation we can consider the model of computation to be goto programs, with the atomic operations of $+$ and $\times$ taking 1 unit of time; and up to a polynomial of small degree the time is independent of the model of computation.

For basic operations such as $+$, $\times$, and integer division $m/n$ the time in “bit operations” is of interest; circuit depth or time on a multitape Turing machine are commonly used to give a formal notion ([Knuth], [AHU]). In such measures, $m \pm n$ can be computed in $O(max(\ell(m), \ell(n)))$ bit operations, $m \times n$ in $O(\ell(m)\ell(n))$ bit operations, and $m = qn + r$ in $O((\ell(m) - \ell(n))(\ell(n)))$ bit operations, where $i - j$ is as in section 5.

Euclid’s algorithm for computing the gcd of two nonzero integers may be described as follows. Given integers $u_1, u_2$ with $u_1 \geq u_2 > 0$, inductively write $u_{n-1} = q_nu_n + u_{n+1}$, where $0 \leq u_{n+1} < u_n$. This algorithm may be performed in $\sum_n O(\ell(u_{n+1})(\ell(u_n) - \ell(u_{n+1})))$ bit operations. As noted in section 23.8, $O(\ell(u_1))$ iteration suffice, so the bound is $O(\ell(u_1)(\ell(u_2)))$.

A version of the extended Euclidean algorithm also can be performed in $O(\ell(u_1)(\ell(u_2)))$ bit operations. Let $\alpha_n, \beta_n$ be such that $\alpha_nu_1 + \beta_nu_2 = u_n$; then $\alpha_{n+2} = \alpha_n - q_n\alpha_{n+1}$ and $\beta_{n+2} = \beta_n - q_n\beta_{n+1}$. Arguing as above, $\alpha_n$ or $\beta_n$ can be computed in $O(\ell(u_1)(\ell(u_2)))$ bit operations.

More sophisticated algorithms can be given to asymptotically reduce the number of bit operations for $m \times n$ and $m/n$; again see [Knuth] and [AHU]. In bounding the number of bit operations in a computation, one may bound the number of atomic operations, and the length of the intermediate values; the latter may be less than the general bound given in chapter 12.

From the foregoing, if elements of $\mathbb{F}_p$ are represented by the integers mod $p$, $\pm$ can be performed in $O(\log p)$ bit operations, and $\times$ and $x^{-1}$ in $O((\log p)^2)$, since $x^{-1}$ can be accomplished using the extended Euclidean algorithm. If “fast” multiplication and division are used, times can be reduced asymptotically. From hereon in this section bounds on the computation time will be given as the number of field operations.

The operation $x^k$ where $k$ is an integer can be performed with at most $2\log(k)$ multiplications rather than the more obvious $k - 1$, using the recursion $x^{2k} = x^k \times x^k$, $x^{2k+1} = x \times x^k \times x^k$. 

412
We recall some facts about finite fields, and make some further observations. For any $a \in \mathcal{F}_p$, $a^p = a$. By exercise 9.5.a, for $f, g \in \mathcal{F}_p[x]$, $(f + g)^p = f^p + g^p$. By these two facts, for $f \in \mathcal{F}_p[x]$, $f(x)^p = f(x^p)$. The polynomial $x^p - x$ splits as $\Pi_{a \in \mathcal{F}_p}(x - a)$. Thus, for any polynomial $g \in \mathcal{F}_p[x]$, $g^p - g = \Pi_{a \in \mathcal{F}_p}(g - a)$.

A polynomial $f$ with coefficients in $\mathcal{F}_p$ can be factored by a method due to Berlekamp (see [Cohen, Knuh], and [LidNie] for additional discussion, and [Gao] for more recent results). Let $f$ be of degree $n$. We may assume that $f$ is monic. We may also assume $f$ is squarefree, by first factoring $f / \gcd(f, f')$ and then factoring $\gcd(f, f')$. In the special case where $f' = 0$, the coefficient of $x^k$ can be nonzero only if $p | k$, so $f(x) = f_1(x^p)$. Further, $f_1(x^p) = f_1(x)^p$. Thus in this case $f$ may be replaced by $f_1$.

Step 1 of Berlekamp’s algorithm for factoring a squarefree monic polynomial $f$ of degree $n$ is to form the $n \times n$ matrix $Q$ over $\mathcal{F}_p$, whose $ith$ row for $0 \leq i < n$ is the coefficients of $x^p$ mod $f$; indexing columns from 0 to $n - 1$, $x^p$ mod $f$ equals $\sum_j Q_{ij} x^j$. Multiplying two polynomials of degree $< n$ mod $f$ can be done in $O(n^2)$ field operations. Row 0 can be computed in $O(\log p)$ polynomial multiplications. The remaining rows can be computed with $n$ multiplications per row. Thus, $Q$ can be computed with $O(n^2 \log p + n^3)$ field operations.

Step 2 is to compute a matrix $V$ whose rows are a basis for the left null space $N = Q - I$. Let $k$ denote the nullity of $Q - I$; then $V$ will be $k \times n$, and $|N| = p^k$. Using elementary column operations, $V$ can be computed in $O(n^3)$ field operations. Indeed, ignoring a row permutation, $Q - I$ can be brought to the form $$\begin{bmatrix} I & 0 \\ A & 0 \end{bmatrix},$$ and (up to a column permutation) $V = [-A \ I]$. Let $v_j$ for $1 \leq j \leq k$ denote row $j$ of $V$. We may assume that $v_1 = [1, 0, \ldots, 0]$.

The polynomials $g \in \mathcal{F}_p[x]$ of degree $< n$ are in obvious 1-1 correspondence with the vectors $[g_0, \ldots, g_{n-1}] \in \mathcal{F}_p^n$; we use $g$ indifferently to denote the vector. Letting $\equiv$ denote congruence mod $f$, $\mathcal{F}_p^n$ is a system of coset representatives. Also, $gQ = \sum_{i<n} g_i x^{pi} \equiv g^p$; thus, $N$ is those $g \in \mathcal{F}_p^n$ such that $g^p \equiv g$.

Suppose $f = f_1 \cdots f_l$ is the prime factorization of $f$. Given $[a_1, \ldots, a_l] \in \mathcal{F}_p^l$, by theorem 6.8 there is a unique $g \in \mathcal{F}_p[x]$ with $\deg(g) < n$ and $g \equiv a_i \pmod{f_i}$. Further, $g^p \equiv a_i^p \equiv a_i \equiv g \pmod{f_i}$, so $g^p \equiv g$. Thus, $l \leq k$. On the other hand if $g \in N$, then $f | g^p - g$, so for each $i$ there is an $a_i \in \mathcal{F}_p$ with $f_i | g - a_i$, and $g$ results from $[a_1, \ldots, a_l]$ by the preceding correspondence. Thus, $k = l$.

The relevance of $N$ to factoring $f$ is contained in the following fact. If $f \in \mathcal{F}_p[x]$ is monic and $g \in N$ then $f = \Pi_{a \in \mathcal{F}_p} \gcd(f, g - a)$. Indeed, the factors on the right side $p$ are relatively prime, so $p | f$. Also, $f_i | g - a_i$ for some $a_i$; and so $p | f$. Since $f$ is monic, and $p$ also, $f = p$.

For step 3 of Berlekamp’s algorithm, set $S = \{ f \}$, and repeat the following for $j = 2, \ldots, k$, quitting if $|S|$ reaches $k$.

a. For each $h \in S$, find factors by computing $\gcd(h, v_j - a)$ for $a \in \mathcal{F}_p$.

b. For each $h \in S$, replace $h$ by its factors as found in step a.

Given an irreducible factor $f_i$, let $A_i$ be the vector of values $a_k \in \mathcal{F}_p$ such that $f_i | v_k - a_k$. Then $A_i$ determines the value $a \in \mathcal{F}_p$ such that $f_i | g - a$ for each $g \in N$. It follows that $A_i$ and $A_j$ are distinct for $i \neq j$. It then follows that at some stage of step 3, $f_i$ and $f_j$ will end up in different members of $S$.

For fixed $j$ and $a$, step 3a can thus be performed in $O(n^2)$ steps, whence step 3 can be performed in $O(pk^n)$ field operations. It follows that a squarefree polynomial over $\mathcal{F}_p$ can be factored in $O(pn^3)$ field operations. An arbitrary polynomial can be also, because $n_1^3 + \cdots + n_r^3 \leq (n_1 + \cdots + n_r)^3$.

8. Factoring integer polynomials. We show in this section that a primitive polynomial with integer coefficients can be factored into irreducible factors in polynomial time. Given any $p \in \mathbb{Z}[x]$, its content and primitive part can be computed in polynomial time; however it is open whether an integer can be factored in polynomial time.

To begin with, we ensure that $p$ is squarefree, by dividing $p$ by $\gcd(p, p')$. Also, the resultant $R$ of $p$ and $p'$ is computed. By results in section 6, these steps can be performed in polynomial time. Also,
\[ \ell(R) = O(n(L + \log n)) \] where \( L \) denotes the maximum length of a coefficient in \( p \).

The next step is to find the smallest prime \( P \) such that \( P \nmid R \). We refer the reader to [HardW r] for a proof that the product of the primes \( P \leq N \) exceeds \( e^{cN} \) for some constant \( c \). It follows that \( P = O(\ell(R)) \). \( P \) can be computed in polynomial time using the “sieve of Eratosthenes” (q.v. see [HardW r]), or faster sieves which can be found in the literature.

We introduce some notation to be used below. Given a prime \( P \) and an integer \( k > 0 \), let \( Q \) denote \( P^k \). Given \( p \in \mathbb{Z}[x] \), let \( \bar{p} \) denote \( p \) with its coefficients reduced mod \( P \), and \( \bar{p} \) \( p \) with its coefficients reduced mod \( Q \). Let \( p_P \) denote any polynomial in \( \mathbb{Z}_p[x] \); these polynomials will be considered indistinguishable from the polynomials in \( \mathbb{Z}[x] \), where \( 0 \leq a_i < P \) for each coefficient \( a_i \). Similarly, \( p_Q \) denotes a polynomial in \( \mathbb{Z}_Q[x] \). Recall that \( |p| \) denotes the Euclidean norm of the vector of coefficients of \( p \).

Since \( P \nmid R \), \( \bar{p} \) is squarefree; also it has degree \( n \), since the leading coefficient of \( p \) divides \( R \). The main loop of the factorization algorithm is as follows.

Factor \( \bar{p} \) in \( \mathbb{F}_P \). Set \( p_1 = 1 \), \( p_2 = p \), and repeat the following steps until \( p_2 = \pm 1 \).
1. Let \( h_P \) be a monic irreducible factor of \( \bar{p}_2 \).
2. Execute the procedure “LiftP” given below, to obtain the irreducible factor \( h \) of \( p_2 \) such that \( h_P|\bar{h} \).
3. Set \( p_1 \) to \( p_1 h \), set \( p_2 \) to \( p_2/h \), and remove from the list of factors of \( \bar{p}_2 \) those which divide \( h \).

The procedure LiftP computes \( h \) as follows.
1. Let \( n = \deg(p_2) \) and \( l = \deg(h_P) \). If \( l = n \) let \( h = p_2 \) and exit from LiftP.
2. Let \( k \) be least such that
\[ P^{kl} > 2^{n(n-1)/2} \left( \frac{2(n-1)}{n-1} \right)^{n/2} |p_2|^{2n-1}. \]
3. Use Hensel lifting to obtain \( h_Q \) with \( h_Q \) monic, \( \bar{h}_Q = h_P \), \( h_Q|\bar{p}_2 \), and \( \gcd(h_P, \bar{p}_2/h_P) = 1 \).
4. Let \( u \) be largest such that \( l \leq (n-1)/2^u \). Execute the procedure “LiftQ” given below, for \( m = [(n-1)/2^{u-i}] \), for \( i = 0, 1, \ldots, u \) successively, exiting from LiftP if a value for \( h \) is obtained.
5. Set \( h = p_2 \) and exit from LiftP.

In step 4 of LiftP, it is only necessary to call LiftQ with \( m = n-1 \); smaller values may be tried first to reduce the time required to find smaller degree factors. For \( a = (a_m, \ldots, a_0) \in \mathbb{Z}^{m+1} \) let \( p_a \) be the polynomial \( \sum_{i=0}^{m} a_i x^i \). It is easily seen that \( L = \{a : h_Q|\bar{p}_a\} \) is an integer lattice. Note that \( |a| = |p_a| \), where \( |a| \) is the Euclidean norm. The vectors of the polynomials \( \{Qx^i : 0 \leq i \leq l\} \) and \( \{h_Qx^i : 0 \leq i \leq m-l\} \) are readily seen to form a basis of \( L \). Indeed, by subtracting multiples of \( h_Q \) a polynomial of degree \( l \) may be obtained; its coefficients must all be divisible by \( Q \). Writing a matrix whose rows are the above polynomials in order, one sees that \( \det(L) = Q^l = P^{kl} \).

The procedure LiftQ computes \( h \) or “none” as follows.
1. Let \( L \) be the lattice defined above. Let \( \langle b_1, \ldots, b_{m+1} \rangle \) be an LLL-reduced basis.
2. Let \( B = \langle P^{kl}/|p_2|^{m} \rangle^{1/n} \). If \( |b_1| \geq B \) report “none” and exit from LiftQ.
3. Let \( t \) be largest \( i \) such that \( |b_i| < B \). Set \( h = \gcd(b_1, \ldots, b_t) \) and exit from LiftQ.

If \( p \) is not squarefree, let \( g = \gcd(p, p') \); then the factorization of \( p \) may be obtained from the factorizations of \( p/g \) and \( g \). The factorization of \( g \) is readily obtained once the factorization of \( p \) is known, since the irreducible factors of \( g \) are factors of \( p \). It is not difficult to verify that the above algorithm may be carried out in a number of bit operations polynomial in \( \ell(p) \). A detailed estimate may be found in [LLL].

It remains to show that LiftP finds \( h \), the irreducible factor of \( p_2 \) with \( h_P|\bar{h} \), for which it suffices to show that LiftQ finds \( h \) if \( \deg(h) \leq m \).

**Lemma 14.** Suppose \( g|p_2 \); then
1. \( h|g \) if \( h \) is not squarefree.
2. \( h_Q|g \) if \( h_P|\bar{g} \).

414
Proof: That (1)⇒(3) and (2)⇒(3) are straightforward. Suppose \( h_P | \tilde{g} \); then \( h_P \nmid (p_2/\tilde{g}) \), so \( h \nmid (p_2/g) \), so \( h | g \). Also, since \( h_P \) is irreducible, \( \lambda_P + \mu (p_2/\tilde{g}) = 1 \) for some \( \lambda, \mu \in \mathbb{Z}[x] \), so \( \lambda Q + \mu (p_2/g) = 1 - PV \) for some \( v \in \mathbb{Z}[x] \). Multiplying by \( 1 + PV + \cdots + (PV)^{k-1} \) and \( g \), \( \lambda h Q + \mu p_2 \equiv g \mod Q \) for some \( \lambda', \mu', v \in \mathbb{Z}[x] \). Thus, \( h Q | \tilde{g} \).

Lemma 15. Suppose \( b \in L \) and \( |b| < B \); then \( h | b \).

Proof: Let \( g = \gcd(p_2, b) \); we will show that if \( h \nmid g \) then \( |b| \geq B \). Let \( r = \deg(b) \) and \( s = \deg(g) \). Let \( M \) be the lattice of vectors of the polynomials \( \lambda_P + \mu \) where \( \deg(\lambda) < r - s \), \( \deg(\lambda) < n - s \). Let \( N \) be the lattice of projections onto the components \( n + r - s - 1, \ldots, s \) (numbering from 0 on the right). We will show that \( N \) is a full rank lattice, and \( p^{kl} \leq \det(N) \leq |p_2|^{n-1}|b|^{n-1} \). Suppose \( \kappa = \lambda_P + \mu \in M \), and its projection to \( N \) is 0; then \( \deg(\kappa) < s \) and \( g | s \), so \( \kappa = 0 \). Thus, \( \lambda(p_2/g) + \mu (b/g) = 0 \), whence, since \( \gcd(p_2/g, b/g) = 1 \), \( (p_2/g) | \mu \); and since \( \deg(p_2/g) = g - s \) and \( \deg(\mu) < n - s \), \( \mu = 0 \). It follows that \( \lambda = 0 \), whence \( \kappa = 0 \). From this, the projections of the vectors for the polynomials \( \{ x^i f : 0 \leq i < r - s \} \) and \( \{ x^i b : 0 \leq i < n - s \} \) are a basis for \( N \), from which \( N \) has full rank \( n + r - 2s \), and \( \det(N) \leq |f'|^{n-1} |b|^{n-s} \leq |f|^{n-1} |b|^n \). Now, since \( h | p_2 \), by lemma 14 \( h_Q | p_2 \); and since \( b \in L \), \( h_Q | b \). Thus, \( h_Q \nmid \kappa \). By the hypothesis that \( h_p \nmid \tilde{g} \), \( \lambda h Q + \mu P = 1 - PV \) for some \( \lambda, \mu, \nu \). Multiplying by \( 1 + PV + \cdots + (PV)^{k-1} \) and \( \kappa/g \), \( \lambda' h Q + \mu' \kappa = \kappa/g \mod Q \) for some \( \lambda', \mu' \). Thus, \( h_Q | \kappa/g \), and so if \( \deg(\kappa) < l + s \) then \( \deg(\kappa/g) < l \), whence \( \kappa/g = 0 \). In particular, \( Q | \kappa/g \), whence \( Q | \kappa \). It follows from this and \( s + l \leq n_r \) that the first \( l \) diagonal elements of a Hermite normal form basis for \( N \) are divisible by \( Q \), whence \( \det(N) \geq p^{kl} \). To see that \( s + l \leq n - s \), note that \( l \leq n - s \) because \( g | b \), so it suffices to show that \( s + l \leq n \). This follows because \( h \) is an irreducible factor of \( p_2 \), and \( h | g \) so \( h | (p_2/g) \).

Lemma 16. Suppose \( L \subseteq \mathcal{R}^n \) is a lattice, \( b_1, \ldots, b_n \) is an LLL-reduced basis, and \( x_1, \ldots, x_t \in L \) are linearly independent. Then for \( 1 \leq i \leq t \), \( |b_i| \leq 2^{-n/2} \max|x_1, \ldots, x_t| \).

Proof: For \( 1 \leq i \leq n \) and \( 1 \leq j \leq t \) there are \( r_{ij} \in \mathbb{Z} \) (\( r_{ij} \in \mathbb{R} \)) such that \( x_j = \sum r_{ij} b_i \) (\( x_j = \sum r_{ij} b_i \)), where the \( b_i \) are the orthogonalized basis as in section 23.5. Let \( i \) be the largest \( i \) such that \( r_{ij} \neq 0 \); we claim that \( |x_j| \geq |b_i| \). Indeed, using the identity \( G = GU \) of section 23.5, \( |x_j|^2 \geq r_{i1,j}^2 |b_i|^2 = r_{i1,j}^2 |b_i|^2 \geq |b_i|^2 \). We may assume that the \( i_j \) are nondecreasing; it then follows from the linear independence of the \( x_i \) that \( x_{ij} \geq j \). Thus, by theorem 23.18.b, \( |b_j|^2 \geq 2^{-1} |b_i|^2 \leq 2^{-n} |x_j|^2 \), from which the lemma follows.

Suppose that \( b_1, \ldots, b_n \) is an LLL-reduced basis for the lattice \( L \) of LiftQ. Let \( J = \{ j : |b_j| < B \} \). We will show that if \( J \) is nonempty then \( 1 \in J \), and furthermore \( h = \gcd(\{ b_j : j \in J \}) \).

Let \( h_1 = \gcd(\{ b_j : j \in J \}) \); by lemma 15 \( h b_j \) for \( j \in J \), so \( h | h_1 \). For \( j \in J \), \( b_j \in \{ \lambda h_1 : \deg(\lambda) \leq m - \deg(h_1) \} \); it follows that \( J \leq m + 1 - \deg(h_1) \).

For \( 0 \leq i \leq m - \deg(h) \), \( h x_i \in L \). Suppose \( 1 \leq j \leq m + 1 - \deg(h) \). By lemma 16, and the fact that \( |h x_i| = |h|, |b_j| \leq 2^{m/2} |h| \), whence by Mignotte’s theorem \( |b_j| \leq 2^{m/2} (2^m |p_2|) \). The inequality \( 2^{m/2} (2^m |p_2|) < B \) is equivalent to \( P^{kl} > 2^{m/2} (2^m |p_2|)^{n/2} \); by step 2 of LiftP we may assume this holds. Thus, \( j \in J \).

From the preceding two paragraphs it follows that \( \deg(h) = \deg(h_1) = \max \{ j : j \in J \} \). By lemmas 14 and 15, \( h \in L \). Since \( h | p_2 \), \( h \) is primitive. Letting \( \bar{p} \) denote the primitive part of \( p \) as above, for \( j \in J \) \( h | b_j \), so \( h | b_j \); thus, \( b_j \in L \). Since the \( b_i \) are a basis, \( b_j = b_j \). In particular, \( h_1 \) is primitive. It follows that, up to sign, \( h = h_1 \).

Exercises.

1. Prove theorem 4. Hint: Fix values for the variables other than \( x \). First, \( F(x) = F_{-\infty}(x) \) for \( x \) sufficiently small. Second, \( F_{-\infty}(x) = F_{-\infty}(x - m N) \) for any \( m \). From these two facts, if the expression is true then \( \exists x F \) is true. Suppose \( \exists x F(x) \) is true. If \( F(x - N) \) were false then by positivity some atomic
formula $k < x - N$ must be false, while $k < x$ is true. But then $x = k + j$ where $k \in S$ and $1 \leq j \leq N$, and the expression is true. Continuing to subtract $N$, we must obtain such an $x$, or reach a low enough value that $F_{-\infty}(x)$ is true. Adding a multiple of $N$ to $x$ makes the expression true.

2. Show that if $F$ is a real closed field and $E = F[\sqrt{-1}]$, then every element of $E$ has a square root in $E$. Hint: From $(c + di)^2 = a + bi$, $c^2 - d^2 = a$ and $2cd = b$. The case $c = 0$ is easily handled; otherwise $\gamma^2 - a\gamma - b^2/4 = 0$ where $\gamma = c^2$.

3. Show that if $F$ is a real closed field and $q \in F[x]$ is irreducible and monic then $q(x) > 0$ for all $x$. Hint: If $g = x^2 + bx + c$ then $b^2 < 4c$, whence $g(x) = (x + b/2)^2 + (4c - b^2)/4$ is positive for all $x$.

4. Prove that $\sum_{k=0}^m \binom{m}{k}^2 = \binom{2m}{m}$. Hint: Using $(1 + x)^{m_1 + m_2} = (1 + x)^{m_1} (1 + x)^{m_2}$ and the binomial theorem, show that $\sum_{k_1+k_2=n} \binom{m_1}{k_1} \binom{m_2}{k_2} = \binom{m_1+m_2}{n}$.
Appendix 1. Set theory.

In this appendix, the Bernstein-Cantor-Schröder theorem is proved. Ordinal and cardinal numbers are defined, and some basic facts about them given. A discussion is given of transfinite induction. A more complete treatment can be found in any of numerous introductory set theory texts.

Suppose injective functions \( f : S \rightarrow T \) and \( g : T \rightarrow S \) are given; we wish to construct a bijection between \( S \) and \( T \). By identifying \( T \) with its image under \( g \), we may consider \( T \) to be a subset of \( S \); we thus have \( T \subseteq S \) and \( f : S \rightarrow T \) injective. Define \( S_0 = X \), \( S_{i+1} = f[S_i] \); and \( T_0 = T \), \( T_{i+1} = f[T_i] \). Since \( f \) is injective, \( f[S_i - T_i] \supseteq f[S_i] - f[T_i] \). It follows that each point either belongs to some \( S_i - T_i \), some \( T_{i+1} - S_i \), or every \( S_i \). We may map points in the first category to \( f(x) \), and the remaining points to themselves.

The ordinal numbers, or simply ordinals, are one of the most important constructions in set theory. Many of their basic properties were discovered by Cantor. As we will see, every well-ordered set is isomorphic to a unique ordinal number, so that the ordinals form a system of representatives of the “order types” of the well-orderings. The modern definition of an ordinal yields a particularly well-behaved such system.

If \( x \) is a set \( \in \) is a relation on its members. The axioms of set theory ensure that this relation is well founded, that is, there are no infinite descending chains, or equivalently any set \( x \) has elements which are minimal under the \( \in \) relation.

A set \( x \) is called transitive if \( w \in x \) and \( v \in w \) imply \( v \in x \). Although the notion of a transitive set is a technical one, it has turned out to be quite useful in set theory. A transitive set \( x \) is called an ordinal if for any \( v, w \in x \) either \( v \in w \), \( v = w \), or \( w \in v \). This requirement is quite restrictive, as we shall see. We use \( \alpha, \beta, \gamma, \delta \) to denote ordinals, as is commonly done in set theory.

**Lemma 1.**

a. If \( x \subseteq \alpha \) is transitive then \( x \) is an ordinal.

b. If \( x \in \alpha \) then \( x \) is an ordinal.

c. If \( \beta \subseteq \alpha \) then \( \beta \in \alpha \).

d. Either \( \alpha \subseteq \beta \) or \( \beta \subseteq \alpha \).

**Proof:** Part a follows because if \( v, w \in x \) then \( v, w \in \alpha \), so they are related by \( \in \). For part b, \( x \subseteq \alpha \) because \( \alpha \) is transitive. If \( w \in x \) and \( v \in w \) then \( v, w \in \alpha \), so \( w \in x \), \( w = x \), or \( x \in w \); but the latter two possibilities contradict well foundedness. For part c, suppose \( \gamma \) is a minimal element of \( \alpha - \beta \). If \( \delta \in \gamma \) then \( \delta \in \alpha \), so \( \delta \in \beta \) else \( \gamma \) is not minimal. If \( \delta \in \beta \) then \( \delta \in \gamma \), since \( \delta = \gamma \) or \( \gamma \in \delta \) both imply \( \gamma \in \beta \). Thus, \( \beta \) and \( \gamma \) have the same elements and so are the same set (by an axiom of set theory), and \( \beta \in \alpha \) as was to be shown. For part d, \( \alpha \cap \beta \) is readily verified to satisfy the defining properties of an ordinal. If \( \alpha \cap \beta = \alpha \) then \( \alpha \subseteq \beta \), and if \( \alpha \cap \beta = \beta \) then \( \beta \subseteq \alpha \); in the remaining case by part c \( \alpha \cap \beta \in \alpha \) and \( \alpha \cap \beta \in \beta \), so \( \alpha \cap \beta \in \alpha \cap \beta \), a contradiction.

The collection of ordinals is not a set; it is an example of a type of collection in set theory called a proper class. The fundamental example if a proper class is the collection of all sets. That this is not a set follows by the argument of Russell’s paradox. That the ordinals are not a set also follows by contradiction. For most purposes a proper class may be thought of as an additional predicate on the universe of sets, and treated as any first order predicate.

In particular, the notation \( \alpha < \beta \) is used for \( \alpha \in \beta \), and satisfies the laws for a well-order, on a proper class rather than a set. If \( S(\alpha) \) is a statement about ordinals, if it is false there is a least ordinal \( \alpha \) for which it is false. So if we can show that \( S(\alpha) \) is a consequence of the fact that \( S(\beta) \) is true for \( \beta < \alpha \), then \( S(\alpha) \) must be true for all ordinals. This is the principal of transfinite induction.

**Lemma 2.**

a. \( \emptyset \) is an ordinal, denoted \( 0 \). \( 0 \leq \alpha \) for any \( \alpha \).
b. $\alpha \cup \{\alpha\}$ is an ordinal, called the successor of $\alpha$ and denoted $\alpha + 1$. If $\alpha \leq \beta \leq \alpha + 1$ then either $\beta = \alpha$ or $\beta = \alpha + 1$.

**Proof:** For part a, $\emptyset$ satisfies the requirements for an ordinal vacuously; and $\emptyset \subseteq \alpha$. For part b, the requirements are readily verified, and $\alpha \subseteq \beta \subseteq \alpha \cup \{\alpha\}$.

The ordinals obtained from 0 by repeatedly taking the successor are precisely the (standard representations as sets of the) natural numbers. In particular induction on the integers is a special case of transfinite induction, because the natural numbers are an initial segment of the ordinals.

An ordinal $\alpha$ is called a successor ordinal if there is a $\beta < \alpha$ such that $\alpha = \beta + 1$; it is called a limit ordinal if it is not 0, and for all $\beta < \alpha$ $\beta + 1 < \alpha$. An ordinal is either 0, a successor ordinal, or a limit ordinal. Transfinite inductions are commonly broken down into these three cases.

Transfinite induction is often used in conjunction with transfinite recursion, where a function $F$ is defined from the ordinals to the sets. To define $F(\alpha)$, apply a function $G$ from sets to sets, to the set $f$, which is the restriction of $F$ to $\alpha$. More formally, $F(\alpha) = x$ if there exists an $f$ such that $f$ is a function with domain $\alpha$, $f(\xi) = G(f|\xi)$ for all $\xi < \alpha$, and $x = G(f)$. Using transfinite induction and some other basic facts from set theory it can be shown that for all $\alpha$ there is a unique $f$ and $x$.

Next we outline a proof that any well-order is isomorphic to a unique ordinal (further the isomorphism is unique). First, the only automorphism (say $f$) of a well-order (on $S$) is the identity. If not there is a least $x \in S$ such that $f(x) \neq x$. If $f(x) = y$ then $y < x$ is impossible because then $f(y) = y$. On the other hand if $x > y$ then if $z = f^{-1}(x)$, $z < x$, so $f(z) = z = x$, a contradiction.

Let $F$ be the class of functions $f : \alpha \to S$ where $\alpha$ is an ordinal and $f$ is an order isomorphism between $\alpha$ and $f[\alpha]$ such that $f[\alpha]$ is $\prec$-closed. By the fact in the preceding paragraph, these are a chain. The union of the functions in $F$ is a set, and is the unique isomorphism, and its domain the unique ordinal.

The well-ordering principle states that any set $S$ has a well-order. Transfinite induction can be used to prove this fact, using a choice function for $S$; this is a function whose domain is the nonempty subsets $T \subseteq S$, where $f(T) \in T$. A well-ordering $g$ may be defined as the maximal function such that $g(\beta) = f(S - g(\beta))$ for all $\beta < \alpha$ where $\alpha$ is the domain of $f$. A fine point is that there is a limit to how long the recursion continues; this is ensured by the axioms of set theory, in particular the axiom of replacement.

The least ordinal $\alpha$ such that $f[\alpha] = S$ for some $f$ is called the cardinality of $S$. Those ordinals which are the cardinality of some set are called cardinal numbers. An ordinal $\alpha$ is a cardinal iff it is not in one-to-one correspondence with any smaller ordinal, iff it has no injection into a smaller ordinal, iff it has no surjection from a smaller ordinal.

It is not difficult to verify that the integers are cardinal numbers. The set of integers is also a cardinal, denoted $\omega$. A set is called finite if its cardinality is an integer, otherwise infinite. An infinite cardinal is easily seen to be a limit ordinal (the integers can be “shifted” to make room for the last element of $\alpha + 1$ in an injection into $\alpha$). An infinite set is called countable if its cardinality is $\omega$, otherwise it is called uncountable.

To prove that $|S| = |S \times S|$ for infinite $S$ one can construct an explicit bijection from $\kappa$ to $\kappa \times \kappa$ for any infinite cardinal $\kappa$. Indeed, an explicit bijection can be constructed from ordinals to ordinal pairs, which “closes off” (maps $\alpha$ bijectively to $\alpha \times \alpha$) frequently, in particular at cardinals. The details of this construction would take us too far afield and are omitted; see [Jech] for example. Various consequences are readily drawn. For example if $S$ has infinite cardinality $\kappa$ then $S^k$ (the set of ordered $k$-tuples) does also, for any $k$. The set of finite sequences does also; there is an injection to $\omega \times \kappa$ mapping a sequence of length $k$ to the pair $(k, \alpha)$ where $\alpha$ is the code for the $k$-tuple.

Next we prove the claim of chapter 3, that if $L$ is a closure system which is closed under unions of chains then it is closed under unions of directed sets. Let $D$ be a directed set, and let $D'$ be the set $\{S \in L : \exists T \in D(S \subseteq T)\}$; that is, add to $D$ any subsets of its sets, which are in $L$. Then
1. $D \subseteq D' \subseteq L$,
2. if $S$ is an upper bound for $D$ then $S$ is an upper bound for $D'$,
3. $D'$ is directed, and
4. $D'$ is closed under $\subseteq$.

These are straightforward from the definitions. They imply that $D'$ is closed under the pairwise join of $L$.

From hereon assume that $D$ is directed and closed under $\subseteq$. Let $D'$ be $D$, with $\cup C$ added for any $C \subseteq D$. $D'$ again satisfies 1-4; we prove 3 and leave the rest to the reader. Suppose $C_1, C_2$ are chains in $D$. Using the axiom of choice and transfinite recursion sequences $S_{i\xi}$ for $\xi < \alpha$, may be defined, such that $S_{i\xi}$ is “cofinal” in $C_i$, meaning that $S_{i\xi} \in C_i$, if $\xi < \eta$ then $S_{i\xi} \subseteq S_{i\eta}$, and if $S \in C_i$ then for some $\xi$ $S_{i\xi} \subseteq S$.

By “padding” $\alpha_1 = \alpha_2$ can be ensured (the sequences need only be nondecreasing). Finally, the sequence $S_{1\xi} \cup S_{2\xi}$ is in $D$, so its union is in $D'$, and is an upper bound for $\cup C_1$ and $\cup C_2$.

Now let $D_0 = D$, $D_{\alpha+1} = D'_\alpha$, and $D_\alpha = \cup_{\beta < \alpha} D_\beta$ if $\alpha$ is a limit ordinal. The process terminates at some ordinal, resulting in a set $E$ with properties 1-4, which further is closed under unions of chains. Note that its join equals that of $D$. Finally, using transfinite recursion and the fact that $E$ is directed, and taking unions at limits, choose a cofinal chain in $E$. Its union is the join of $E$, and hence of $D$.

Next we prove the infinite case of theorem 8.6. Let $\{x_\beta < \kappa\}$ be a basis for $M$ where $\kappa$ is the cardinality of $M$ (although a well-ordering of any order type will do). For $\beta < \kappa$ let $N_\beta = N \cap \text{Span}\{x_\gamma: \gamma < \beta\}$. If $\beta$ is a limit ordinal then $N_\beta = \cup_{\gamma < \beta} N_\gamma$, since anything in $N_\beta$ is a linear combination of some $x_\gamma$’s with $\gamma < \beta$. A basis $B_\beta$ for $N_\beta$ may be defined by transfinite recursion; at successor stages the argument is as in theorem 8.6, and at limit stages $B_\beta = \cup_{\gamma < \beta} B_\gamma$. 

419
Appendix 2. Basic linear algebra.

This appendix assumes the reader is familiar with the notion of a finite dimensional vector space $V$ over a field $F$, in particular the definitions of linear independence, basis, and subspace; and the fact that the dimension $n$ is the size of any basis. This material is covered in section 8.7. The appendix covers linear equations, elementary operations, Gaussian elimination, matrices, and the determinant.

We define a linear combination over $F$ to be an expression $\sum_{0 \leq i \leq k} a_i x_i$, where $a_i$ is an element of $F$ and the $x_i$ are distinct variables. An element from $F$ is called a scalar in this context; and $a$ in a term $a x$ is called the coefficient of $x$. If vectors $v_i \in V$ are substituted for the variables, the expression evaluates to a vector, which is called a linear combination of the $v_i$. (This terminology applies for a module over a commutative ring).

A linear equation is defined to be a formula of the form $C = b$, where $C$ is a linear combination and $b \in V$. More general linear equations (such as $C_1 + b_1 = C_2 + b_2$) can be “brought to” this form. A system of linear equations in the variables $x_1, \ldots, x_n$ is some number $m$ of linear equations $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$, $1 \leq i \leq m$ (note that some $a_{ij}$ may be 0, and by convention such terms can be omitted when writing the equation).

An $n$-tuple of values $v_i \in V$ is said to satisfy the system or be a solution to the system if the equations are all true when $v_i$ is substituted for $x_i$.

In many situations in mathematics, both theoretical and applied, one wishes to find the solutions to a system of linear equations. Sometimes one expects there to be a single solution, and sometimes a larger set. When one expects a single solution, the number of equations often equals the number of unknowns. There is a simple general method for solving systems of linear equations, called Gaussian elimination. Carl Friedrict Gauss was a great German mathematician, for whom many things in mathematics are named. The term “degaussing” for example is used for eliminating unwanted magnetic fields; this stems from the fact that Gauss did fundamental work in the mathematics of electromagnetism.

The fundamental fact underlying Gaussian elimination is that any solution to a system of linear equations is a solution to any linear combination of equations of the system. An equation can be multiplied by a scalar; the equation $c a_1 x_1 + \cdots + c a_n x_n = c b$ is the result of multiplying the equation $a_1 x_1 + \cdots + a_n x_n = b$ by the scalar $c$. Any solution to the original equation is still a solution to the new equation; this is easily seen from the vector space axioms, in particular the generalized distributive law $x(y_1 + \cdots + y_n) = xy_1 + \cdots + xy_n$, an easy consequence of the axioms.

Two equations can be added; the equation $(a_1 + a'_1)x_1 + \cdots + (a_n + a'_n)x_n = b + b'$ is the result of adding the equations $a_1 x_1 + \cdots + a_n x_n = b$ and $a'_1 x_1 + \cdots + a'_n x_n = b'$. Any solution to the original equations is a solution to the new equation, again an easy consequence of the axioms. It follows that a solution to a system of equations is a solution to any linear combination of equations from the system.

Thus, if a linear combination of equations of a system is added to the system, the solutions will not change, that is, the system with the new equation added will have same the solutions, so it is redundant and can be omitted.

Suppose now that we modify the system by adding to some equation (say equation number $i$, denoted $E_i$) a multiple of some other (say $c E_j$). Such an operation is called an elementary operation; we claim it does not change the solution set. To see this, adding the equation $E_i + c E_j$ to the system does not change the solutions. Further $E_i$ is now redundant (it equals $(E_i + c E_j) - c E_j$) and can be omitted.

Elementary operations may be used to “eliminate” variables from equations in the system, thereby resulting in a simpler system. If $a_{ij} \neq 0$, subtracting $a_{ij}/a_{ij}$ times $E_i$ from $E_{i'}$ (where $E_i$ is $a_{i1} x_1 + \cdots + a_{in} x_n = b_i$) eliminates $x_j$ from equation $E_{i'}$ by setting its coefficient to 0.

Gaussian elimination is a strategy for eliminating variables which eventually will result in a system...
in a special form which is easy to solve. Actually, first we will describe a variation called Gauss-Jordan elimination. A step of the Gauss-Jordan elimination procedure consists of eliminating some variable from all but one equation. Say we wanted to eliminate \(x_j\); find some \(a_{ij} \neq 0\), and for each \(i' \neq i\) subtract from the \(i'\)th equation the \(i\)th equation times \(a_{i'j}/a_{ij}\). This has the effect of setting all the coefficients of \(x_j\), except \(a_{ij}\), equal to 0.

An example is helpful. Following is a system of 5 equations in 5 “unknowns”, an alternative term for the variables.

\[
\begin{align*}
(1) & \quad 100x_1 + 50x_3 + 70x_4 = 20 \\
(2) & \quad 50x_2 + 50x_3 + 80x_5 = 40 \\
(3) & \quad x_1 + x_2 - x_3 = 0 \\
(4) & \quad x_3 - x_4 - x_5 = 0 \\
(5) & \quad -x_1 + x_4 = 0 \\
\end{align*}
\]

Adding 1 times equation 5 to equation 3 eliminates \(x_1\) from equation 3, and adding 100 times equation 5 to equation 1 eliminates \(x_1\) from equation 1. This eliminates \(x_1\) from all equations but number 5.

The resulting system is as follows.

\[
\begin{align*}
(1) & \quad 50x_3 + 170x_4 = 20 \\
(2) & \quad 50x_2 + 50x_3 + 80x_5 = 40 \\
(3) & \quad +x_2 - x_3 + x_4 = 0 \\
(4) & \quad x_3 - x_4 - x_5 = 0 \\
(5) & \quad -x_1 + x_4 = 0 \\
\end{align*}
\]

As we have shown this has the same solutions as the original system.

To solve a system of linear equations, simply eliminate variables one after another until no more eliminations are possible. However, once an equation has been used to eliminate some variable, it should be marked in some way and not used again. Once a variable \(x_j\) has been eliminated it has a nonzero coefficient in only one equation, say equation \(i\). Succeeding elimination steps will not disturb this property if equation \(i\) is not used again; any other equation has 0 for its \(x_j\) coefficient, so the coefficients of \(x_j\) in any equation remain unaltered by succeeding steps.

In the example equation 5 should not be used again, say to eliminate \(x_4\) from the other equations. If it is not the coefficients of \(x_1\) will remain unaltered. A step which can be done is to eliminate \(x_3\) using equation 3, resulting in the following.

\[
\begin{align*}
(1) & \quad 50x_2 + 220x_4 = 20 \\
(2) & \quad 100x_2 + 50x_4 + 80x_5 = 40 \\
(3) & \quad x_2 - x_3 + x_4 = 0 \\
(4) & \quad x_2 - x_5 = 0 \\
(5) & \quad -x_1 + x_4 = 0 \\
\end{align*}
\]

In general any nonzero \(a_{ij}\) where equation \(i\) has not yet been used can be selected. In specific examples there might be some advantage to one choice over another, but any choice will work. Computer implementations would use some fixed strategy, sometimes for “numerical” reasons (for example the entry with largest absolute value; this strategy goes by the prosaic name of “total pivoting”).

When an impasse is reached, there might be some equations where the coefficients are all zero. If so, if such an equation has a nonzero right hand side, the system of linear equations has no solution. For example, \(x = 2\) and \(x = 3\) is a system of linear equations with no solution. If such an equation has a right hand side of zero it is redundant and may be ignored.
We may thus assume that no equation has all coefficients zero. This implies that every equation (excluding any discarded) has been used, since as noted above any nonzero $a_{ij}$ in an unused equation can be used. By renumbering the variables and rearranging the system we may assume it is in the form of figure 1, which we will call standard form: $1 \leq m \leq n$, $x_1, \ldots, x_m$ are the variables used in the elimination, and $a_{ii} \neq 0$ for $1 \leq i \leq m$.

\[
\begin{align*}
a_{11}x_1 &+ a_{12}x_2 + \cdots + a_{1m+1}x_{m+1} + \cdots + a_{1n}x_n = b_1 \\
a_{22}x_2 &+ a_{2m}x_m + \cdots + a_{2n}x_n = b_2 \\
\vdots & \quad \vdots \\
a_{mm}x_m &+ a_{m,m+1}x_{m+1} + \cdots + a_{mn}x_n = b_m
\end{align*}
\]

Figure 1 - Standard form

A standard form for a given set of equations is not unique, depending on which variables were used, and on the order in which they were used. However the solutions to the system will be the same, namely the solutions to the original system.

Once a system is in standard form it is easy to determine the solutions, of which there may be more than one (or as we have seen none). Make any assignment whatever to the unused variables $x_{m+1}, \ldots, x_n$. Each equation then has a single variable left and may be solved for the value of that variable. If $m = n$ there are no unused variables and the system has a single solution.

The term “underdetermined” is used to refer to a system where there is more than one solution; but sometimes one is expecting many solutions rather than one. The term “overdetermined” refers to a system with more equations than unknowns. In this case Gauss-Jordan elimination will produce equations with all coefficients 0. If any right hand side of these is nonzero the system has no solution; if they are all 0 they may be considered redundant and thrown away.

Gauss-Jordan elimination works in any field. This has as a consequence the fact that if a system where all the $a_{ij}$ and $b_i$ are rational has a real solution it has a rational solution. To see this, note that the Gauss-Jordan elimination procedure will never generate any coefficients which are not rational, since the rationals are closed under the field operations.

There is a comment worth making on the notation we have used in this chapter. The symbols $x$, $x_1$, etc. are used to denote variables, and $a$, $a_{ij}$, etc. to denote coefficients, i.e., members of the ring. One sometimes says that the latter are “metavariables” of the language used to discuss the expressions, and the former are the actual variables of the object language. This point was discussed at great length by early logicians. One can see from this chapter that in many cases the distinction becomes blurred; for example the identities used to justify Gauss-Jordan elimination treat the coefficients and variables on an equal footing.

There is a variation of the algorithm given above, which saves computational steps. Once an equation has been used, we may ignore it completely, in the sense that when a later step eliminates, say, $x_j$, it is only eliminated from unused equations and not from used ones. This results (after rearrangement if necessary) in a system in the following form.

\[
\begin{align*}
a_{11}x_1 &+ a_{12}x_2 + \cdots + a_{1m+1}x_{m+1} + \cdots + a_{1n}x_n = b_1 \\
a_{22}x_2 &+ a_{2m}x_m + \cdots + a_{2n}x_n = b_2 \\
\vdots & \quad \vdots \\
a_{mm}x_m &+ a_{m,m+1}x_{m+1} + \cdots + a_{mn}x_n = b_m
\end{align*}
\]

After an assignment to the unused variables, a unique solution for the used variables may be obtained, in the order $x_m, x_{m-1}, \ldots, x_1$, by the method known as back substitution. Certainly $x_m$ can be determined using equation $m$, and after that $x_{m-1}$ can be determined using equation $m - 1$, and so forth.
This latter algorithm is usually called Gaussian elimination. Gaussian elimination involves fewer computational steps than Gauss-Jordan elimination, so is preferable in algorithms for solving systems of linear equations. Gauss-Jordan elimination results in a standard form which is more useful in mathematical proofs in linear algebra.

\( F^n \) may be considered a vector space over \( F \), with componentwise operations; its dimension is \( n \), since the vectors which are 1 in one component and 0 in the remaining components (we call such vectors standard unit vectors) comprise a basis. Figure 2 shows the unique way of writing a vector as a linear combination of standard unit vectors; it uses the convention of writing the vector as a column, with the first component at the top.

Let \( \sum_j a_{ij} x_j = b_i \), \( 1 \leq i \leq m \), \( 1 \leq j \leq n \), be a system of \( m \) linear equations in \( n \) unknowns; the solutions may be considered to be vectors in \( F^n \). The equations may also, with the \( j \)th component of the vector for equation \( i \) being \( a_{ij} \). These vectors are called the rows of the system, from arranging the system in a array where the equations are in rows, with a column for each variable, as in the example above.

To characterize the set of solutions, first, replace all the \( b_i \) by 0; we will call the solutions to this system the homogeneous solutions. Applying Gauss-Jordan elimination will result in a system of \( r \leq m \) equations in standard form; the right sides are all 0. The number \( r \) is called the rank of the system; \( r \) is the dimension of the space that the rows generate, the so-called row space of the system. This is easily seen; indeed an elementary operation does not change the row space, since it may be reversed, and further the rows in standard form are linearly independent, since each contains a nonzero position which is 0 in the others. Note that \( r \leq n \), since in standard form each used variable is in only one row, so \( r \) is the number of used variables, and there are only \( n \) variables.

The homogeneous solutions form a subspace of dimension \( m - r \) where \( r \) is the rank of the system. To see this, choose an unused variable; assign it the value 1, assign the remaining unused variables 0, and determine the assignment to each used variable from the row where it has nonzero coefficient. Doing this for each unused variable yields a basis for the set of solutions, as in figure 3, for the system in figure 1. Any assignment to the unused variables occurs in exactly one linear combination of these solutions, and an assignment to the unused variables determines the solution.

The original system (with possibly nonzero \( b_i \)) may not have solutions; this is the case iff Gauss-Jordan elimination yields an equation with all 0 coefficients and a nonzero right side. If it does have solutions, let \( y \) be any solution vector, a so-called particular solution. We claim that the solutions are exactly the vectors \( y + x \) where \( x \) ranges over the homogeneous solutions. Certainly any such is a solution, because \( \sum_j a_{ij} y_j = b_i \).
and $\sum_j a_{ij}x_j = 0$, so $\sum_j a_{ij}(y_j + x_j) = b_i$. Conversely given any solution $z$, $x = z - y$ is a homogeneous solution, and $z = y + x$.

To repeat, the solutions to a system of linear equations consist of the subspace of homogeneous solutions, translated by adding some fixed particular solution to each vector in the subspace. The Gauss-Jordan elimination procedure furnishes both a particular solution and a basis for the subspace.

Suppose $\sum_j a_{ij}x_j = b_i$, $1 \leq i \leq m$, $1 \leq j \leq n$ is a system of linear equations. Each vector $x \in F^n$ gives rise to a vector $u \in F^m$, namely that where the $i$th component $u_i$ equals $\sum_j a_{ij}x_j$. The array of field elements $a_{ij}$ thus determines a function from $F^n$ to $F^m$. Letting $A$ denote this function, the system of linear equations may be written as $A(x) = b$; it has a solution if and only if $b$ is in the range of $A$.

We claim that $A(cx) = cA(x)$, and that $A(x + y) = A(x) + A(y)$. To see the first, if $u = A(x)$ then $u_i = \sum_j a_{ij}x_j$, and so $\sum_j a_{ij}cx_j = cu_i; \text{ that is, the } i \text{th component of } u = A(x), \text{ which is exactly the claim. If also } v = A(y) \text{ then } v_i = \sum_j a_{ij}y_j, \text{ and so } \sum_j a_{ij}(x_j + y_j) = u_i + v_i, \text{ which proves the second claim.}

A function $A : F^n \rightarrow F^m$ such that $A(cx) = cA(x)$ and $A(x + y) = A(x) + A(y)$ is called a linear transformation. An array of field elements $A_{ij}$, $1 \leq i \leq m$, $1 \leq j \leq n$ is called an $m \times n$ matrix. To see this, any $x \in F^n$ equals $x_1e_1 + \cdots + x_ne_n$ where $x_j$ is the $j$th component of $x$: by the linearity of $A$, $A(x) = x_1A(e_1) + \cdots + x_nA(e_n)$. The $i$th component of $A(x)$ is thus $A_{1i}x_1 + \cdots + A_{ni}x_n$; if $x = e_j$ this equals $A_{ij}$, as was to be shown.

The vector of elements $A_{ij}$ for $i$ fixed is called a row, of the matrix $[A_{ij}]$; the vector $A_{ij}$ for $j$ fixed is called a column. This terminology is quite descriptive when the elements of the matrix are arranged on paper in the obvious fashion (figure 4). The operation of applying the linear transformation $A$ determined by the matrix to a vector $x \in F^n$ is easy to visualize if $x$ is written as a column to the right of the matrix. In most contexts the symbol $A$ may be used to denote indifferently the matrix or the linear transformation. We use the notation $Ax$ for $A(x)$, considering $A$ as the matrix and $x$ a column vector, although many other conventions are in use.

$$
\begin{bmatrix}
  a_{11}x_1 + \cdots + a_{1n}x_n \\
  \vdots \\
  a_{m1}x_1 + \cdots + a_{mn}x_n
\end{bmatrix}
= 
\begin{bmatrix}
  a_{11} & \cdots & a_{1n} \\
  \vdots & \ddots & \vdots \\
  a_{m1} & \cdots & a_{mn}
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  \vdots \\
  x_n
\end{bmatrix}
$$

Figure 4 - $Ax$

The bijective correspondence between $m \times n$ matrices and linear transformations from $F^n$ to $F^m$ is in fact an isomorphism of vector spaces. The linear transformations form a vector space under “pointwise” addition and scalar multiplication. The corresponding operations on the matrices are matrix addition and scalar multiplication; $(M + N)_{ij} = M_{ij} + N_{ij}$, and $(cM)_{ij} = cM_{ij}$. The additive identity is the “zero” matrix $M$ where $M_{ij} = 0$ for all $i,j$. These facts follow immediately and are left to the reader.

Composition of linear transformations induces a corresponding operation on matrices, called matrix multiplication. We already know from what we have shown that if $A$ is a $m \times n$ matrix (determining a linear
transformation \( A : F^n \mapsto F^m \), and \( B \) is a \( k \times m \) matrix (determining a linear transformation \( B : F^m \mapsto F^k \)), then there is a \( k \times n \) matrix \( C \) (determining a linear transformation \( C : F^n \mapsto F^k \)), which is the composition of \( B \) and \( A \), that is, such that \( C(x) = B(A(x)) \) for all \( x \in F^n \). The matrix \( C \) is called the product of the matrices \( B \) and \( A \), and written \( BA \).

An explicit formula for the entries of \( C = BA \) in terms of those of \( A \) and \( B \) is easily obtained by considering what happens to the standard unit vector \( e_j \in F^n \). The element \( C_{ij} \) is the \( i \)th component of \( C(e_j) \). The vector \( A(e_j) \) equals \( \langle A_{1j}, \ldots, A_{mj} \rangle \); the \( i \)th component of \( B(A(e_j)) \) thus equals \( B_{1i}A_{1j} + \cdots + B_{mi}A_{mj} \). Using the summation notation,

\[
C_{ij} = \sum_{l=1}^{m} B_{il}A_{lj}.
\]

Suppose \( A \) is an \( m \times n \) matrix; the above convention of writing \( Ax \) for \( A(x) \) amounts to considering the vector \( x \) to be a \( n \times 1 \) matrix. By definition, the \( i \)th component of \( A(x) \) is \( \sum_j a_{ij}x_j \), which is exactly the same as the \( i \)th component of \( Ax \) when \( x \) is considered an \( n \times 1 \) matrix and \( Ax \) an \( m \times 1 \) matrix. Matrices with only one column are called column vectors, and matrices with only one row are called row vectors.

The \( n \times n \) matrix \( I \) where \( I_{i,j} = 1 \) if \( i = j \) else 0, is called the identity matrix; this represents the linear transformation \( I(x) = x \) on \( F^n \).

Matrix multiplication is associative, that is, \((AB)C = A(BC)\) when either side (and hence both sides) is defined. This follows since composition of functions is associative, or may be verified directly. The laws \( A(B + C) = AB + AC \) and \((B + C)A = BA + CA\), and \((aA)B = A(ab) = a(AB)\), hold whenever the multiplication is defined. The \( n \times n \) matrices form a (noncommutative) ring with the operations of matrix addition and multiplication; the multiplicative identity is the identity matrix. This ring is isomorphic to the ring of linear transformations with composition as the multiplication operation; and is in fact an \( F \)-algebra (i.e., the multiplication operation is defined on a vector space over \( F \)).

A matrix \( M \) is called invertible if it is invertible in the multiplicative monoid of the matrix ring, that is, if there is a matrix \( N \) such that \( MN = NM = I \). As usual the inverse is denoted \( M^{-1} \).

Another operation defined on matrices is the transpose; if \( M \) is an \( m \times n \) matrix the transpose \( M^t \) of \( M \) is the \( n \times m \) matrix where \( M_{i,j}^t = M_{j,i} \). Matrix transpose obeys the identities \( (M + N)^t = M^t + N^t \) and \( (MN)^t = N^t M^t \).

An important operation on vectors is the dot product (or inner product or scalar product). It has two vector arguments and a scalar value. It is quite useful for making computations with vectors shorter and more transparent. The dot product of two vectors \( x = (x_1, \ldots, x_n) \) and \( y = (y_1, \ldots, y_n) \) in \( F^n \), denoted \( x \cdot y \), is defined to be \( x_1y_1 + \cdots + x_ny_n \). This is in fact a special case of matrix multiplication, if it is rewritten as \( x^t y \). It obeys the identities \( x \cdot (cy + dz) = c(x \cdot y) + d(x \cdot z) \) and \( (cy + dz) \cdot x = c(y \cdot x) + d(z \cdot x) \). These are referred to as the bilinearity of the operation, since they state that the function is linear in each argument, when the other is held fixed.

The determinant of a matrix is a scalar derived from the matrix which has many useful properties. Recall from chapter 5 that a permutation of a set \( S \) is a bijection; the permutations of \( S \) form a group under composition. Define a transversal of an \( n \times n \) matrix \( M \) to be a set of positions, with exactly one position in each row and column. If \( \sigma \) is a permutation of \( \{1, \ldots, n \} \) then the set of positions \( \{ \langle i, \sigma(i) \rangle \} \) is clearly a transversal; on the other hand every transversal is of this form. In the rest of this appendix, unless otherwise specified by matrix we mean \( n \times n \) matrix, and by permutation a permutation of \( \{1, \ldots, n \} \); \( M, N \) will denote matrices, and Greek letters permutations. A matrix corresponds to a linear transformation of \( F^n \), which may also be referred to as \( M \).
The determinant \( \det(M) \) of an \( n \times n \) matrix \( M \) is defined to be

\[
\det(M) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n M_{i, \sigma(i)},
\]

where \( S_n \) is the symmetric group of permutations of \( \{1, \ldots, n\} \) and \( \text{sgn}(\sigma) \) is the sign of \( \sigma \), as defined in chapter 5. That is, the determinant is the sum over the transversals of the signed products of the elements selected by the transversal, where the sign is + or − according to the sign of the permutation of the transversal.

For example if \( n = 2 \) then \( \det(M) = M_{11}M_{22} - M_{12}M_{21} \), and if \( n = 3 \),

\[
\det(M) = M_{11}M_{22}M_{33} - M_{11}M_{23}M_{32} + M_{12}M_{21}M_{33} - M_{12}M_{23}M_{31} + M_{13}M_{21}M_{32} - M_{13}M_{22}M_{31}.
\]

Note that there are \( n! \) permutations, so there are \( n! \) terms in the sum. Note also that the determinant may be considered as a polynomial in the \( n^2 \) variables \( M_{ij} \).

**Lemma 1.**

\[
\prod_{i=1}^n \prod_{k=1}^n x_{ik} = \sum_{\pi} \prod_{i=1}^n x_{i, \pi(i)}.
\]

**Proof:** This is an immediate consequence of the distributive law; to form the product of the sums, arrange the terms in a square, with the rows being the terms of a sum. Then sum the products along each transversal. A more tedious proof can be given by induction.

**Theorem 2.** \( \det(MN) = \det(M) \det(N) \).

**Proof:** Let \( L = MN; \) then

\[
\det(L) = \sum_{\sigma} \text{sgn}(\sigma) \prod_{i=1}^n L_{i, \sigma(i)} = \sum_{\sigma} \text{sgn}(\sigma) \prod_{i=1}^n \sum_{k=1}^n M_{ik} N_{k, \sigma(i)}
\]

\[
= \sum_{\sigma} \text{sgn}(\sigma) \prod_{i=1}^n M_{i, \pi(i)} N_{\pi(i), \sigma(i)} = \sum_{\sigma} \prod_{i=1}^n \text{sgn}(\sigma) M_{i, \pi(i)} N_{i, \sigma^{-1}(i)}
\]

\[
= \sum_{\nu} \prod_{i=1}^n M_{i, \pi(i)} N_{i, \nu(i)} = \sum_{\nu} \prod_{i=1}^n \text{sgn}(\nu) N_{\nu(i)}
\]

\[
= (\prod_{i=1}^n \text{sgn}(\pi) M_{i, \pi(i)}) \left( \sum_{\nu} \prod_{i=1}^n \text{sgn}(\nu) N_{\nu(i)} \right) = \det(L) \det(N).
\]

An \( n \times n \) matrix \( M \) is called upper triangular if \( M_{ij} = 0 \) for \( i < j \); lower triangular if \( M_{ij} = 0 \) for \( i > j \); and diagonal if \( M_{ij} = 0 \) for \( i \neq j \). It is easily seen that for any of these matrices, the only transversal which does not contain a 0 is the diagonal (identity permutation), and hence the determinant is the product of the diagonal entries.

For a permutation \( \sigma \) let \( P_\sigma \) be the matrix with \( P_{\sigma(i), \sigma(i)} = 1 \) and \( P_{\sigma(i), \sigma(j)} = 0 \) for \( j \neq \sigma(i) \). Such a matrix is called a permutation matrix. Multiplying a matrix \( M \) by a permutation matrix \( P_\sigma \) on the left moves row \( i \) to row \( \sigma(i) \); multiplying on the right moves column \( i \) to column \( \sigma(i) \). It is clear that \( \det(P_\sigma) = \text{sgn}(\sigma) \); the only transversal not containing a 0 entry is that corresponding to \( \sigma \).

Let \( E_{ij} \) be the matrix where \( (E_{ij})_{ij} = 1 \) and which has 0’s elsewhere. If \( E = I + cE_{ij} \) where \( j \neq i \), multiplying a matrix by \( E \) on the left adds \( c \) times row \( j \) to row \( i \); this is called an elementary row operation (the notion of an elementary operation on a system of equations considered earlier is essentially the same thing). If \( E = I + cE_{ji} \) where \( j \neq i \), multiplying a matrix by \( E \) on the right adds \( c \) times column \( j \) to column \( i \); this is called an elementary column operation. Matrices \( M, N \) are said to be similar if \( N = S^{-1}MS \) for some invertible matrix \( S \).

426
Proof:
Part a follows by the remarks above concerning permutation matrices. If \( E = I + cE_{ij} \) where \( j \neq i \), the only transversal of \( E \) which does not contain a 0 is the diagonal, so the determinant of \( E \) is 1, proving part b. If \( E \) is a matrix such that \( E_{ij} = 1 \) if \( i = j \) and 0 otherwise, except \( E_{kk} = c \) for some \( k \), multiplying a matrix by \( E \) on the left multiplies row \( k \) by \( c \); multiplying on the right multiplies column \( k \) by \( c \). The determinant of \( E \) is \( c \), proving part c. Parts d and e are immediate.

Parts a and b provide an easy way to compute the determinant. Using Gaussian elimination (and permutations if necessary), an \( n \times n \) matrix may be placed in lower triangular form without changing the determinant. For a \( 2 \times 2 \) matrix it is simpler to use the formula, and for a \( 3 \times 3 \) matrix the next theorem provides an alternative.

A minor of a matrix \( M \) is defined to be a matrix obtained by keeping only the entries in certain rows and columns. More formally, if \( S \subseteq \{1, \ldots, n\} \) is of cardinality \( k \) let \( \rho_S \) be the unique increasing function from \( \{1, \ldots, k\} \) to \( \{1, \ldots, n\} \) whose range is \( S \), so that \( \langle \rho_S(1), \ldots, \rho_S(k) \rangle \) is the elements of \( S \) in increasing numerical order. If \( M \) is an \( m \times n \) matrix, \( S \subseteq \{1, \ldots, m\} \) is of cardinality \( s \), and \( T \subseteq \{1, \ldots, n\} \) is of cardinality \( t \), let \( M_{ST} \) be the \( s \times t \) matrix \( N \) where \( N_{ij} = M_{\rho_S(i), \rho_T(j)} \). A minor of \( M \) is exactly an \( M_{ST} \) for some \( S, T \).

Choosing a particular \( i, j \) the sum of the terms which contain \( M_{ij} \) may be written as

\[
M_{ij} \sum_{\sigma, \sigma(i)=j} \sg(\sigma) \prod_{1 \leq l \leq n, l \neq i} M_{i, \sigma(l)}.
\]

The other factor to \( M_{ij} \) is called its cofactor. Letting \( i \) denote \( \{1, \ldots, n\} - \{i\} \) we have the following.

Lemma 4. If \( M \) is an \( n \times n \) matrix then

\[
\sum_{\sigma, \sigma(i)=j} \sg(\sigma) \prod_{1 \leq l \leq n, l \neq i} M_{i, \sigma(l)} = (-1)^{i+j} \det(M_{ij})
\]

Proof: A \( \sigma \) such that \( \sigma(i) = j \) corresponds in an obvious way to a permutation \( \tau \) of \( \{1, \ldots, n-1\} \), namely the unique \( \tau \) such that \( \sigma(\rho_i(k)) = \rho_j(\tau(k)) \) for \( k = 1, \ldots, n-1 \). To prove the lemma it is only necessary to show that \( \sg(\sigma) = (-1)^{i+j} \sg(\tau) \). For this, let \( \beta(1) = i, \beta(k) = k-1 \) for \( 2 \leq k \leq i \), and \( \beta(k) = k \) for \( i < k \leq n \); clearly \( \sg(\beta) = (-1)^{i-1} \). Let \( \gamma \) be the similarly defined function for \( j \), so that \( \sg(\gamma) = (-1)^{j-1} \). Clearly \( \sg(\tau) = \sg(\gamma^{-1} \sigma \beta) \). For any permutation \( \gamma \sg(\gamma^{-1}) = \sg(\gamma) \) (because their product is +1), so \( \sg(\tau) = (-1)^{i+j-2} \sg(\sigma) \).

That is, the cofactor of an entry is, up to sign, the determinant of the minor obtained by deleting the row and column of the entry. Now, fixing a row \( i \), for each term of the determinant there is one and only one \( j \) such that the term contains \( M_{ij} \); similarly if \( j \) is fixed the terms may be grouped according to \( i \). We have thus proved the following, called expansion by minors of \( \det(M) \).

Theorem 5. If \( M \) is an \( n \times n \) matrix then for any row \( i \)

\[
\det(M) = \sum_{j=1}^{n} (-1)^{i+j} M_{ij} \det(M_{ij}).
\]
and for any column $j$

$$\det(M) = \sum_{i=1}^{n} (-1)^{i+j} M_{ij} \det(M_{ji}).$$

**Theorem 6.** Suppose $M$ is an $n \times n$ matrix, and $A$ is defined by

$$A_{ij} = (-1)^{i+j} \det(M_{ji}).$$

Then $AM = MA = \det(M)I$.

**Proof:** The $i, i$ entry of $MA$ is given by the sum of lemma 3 for row $i$, so equals $\det(M)$. If $j \neq i$ the $i, j$ entry of $AM$ is given by the sum of lemma 3, for row $j$ if it were made equal to row $i$. Since the determinant of a matrix with two equal rows is 0, the sum is 0 and we have proved that $AM$ equals $\det(M)I$. A similar argument involving columns proves that $MA$ does also.

The matrix $A$ of the theorem is called the adjugate of $M$. Its $i, j$ entry is the cofactor of $M_{ji}$. Note that if $M$ is invertible then then $M^{-1} = (1/\det(M))A$. Thus, given an equation $Mx = b$, the solution is given by $x = M^{-1}b = Ab/\det(M)$. By expansion by minors, $x_i = \det(M_{ij})/\det(M)$, where $M_{ij}$ is obtained from $M$ by replacing column $i$ by $b$. This is called Cramer’s rule.

The next two lemmas apply to rectangular matrices. Recall that the row space of an $m \times n$ matrix $M$ is the subspace of the domain space $F^n$ generated by the rows; similarly the column space is the subspace of the codomain space $F^m$ generated by the columns. The column space is also called the image space, since it equals $\{Mx : x \in F^n\}$. It is readily verified that $\{x : Mx = 0\}$ is a subspace of $F^n$; it is called the kernel space or right nullspace. The next lemma shows that the dimensions of the row and column spaces are the same; recall that the dimension of the row space is called the rank.

**Lemma 7.** The dimensions of the row space and of the column space of an $m \times n$ matrix $M$ are equal.

**Proof:** Apply Gauss-Jordan elimination to place $M$ in standard form. We already know that this does not change the dimension of the row space. It does not change the dimension of the column space either. Indeed, if $I$ is the coefficients of the linear combination, written as a vector, and row $r_i$ is replaced by $r_i + cr_j$, then if $r_i \cdot 1 = 0$ then $r_j \cdot 1 = 0$ iff $(r_i + cr_j) \cdot 1 = 0$. If $M$ is in standard form the claim of the lemma is clear; there are $r$ nonzero rows where $r$ is the rank. The first $r$ columns are linearly independent, and the remaining columns are linear combinations of these.

**Lemma 8.** The sum of the dimensions of the image space and the kernel space of an $m \times n$ matrix $M$ is $n$.

**Proof:** Let $K \subseteq F^m$ be the kernel space. Let $u_1, \ldots, u_r$ be a basis for the image space, and let $v_1, \ldots, v_r$ be vectors of $F^n$ which map to them. Let $w_1, \ldots, w_t$ be a basis for $K$. We claim that $B = \{v_1, \ldots, v_r, w_1, \ldots, w_t\}$ is a basis for $F^n$, which proves the lemma. First, no linear combination of $v_1, \ldots, v_r$ is a member of $K$, else the corresponding linear combination of $u_1, \ldots, u_r$ would be 0 in the image space. It follows that $B$ is linearly independent, because a linear dependency would result in a linear combination of the $v_i$ in $K$. Now, given $x \in F^n$, find the linear combination of the $u_i$ equaling $M(x)$; let $y$ be the corresponding linear combination of the $v_i$. Clearly, $M(y) = M(x)$, whence $x - y \in K$, and is a linear combination of the $w_j$. Since $x$ was arbitrary, it follows that $B$ generates $V$, and so is a basis.

Summarizing, the dimension of the row and column spaces of an $m \times n$ matrix are the same, and this value is called the rank of the matrix. The dimension of the right nullspace (the space of solutions to the homogeneous system $Ax = 0$) is called the nullity of the matrix. The sum of the rank and the nullity equals $n$. 

428
Theorem 9. The following are equivalent, for an \( n \times n \) matrix \( M \).

a. \( M \) is surjective;

b. the dimension of the image space is \( n \);

c. \( M \) is one-to-one;

d. the dimension of the kernel space is 0;

e. \( M \) is a one-to-one correspondence;

f. the rank of \( M \) is \( n \);

g. \( M \) is invertible;

h. \( \det(M) \neq 0 \).

Proof: That a is equivalent to b is immediate from the definition of the image and the fact that a subspace is of dimension \( n \) iff it is the whole space. That c is equivalent to d follows from the definition of the kernel and the fact that a linear transformation is one-to-one iff the only vector it maps to 0 is 0. That b is equivalent to d follows since the sum of the dimensions of the image space and kernel space is \( n \). Part e implies parts a and c, and since these are equivalent follows from either of them. Part f is a restatement of part b. Now, a function between sets is a one-to-one correspondence iff it has an inverse function. If an inverse matrix exists, clearly an inverse function does; this shows that part g implies part e. If an inverse function exists, it is in fact a linear transformation; its matrix is the inverse matrix, showing that part e implies part g. To see that the inverse function \( g \) to a linear transformation \( f \) is a linear transformation, suppose \( g(\mathbf{y}_1) = \mathbf{x}_1 \) and \( g(\mathbf{y}_2) = \mathbf{x}_2 \); then \( f(a\mathbf{x}_1 + b\mathbf{x}_2) = a\mathbf{y}_1 + b\mathbf{y} + 2 \), so \( g(a\mathbf{y}_1 + b\mathbf{y}_2) = a\mathbf{x}_1 + b\mathbf{x} + 2 \). For part h, if \( M \) is invertible \( \det(M) \) cannot be 0, else \( \det(I) = \det(M)\det(M^{-1}) = 0 \) which is false; conversely theorem 3 shows that if \( \det(M) \neq 0 \) then \( M \) is invertible.

The rank of an \( m \times n \) matrix \( M \) can also be characterized as the largest \( r \) such that \( M \) has a nonsingular \( r \times r \) minor. If the rank is \( r \) choose any \( r \) linearly independent rows. The resulting minor has rank \( r \), so has \( r \) linearly independent columns, and these give a nonsingular \( r \times r \) minor. On the other hand any \( s > r \) rows of \( M \) are linearly dependent, and deleting columns does not change this.
Appendix 3. Basic graph theory.

There are many conventions in use concerning graphs. We begin by choosing some basic conventions, and discuss variations below. In chapter 13 the conventions used in category theory were given; in the remainder of the text the conventions given here are used.

An (undirected) graph \( G \) is a pair \( (V, E) \) where \( V \) is a finite set of vertices, and \( E \) is a set of unordered pairs (two element subsets) from \( V \), called edges. A directed graph is a finite set \( V \) of vertices, and a set of ordered pairs \( E \subseteq V \times V \), called directed edges. We have allowed a directed edge from a vertex to itself; but not an undirected edge.

Graphs are very useful objects; they clearly provide a model for networks, such as airline routes, communication networks, or electrical networks. They are also mathematically interesting, the branch of discrete mathematics known as graph theory being devoted to their study. Graphs have provided endless problems in recreational mathematics. An early paper in graph theory was written by the great Swiss mathematician Leonhard Euler in 1736. In it Euler solved the “Konigsberg bridge problem”, whether it was possible to walk along seven bridges between two islands in, and the banks of, the Pregel River in Konigsberg, so as to cross each bridge exactly once. Euler showed that it was not by giving necessary and sufficient conditions on a graph, that such a walk be possible.

Graphs have a “pictorial” quality; indeed several packages for drawing them are available within the TeX typesetting system. Edges may be drawn as lines, with directed edges being equipped with an arrowhead; figure 1 is an example. It may or may not be possible to draw a graph in the plane without crossing edges. (It does not matter whether edges are required to be drawn as straight lines, although this is a nontrivial fact). A graph which can be drawn without crossing edges is called planar; a simple example of a graph which is not planar is the graph with 5 vertices and an edge between every pair of vertices.

![Figure 1](image-url)

Variations in usage concerning graphs include the following. In undirected graphs, self loops may be allowed, where a self loop is an edge between a vertex and itself. Multiple edges between two vertices may be allowed; these are often called multigraphs. Infinite graphs are considered, but most of graph theory is concerned with finite graphs. Vertices are also called points or nodes, and edges lines.

For the remainder of the appendix only undirected graphs will be considered. The two vertices which an edge connects are called its endpoints; these are said to be incident to the edge. The endpoints of an edge are said to be adjacent or to be neighbors. Two distinct edges are called adjacent if they are incident to a common vertex. The number of edges incident to a vertex is called its degree. A vertex may have degree 0, in which case it is called isolated.

A path in a graph may be defined as a sequence of vertices \( p_0, \ldots, p_n \) such that for \( 0 \leq i < n \) vertices \( p_i \) and \( p_{i+1} \) are adjacent. This is the simplest definition, but clearly we are thinking of the edges as comprising the path. The path is said to traverse the edges between \( p_i \) and \( p_{i+1} \), in order of increasing \( i \). Note that two consecutive edges are either incident or identical, in the latter case the edge being traversed in first one direction, then the other. The number \( n \) is called the length of the path; it is the number of edges traversed by the path, counting an edge once each time it is traversed. The case \( n = 0 \) is allowed, and the path is said
to be trivial in this case.

The vertices $p_0$ and $p_n$ are called the endpoints of the path, and the path is said to lead from $p_0$ to $p_n$. If no vertex occurs more than once the path is called simple. If there is a path from $a$ to $b$ then there is a simple path from $a$ to $b$. Indeed, suppose some vertex $c$ occurs as $p_i$ and $p_j$ for $j > i$; the entire sequence of vertices $p_{i+1}, \ldots, p_j$ may be removed and the result is still a path, with the same endpoints. The process may be repeated until a simple path is obtained. The same argument shows that a path of shortest length between two vertices is simple. A nontrivial path from a vertex to itself, which visits no other vertex more than once is called a cycle.

A subgraph of a graph $G$ is a subset of the vertices, together with some of the edges between them; if all the edges are included the subgraph is called full. A path may be considered a subgraph, by including the edges from $p_i$ to $p_{i+1}$.

A graph is called connected if there is a path between any two vertices. If there are paths from $p$ to $q$ and from $q$ to $r$ then clearly there is one from $p$ to $r$. It follows that the vertices of a graph may be grouped into classes, where in each class there is a path between any two vertices, and there are no paths (in fact no edges) between vertices of different classes. The full subgraph on each class of vertices is a maximal connected subgraph, called a connected component of $G$; figure 2 gives an example, with two components.

![Figure 2](image1)

A graph is called a tree if there is exactly one path between any pair of vertices; figure 3 is an example. A graph is a tree if it is connected and contains no cycles, since there is a cycle containing two vertices iff there are distinct paths between them. A vertex of degree 1 in a tree is called a leaf of the tree. There must be at least two of these, unless the tree consists of a single isolated vertex. To see this, start at any vertex; repeatedly follow an edge out of the current vertex, other than the one by which it was reached. Eventually a leaf must be reached. If the original vertex was a leaf, we have found at least one other, and if not, there are at least two edges incident to it so we have found at least two leaves. If a tree contains $n$ vertices it contains $n - 1$ edges. This is clear if $n = 1$; otherwise there is a leaf. Removing the edge containing the leaf reduces both the number of edges and the number of vertices by 1.

![Figure 3](image2)

A subgraph of a connected graph is called a spanning tree if it is a tree, and contains each vertex. That is, it is a set of edges such that each vertex is incident to some edge, where the edges comprise a tree; see figure 4.

**Theorem 1.** Every connected graph has a spanning tree.
**Proof:** Place some vertex in the set $S$ of processed vertices, and repeat the following step. If there are any unprocessed vertices, take those which are adjacent to some vertex in $S$; for each, add an edge to some vertex of $S$; and then add all these vertices to $S$. At each step, the edges which have been added so far form a tree, because the new edges are incident to new vertices and so cannot lie on a cycle. Each vertex must eventually be processed, because it is connected to the start vertex by some path.

Suppose $T$ is a spanning tree in a connected graph $G$. An edge $e$ of $G$ not in $T$ is called a chord of $T$; there are $e - n + 1$ of these. Adding $e$ to $T$ yields a subgraph containing a unique cycle, namely that consisting of the chord and the path between its two endpoints (any cycle must contain the chord). The cycle determined by a chord is called a fundamental cycle of $T$. Note that these are distinct for distinct chords.

As an example of algebraic facts concerned with graphs, consider the vector space over $\mathbb{Z}_2$ with basis $E$; its elements are the 1-chains of the graph. Similarly the elements of the vector space with basis $V$ are the 0-chains. The linear operator which assigns to an edge $\{u, v\}$ the chain $u + v$ is the 1-boundary operator. A 1-chain with empty boundary is called by various names, cycle vector for example (since the term cycle is already in use). One may verify that the fundamental cycles form a basis for the space of cycle vectors.
Bibliography


[HardWi] G. Hardy and E. Wright, An Introduction to the Theory of Numbers, Oxford University Press.


45 (1992), 231–239.
1955.
Form” International Symposium on Symbolic and Algebraic Computation, B. Mourrain (ed.), ACM, 
2001, pp. 231-236.
[Milne] J. Milne, Algebraic Geometry, Algebraic Number Theory, Algebraic groups and Arithmetic groups, 
preprints.
98–114.
of Symbolic Logic 4 (1939), 53–60.
[Schikhof] W. H. Schikhof, Ultrametric Calculus, an Introduction to p-adic Analysis, Cambridge University 
### Index

- Abelian Lie algebra, 393
- absolute value, 241
- action, 27
- addition in a Category, 145
- additive functor, 146
- adjoint operator, 79
- adjoint representation, 393
- adjugate, 428
- affine algebraic set, 357
- affine halfspace, 284
- affine hyperplane, 284
- affine space, 357
- affine subspace, 284
- Alexandrov topology, 278
- algebra, 55
- algebra, free, 55
- algebraic closure, 239
- algebraic closure system, 18
- algebraic element, 62
- algebraic extension, 62
- algebraic independence, 239
- algebraic integer, 378
- algebraic integer ring, 378
- algebraic lattice, 260
- algebraic number field, 378
- algebraic poset, 268
- algebraic set, 355
- alternating multilinear form, 54
- alternating tensor algebra, 209
- amalgamated product, 138
- analytic function, 332, 389
- annihilator, 54
- anti-homomorphism, 30, 162
- antilinear, 75
- antisymmetric multilinear form, 54
- antisymmetric tensor, 210
- approximating auxiliary order, 278
- arrow, 126
- ascending chain condition, 16
- associated sheaf, 339
- atlas, 200
- augmentation, 347
- augmentation arrow, 227
- augmentation preserving, 347
- augmented complex, 227
- automorphism, 9
- automorphism, inner, 24
- automorphism, outer, 24
- auxiliary order, 278
- Baire space, 325
- balanced module, 164
- Banach space, 329
- barycenter, 349
- base for topology, 184
- basic solution, 288
- basis, of module, 55
- Berlekamp's algorithm, 413
- Betti number, 222
- bidual, of ring, 164
- bilinear, 54
- binomial coefficients, 5
- biproduct, 144
- block of imprimitivity, 151
- Boolean algebra, 19
- bound variable, 90
- boundary chain, 219
- braid relation, 307
- Buchberger's algorithm, 375
- Buchberger's criterion, 374
- Caratheodory's theorem, 287
- cardinal, 418
- cardinality, 6
- Cartesian subalgebra, 394
- Cartesian closed category, 204
- category, 126
- category, Abelian, 144
- category, concrete, 127
- category, exact sequence, 143
- category, opposite, 129
- category, small, 127
- category, standard, 126
- Cauchy sequence, 195
- center, of group, 24
- center, of ring, 32
- centralizer, 149
- chain, 222
- chain complex, 219
- chain complex over an object, 227
- chain complex under an object, 227
- chain contractible, 221
- chain homotopy, 221
- chain map, 219
- chain, in a poset, 15
- character on a commutative group, 156
- character on a finite field, 159
- character, of group representation, 166
- characteristic polynomial, 72
- characteristic roots, 72
- characteristic subgroup, 152
- chart, 200
- Cholesky factorization, 84
- Church's thesis, 102
- class choice, 139
- class equation, 149
division ring, 32
domain, 3
dot product, 425
double complex, 230
double coset, 149
double of a real algebra, 256
dual basis, 70
dual space, 70
dual, in category theory, 129

edge, of graph, 430
effective procedure, 105
eigenspace, 73
eigenvalue, 72
eigenvector, 73
eigenspace, defective, 73
eigenvalue, 73
elementarily equivalent, 96
elementary column operation, 426
elementary divisors, 60
elementary embedding, 96
elementary functions, 108
elementary operation, 420
elementary row operation, 69, 426
embedding, 9
embedding, over a subfield, 64
endomorphism, 9
elementary operation, 106, 122
endomorphism, 9
finite extension, 62
finite intersection property, 194
finite presentation, 106, 122
first countable space, 188
first Noether isomorphism, 219
first order language, 89
first order logic, 92
flat module, 233
FOL, 92
formula of first order logic, 89
formula, $\Pi_0$, 97
formula, $\Sigma_0$, 97
formula, atomic, 89
formula, Horn, 91
formula, open, 90
formula, positive, 91
fractional ideal, 251
frame, 267
frame in vector bundle, 338
free resolution, 227
free variable, 90
full functor, 128
function, 3
function, bijective, 3
function, ceiling, 5
function, characteristic, 4
function, delta, 4
function, floor, 5
function, injective, 3
function, partial, 4
function, surjective, 3
functor, 127
fundamental group, 343
fundamental groupoid, 343
fundamental region, 296
fundamental theorem of linear inequalities, 286
fundamental theorem of linear programming, 288
ideal, in a poset, 97
ideal, left, 53
ideal, right, 53
idempotent ring element, 163
identity element, 21
identity matrix, 425
image, 142
improper isometry, 312
incompleteness theorem, first, 113
incompleteness theorem, second, 113
index, 23
induced topology, 185
inductive partial order, 16
linear program, dual, 290
linear program, primal, 290
inertia group, 384
inertia, of matrix, 78
initial object, 129
injection, of matrix, 78
injective object, 223
injective resolution, 227
inner product, 76, 425
integer lattice, 295
integral closure, 245
integral element, 244
integral extension, 244
integrally closed, 245
interior, 185
interpretation, 114
invariant submodule, 71
invariants, 74
inverse, 21
inverse limit, 136
inverse system, 136
invertible ideal, 252
invertible matrix, 425
involution in an algebra, 256
irreducible polynomial, 62
irreducible space, 271
irredundant collection, 356
irrelevant ideal, 363
isometry, 198
isomorphic, 10
isomorphism, 9, 126
isomorphism, second Noether, 53
Jordan measurable set, 299
kernel, 136
kernel chain, 219
Killing form, 393
KKT conditions, 290
Kleene T predicate, 105
Krull dimension, 360
null vector, 76
nullity, 428

object, 126
OFOL, 96
open function, 184
open mapping theorem, 330
open set, 184
orbit, 27, 29
order, of element, 22
order, of group, 22
ordered commutative ring, 37
ordered field, 409
ordered real vector space, 305
ordinal, 417
orthogonal group, 257
orthogonal vectors, 77
ortholattice, 263
orthonormal vectors, 77
Ostrowski's theorem, 253
overdetermined, 422
overloading, 3

$p$-adic valuation, 243
$p$-group, 149
$p$-subgroup, 150
paracompact space, 188
parallelogram law, 75
parameters in a formula, 96
partial derivative, 341
partial order, 15
partial recursive function, 105
particular solution, 423
partition of unity, 190
partition, of set, 4
path connected space, 192
path, in graph, 430
Peano arithmetic, 113
permutation group, 27
permutation matrix, 426
plane crystallographic group, 312
point lattice, 295
polar cone, 286
polar convex set, 292
polar decomposition, 85
polarization identity, 75
polyhedral cone, 287
polyhedron, 286
polynomial function, 358
polynomial time, 104
polynomial, degree of, 42
polytope, 291
positive chain complex, 222
positive definite bilinear form, 76
positive system of roots, 305
power series, 332
predicate, 3

prenex normal form, 97
preorder, 15
Presburger arithmetic, 403
preserves, 9
presheaf, 338
primary decomposition, 249
primary decomposition, reduced, 249
primary ideal, 249
primary submodule, 249
prime filter, in a poset, 98
prime ideal, in a poset, 98
prime spectrum of a ring, 370
prime subfield, 62
prime subring, 32
primitive element, of finite field, 66
primitive permutation group, 151
primitive recursive arithmetic, 113
primitive recursive function, 107
principal bundle, 337
principal ideal, in a poset, 97
product, 136
product, Cartesian, 3, 6
product, of structures, 12
projection map, 13
projection onto subspace, 77
projection operator, 81
projection, of product, 136
projective object, 223
projective resolution, 226
projective space, 363
proper class, 417
proper isometry, 312
pseudo-complement, 264
pseudo-complement, relative, 263
pseudo-metric space, 195
pseudo-metrizable space, 197
pullback, 136

quadratic algebra, 255
quadratic form, 75
quadratic Hensel lift, 408
quadratic surd, 316
quasi-affine algebraic set, 358
quasi-projective algebraic set, 365
quaternions, 257
quotient, 10, 186
quotient space, 186
quotient, in a category, 144
stalk, 339
standard unit vectors, 55, 423
strong deformation retract, 351
strong norm, 327
stronger topology, 184
strongly locally compact space, 277
structure, 8
structure constants, 56
structure group, 337
structure, domain of, 8
Sturm’s theorem, 411
subbase for topology, 184
subfield, 62
submanifold, 340, 391
submodule, 52
subresultant, 406
subspace topology, 186
substructure, 8
summand, 56, 224
supporting hyperplane, 285
suspension functor, 345
Sylow subgroup, 150
symmetric function, 45
symmetric multilinear form, 54
symmetric polynomial, 48
symmetric tensor, 210
symmetric tensor algebra, 209
system of imprimitivity, 151
syzygy, 375

$T_0$ space, 188
$T_1$ space, 189
$T_2$ space, 189
$T_3$ space, 189
$T_4$ space, 189
tangent bundle, 341
tensor algebra, 209
tensor product, 205
term algebra, 96
terminal object, 129
topological group, 319
topological property, 188
topological space, 184
topological vector space, 283, 319
topology, 184
Tor functor, 234
torsion, 54
totally bounded space, 195
totally disconnected, 193
tower, 62

trace, on extension field, 158
transcendence base, 239
transcendence degree, 240
transfinite induction, 6, 417
transitive, 29
transitive set, 417
transitive, $t_-$, 29
transitive, sharply $t_-$, 29
translation invariance, 298
transpose, 4
transversal, 425
tree, 431
triangular form, 69
triangular identities, 134
tuple, of elements, 3
Turing machine, 103
Tychonoff space, 189

ultrafilter, 99
ultrametric inequality, 242
ultrametric metric space, 324
ultraproduct, 99
unbounded linear program, 288
undecidable predicate, 106
underdetermined, 422
uniform norm, 329
uniformly continuous, 198
unimodular matrix, 295
unit vector, 77
unit, of ring, 31
unitary group, 257
universal covering projection, 397
universal Turing machine, 105
unsatisfiable set of sentences, 96

valuation, 243
valuation ring, 243
valuation, real, 243
value group, 243
vector bundle, 335
vector space, 162
vertex of polyhedron, 287
vertex, of graph, 430

weaker topology, 184
Wedderburn’s theorem, 155
well-ordering principle, 16

Zariski topology, 357
zero object, 129
Zorn’s lemma, 16