1

Algebraic Fundamentals

1.1 Sets and Numbers

Perhaps the basic concept in mathematics is that of a **set**, an unambiguously determined collection of mathematical entities. A set can be specified by listing its contents, as for instance the set $A = \{w, x, y, z\}$, or by describing its contents in any other way so long as the members of the set are fully determined. Among all sets the strangest is no doubt the **empty set**, traditionally denoted by $\emptyset$, the set that contains nothing at all, so the set $\emptyset = \{\ \}$; this set should be approached with some caution, since it can readily sneak up unexpectedly as an exception or counterexample to some mathematical statement if that statement is not carefully phrased. There is a generally accepted standard notation dealing with sets, and students should learn it and use it freely. That a belongs to or is a member of the set A is indicated by writing $a \in A$; that b does not belong to the set A is indicated by writing $b \notin A$. A set A is a **subset** of a set B if whenever $a \in A$ then also $a \in B$, and the condition that A is a subset of B is indicated by writing $A \subset B$ or $B \supset A$. There is some variation in the usage though. With the definition adopted here, $A \subset B$ includes the case that $A = B$ as well as the case that there are some $b \in B$ for which $b \notin A$; in the latter case the inclusion $A \subset B$ is said to be a **proper inclusion** and it is denoted by $A \subsetneqq B$ or $B \supsetneqq A$. In particular $\emptyset \subset A$ for any set A; for since there is nothing in the empty set $\emptyset$ the condition that anything contained in $\emptyset$ is also contained in A holds vacuously. Sets A and B are said to be **equal**, denoted by $A = B$ or $B = A$, if they contain exactly the same elements, or equivalently if both $A \subset B$ and $B \subset A$.

The **intersection** of sets A and B, denoted by $A \cap B$, consists of those elements that belong to both A and B, and the **union** of sets A and B, denoted by $A \cup B$, consists of those elements that belong to either A or B or both. More generally, for any collection of sets $\{A_\alpha\}$, the intersection and union of this collection of sets are defined by

$$\begin{aligned} \bigcap_\alpha A_\alpha &= \{\, a \mid a \in A_\alpha \ \text{ for all } A_\alpha \}, \\ \bigcup_\alpha A_\alpha &= \{\, a \mid a \in A_\alpha \ \text{ for some } A_\alpha \}. \end{aligned} \tag{1.1}$$

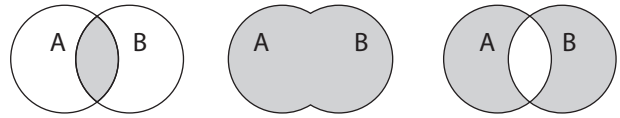

Figure 1.1. Venn diagrams illustrating the sets $A \cap B$, $A \cup B$, $A \Delta B$, respectively

Two sets A and B are **disjoint** if $A \cap B = \emptyset$. The intersection and union operations on sets are related by

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C), \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C). \end{aligned} \tag{1.2}$$

The **difference** of two sets A and B, denoted by $A \sim B$, consists of those $a \in A$ such that $a \notin B$, whether B is a subset of A or not; for clarity $\emptyset \sim \emptyset = \emptyset$, a result that might not be altogether clear from the definition of the difference. Obviously the order in which the two sets are listed is critical in this case; but there is also the **symmetric difference** between these two sets, defined by

$$A \Delta B = B \Delta A = (A \sim B) \cup (B \sim A) = (A \cup B) \sim (A \cap B). \tag{1.3}$$

The symmetric difference, as well as the intersection and union, can be illustrated by **Venn diagrams**, as in the accompanying Figure 1.1. It is clear from the definitions that

$$\begin{aligned} A \sim (B \cup C) &= (A \sim B) \cap (A \sim C), \\ A \sim (B \cap C) &= (A \sim B) \cup (A \sim C). \end{aligned} \tag{1.4}$$

If $A, B, C, \ldots$ are all viewed as subsets of a set E, and that is understood in the discussion of these subsets, then a difference such as $E \sim A$ often is denoted just by $\sim A$ and is called the **complement** of the subset A; with this understanding (1.4) takes the form

$$\begin{aligned} \sim (A \cup B) &= (\sim A) \cap (\sim B), \\ \sim (A \cap B) &= (\sim A) \cup (\sim B). \end{aligned} \tag{1.5}$$

It is worth writing out the proofs of (1.2) and (1.4) in detail if these equations do not seem evident.

A **mapping** $f : A \longrightarrow B$ from a set A to a set B associates to each $a \in A$ its image $f(a) \in B$. The set A often is called the **domain** of the mapping and the set B the **range** of the mapping. The mapping $f : A \longrightarrow B$ is said to be **injective** if $f(a_1) \neq f(a_2)$ whenever $a_1 \neq a_2$; it is said to be **surjective** if for every $b \in B$ there is $a \in A$ such that $b = f(a)$; and it is said to be **bijective** if it is both injective

and surjective. A bijective mapping $f : A \longrightarrow B$ is said to establish a **one-to-one correspondence** between the sets A and B, since it associates to each point $a \in A$ a unique point $b \in B$, and each point of B is associated in this way to a unique point of A. A mapping $f : A \longrightarrow B$ is bijective if and only if there is a mapping $g : B \longrightarrow A$ such that $g(f(a)) = a$ for every $a \in A$ and $f(g(b)) = b$ for every $b \in B$. Indeed if f is bijective then it is surjective, so each point $b \in B$ is the image $b = f(a)$ of a point $a \in A$, and f is also injective, so the point a is uniquely determined by b and consequently can be viewed as the image $a = g(b)$ of a well-defined mapping $g : B \longrightarrow A$, for which $b = f(g(b))$; and for any $a \in A$ substituting $b = f(a)$ in the preceding formula shows that $f(a) = f\big(g(f(a))\big)$, so since f is injective it follows that $a = g(f(a))$. Conversely if there is a mapping $g : B \longrightarrow A$ such that $g(f(a)) = a$ for every $a \in A$ and $f(g(b)) = b$ for every $b \in B$ then the mapping f is clearly both injective and surjective so it is bijective. The mapping g is called the **inverse mapping** to f; it is usually denoted by $g = f^{-1}$, and it is a bijective mapping from B to A. For any mapping $f : A \longrightarrow B$, not necessarily bijective or injective or surjective, there can be associated to any subset $X \subset A$ its **image** $f(X) \subset B$, defined by

$$f(X) = \Big\{ f(a) \in B \,\Big|\, a \in X \Big\}, \tag{1.6}$$

and to any subset $Y \subset B$ its **inverse image** $f^{-1}(Y) \subset A$, defined by

$$f^{-1}(Y) = \Big\{ a \in A \,\Big|\, f(a) \in Y \Big\}. \tag{1.7}$$

The image $f(X)$ may or may not coincide with the range of the mapping f. If $f : A \longrightarrow B$ is injective then clearly it determines a bijective mapping from any subset $X \subset A$ to its image $f(X) \subset B$, so in a sense it describes an injection of the set X into B. If $f : A \longrightarrow B$ is bijective the inverse image $f^{-1}(Y)$ of a subset $Y \subset B$ is just the image of that subset Y under the inverse mapping f^{-1}; it should be kept clearly in mind though that $f^{-1}(Y)$ is well defined even for mappings $f : A \longrightarrow B$ that are not bijective, so mappings for which the inverse mapping f^{-1} is not even defined. Thus the notation $f^{-1}(Y)$ really has two different meanings, and some care must be taken to distinguish them carefully to avoid confusion and errors. For any mapping $f : A \longrightarrow B$ and any subsets $X_1, X_2 \subset A$ and $Y_1, Y_2 \subset B$ it is fairly easy to see that

$$\begin{aligned} f(X_1 \cup X_2) &= f(X_1) \cup f(X_2), \\ f(X_1 \cap X_2) &\subset f(X_1) \cap f(X_2) \ \text{ but it may be a proper inclusion,} \\ f^{-1}(Y_1 \cup Y_2) &= f^{-1}(Y_1) \cup f^{-1}(Y_2), \\ f^{-1}(Y_1 \cap Y_2) &= f^{-1}(Y_1) \cap f^{-1}(Y_2); \end{aligned} \tag{1.8}$$

this is an instance in which the inverse images of sets are better behaved than the images of sets. To any mappings $f : A \longrightarrow B$ and $g : B \longrightarrow C$ there can be associated the **composite** mapping $g \circ f : A \longrightarrow C$ defined by

$$(g \circ f)(a) = g\big(f(a)\big) \quad \text{for all } a \in A. \tag{1.9}$$

The order in which the composite is written should be kept carefully in mind: $g \circ f$ is the operation that results from first applying the mapping f and then the mapping g.

To any set A there can be associated other sets, for example, the set consisting of all subsets of A, sometimes called the **power set** of A and denoted by $\mathfrak{P}(A)$. The power set $\mathfrak{P}(\emptyset)$ of the empty set is the set that consists of a single element, since the only subset of the empty set is the empty set itself; the power set of the set $\{a\}$ consisting of a single point has two elements, $\mathfrak{P}(\{a\}) = \{\{a\},\ \emptyset\}$; and the power set of the set $\{a, b\}$ consisting of two points has four elements, $\mathfrak{P}(\{a, b\}) = \{\emptyset,\ \{a\},\ \{b\},\ \{a, b\}\}$. To any two sets A, B there can be associated the set

$$B^A = \{\, f : A \longrightarrow B \,\} \tag{1.10}$$

consisting of all mappings $f : A \longrightarrow B$. If $A = \{a\}$ consists of a single point there is the natural bijection $\phi : B^{\{a\}} \longrightarrow B$ that associates to any mapping $f \in B^{\{a\}}$ its image $f(a) \in B$, since the mapping f is fully determined by its image and any point $b \in B$ is the image of the mapping f for which $f(a) = b$. On the other hand there is the natural bijection $\phi : \{a\}^B \longrightarrow \{a\}$ since there is a single mapping $f \in \{a\}^B$, the mapping for which $a = f(b)$ for every $b \in B$. If $B = \mathbb{F}_2 = \{0, 1\}$ is the set consisting of two points 0 and 1 there is the natural bijection $\psi : (\mathbb{F}_2)^A \longrightarrow \mathfrak{P}(A)$ that associates to each mapping $f : A \longrightarrow \mathbb{F}_2$ the subset $E_f = f^{-1}(1) \subset A$; for the mapping f is determined uniquely by the subset $E_f \subset A$, since $f(x) = 1$ if $x \in E_f$ and $f(x) = 0$ if $x \in A \sim E_f$, and for any subset $E \subset A$ the **characteristic function** χ_E of the set E, the mapping $\chi_E : A \longrightarrow \mathbb{F}_2$ defined by

$$\chi_E(a) = \begin{cases} 1 & \text{if } a \in E, \\ 0 & \text{if } a \notin E, \end{cases} \tag{1.11}$$

has the property that $\chi_E^{-1}(1) = E$. The bijective mapping ψ can be viewed as the identification

$$\mathfrak{P}(A) = (\mathbb{F}_2)^A; \tag{1.12}$$

sometimes the power set $\mathfrak{P}(A)$ of a set A is denoted just by $\mathfrak{P}(A) = 2^A$.

Yet a different way of associating to a set A another set is through an **equivalence relation** on the set A, a relation between some pairs of elements $a_1, a_2 \in A$ which is denoted by $a_1 \asymp a_2$ and is characterized by the following three properties:

(i) *reflexivity*, $a \asymp a$ for any $a \in A$;
(ii) *symmetry*, if $a_1 \asymp a_2$ then $a_2 \asymp a_1$; and
(iii) *transitivity*, if $a_1 \asymp a_2$ and $a_2 \asymp a_3$ then $a_1 \asymp a_3$.

If $\asymp$ is an equivalence relation on a set A then to any $a \in A$ there can be associated the set of all $x \in A$ that are equivalent to a, the subset

$$A_a = \left\{ x \in A \,\middle|\, x \asymp a \right\} \subset A, \tag{1.13}$$

which by reflexivity includes in particular a itself. The set A_a is called the **equivalence class** of a with respect to the equivalence relation $\asymp$. Similarly to any $b \in A$ there can be associated the set A_b of all $y \in A$ that are equivalent to b. If $c \in A_a \cap A_b$ for an element $c \in A$ then $c \asymp a$ and $c \asymp b$ so by symmetry and transitivity $a \asymp b$; then by symmetry and transitivity again whenever $x \in A_b$ then $x \asymp b \asymp a$ so $x \asymp a$ and consequently $x \in A_a$, hence $A_b \subset A_a$, and correspondingly $A_a \subset A_b$ so that actually $A_b = A_a$. Thus the set A is naturally decomposed into a collection of disjoint equivalence classes; the set consisting of these various equivalence classes is another set, called the **quotient** of the set A under this equivalence relation and denoted by $A/\asymp$. This is a construction that arises remarkably frequently in mathematics. For example, to any mapping $f : A \longrightarrow B$ between two sets A and B there can be associated an equivalence relation on the set A by setting $a_1 \asymp_f a_2$ for two points $a_1, a_2 \in A$ whenever $f(a_1) = f(a_2)$; it is quite clear that this is indeed an equivalence relation and that the quotient $A/\asymp_f$ can be identified with the image $f(A) \subset B$.

The notion of equivalence also is used in a slightly different way, as a relation among sets rather than as a relation between the elements of a particular set. For instance the equality $A = B$ of two sets clearly satisfies the three conditions for an equivalence relation; it is not phrased as an equivalence relation between elements of a given set but rather as an equivalence relation among various sets, to avoid any involvement with the paradoxical notion of the set of all sets. As another example, two sets A, B are simply said to be **equivalent** if there is a bijective mapping $f : A \longrightarrow B$; that these two sets are equivalent is indicated by writing $A \leftrightarrow B$. This notion also clearly satisfies the three conditions for an equivalence relation among sets. Equivalent sets intuitively are those with the same "number of elements," whatever that may mean. One possibility for giving that notion a definite meaning is merely to use the equivalence relation itself as a proxy for the "number of elements" in a set, that is, to define the

cardinality of a set A as the equivalence class of that set,

$$\#(A) = \Big\{ X \,\big|\, X \text{ is a set for which } X \leftrightarrow A \Big\}. \tag{1.14}$$

With this definition $\#(A) = \#(B)$ just means that the sets A and B determine the same equivalence class, that is, that $A \leftrightarrow B$ so that there is a bijective mapping $f : A \longrightarrow B$.

For sufficiently small sets this notion of cardinality can be made somewhat more explicit. Consider formal symbols $\{/, /, \ldots, /\}$, which are constructed beginning with the symbol $\{/\}$, followed by its **successor** $\{/, /\}$ obtained by adjoining a stroke, followed in turn by its successor $\{/, /, /\}$ obtained by adjoining another stroke, and so on; at each stage of the construction the successor of one of these symbols is obtained by adjoining another stroke to that symbol. The initial symbol $\{/\}$ represents the cardinality or equivalence class of sets consisting of the sets that can be obtained by replacing the stroke / by an element of any set; its successor $\{/, /\}$ represents the cardinality or equivalence class of sets consisting of the sets that can be obtained by replacing the strokes $/, /$ by distinct elements of any set, and so on. The cardinalities thus constructed form a set $\mathbb{N}$ called the set of **natural numbers**. The natural number that is the cardinality represented by the symbol $\{/\}$ is usually denoted by 1, so that $1 = \#(X)$ for any set X consisting of a single element; and if $n \in \mathbb{N}$ is the cardinality described by one of the symbols in this construction the cardinality represented by its successor symbol is denoted by n' and is called the **successor** of n. As might be expected, a customary notation is $2 = 1'$, $3 = 2'$, and so on. It is evident from this construction that the set $\mathbb{N}$ of natural numbers satisfies the **Peano axioms**:

(i) *origin*, there is a specified element $1 \in \mathbb{N}$;

(ii) *succession*, to any element $a \in \mathbb{N}$ there is associated an element $a' \in \mathbb{N}$, called the *successor* to a, such that

  (ii′) $a' \neq a$,
  (ii″) $a' = b'$ if and only if $a = b$, and
  (ii‴) 1 is not the successor to any element of $\mathbb{N}$;

(iii) *induction*, if $E \subset \mathbb{N}$ is any subset such that $1 \in E$ and that $a' \in E$ whenever $a \in E$ then $E = \mathbb{N}$.

The axiom of induction is merely a restatement of the fact that the set $\mathbb{N}$ is constructed by the process of considering the successors of cardinalities already in $\mathbb{N}$. A simple consequence of the axiom of induction is that every element $a \in \mathbb{N}$ except 1 is the successor of another element of $\mathbb{N}$; indeed if

$$E = 1 \cup \Big\{ x \in N \,\big|\, x = y' \;\text{ for some }\; y \in \mathbb{N} \Big\}$$

it is clear that $1 \in E$ and that if $x \in E$ then $x' \in E$ so by the induction axiom $E = \mathbb{N}$. The set $\mathbb{N}$ of natural numbers is customarily defined by the Peano axioms; for it is evident that any set satisfying the Peano axioms can be identified with the set of natural numbers, by yet another application of the axiom of induction.

Note particularly that it is not asserted or assumed that the cardinality of any set actually is a natural number. The sets with cardinalities that are natural numbers are called **finite sets**, while the sets that are not finite sets are called **infinite sets**. That a set S is infinite is indicated by writing $\#(S) = \infty$, a slight abuse of notation since it does not mean that ∞ is the cardinality of the set S but just that S is not a finite set. The set $\mathbb{N}$ itself is an infinite set. Indeed it is evident from the definition of the natural numbers that a finite set cannot be equivalent to a proper subset of itself; but the mapping $f : \mathbb{N} \longrightarrow \mathbb{N} \sim \{1\}$ that associates to each natural number n its successor n' is injective by succession and surjective by the preceding discussion, so $\mathbb{N}$ is equivalent to $\mathbb{N} \sim \{1\}$ hence $\mathbb{N}$ cannot be finite. It is traditional to set $\#(\mathbb{N}) = \aleph_0$; so if A is any set that is equivalent to $\mathbb{N}$ then $\#(A) = \aleph_0$ as well. A set A such that $\#(A) = \aleph_0$ is said to be a **countably infinite** set, while a set that is either finite or countably infinite, is called a **countable set**.[1]

Induction is also key to the technique of proof by **mathematical induction**: If $T(n)$ is a mathematical statement depending on the natural number $n \in \mathbb{N}$, if $T(1)$ is true and if $T(n')$ is true whenever $T(n)$ is true, then $T(n)$ is true for all $n \in \mathbb{N}$; indeed if E is the set of those $n \in \mathbb{N}$ for which $T(n)$ is true then by hypothesis $1 \in E$ and $n' \in E$ whenever $n \in E$, so by the induction axiom $E = \mathbb{N}$. Many examples of the application of this method of proof will occur in the subsequent discussion.

There is a natural comparison of cardinalities of sets defined by setting

$$\#(A) \leq \#(B) \text{ if there is an injective mapping } f : A \longrightarrow B. \tag{1.15}$$

This is well defined, since if $A' \leftrightarrow A$, $B' \leftrightarrow B$, and $\#(A) \leq \#(B)$ there are a bijective mapping $g : A \longrightarrow A'$, a bijective mapping $h : B \longrightarrow B'$, and an injective mapping $f : A \longrightarrow B$; and the composite mapping

$$f' = h \circ f \circ g^{-1} : A' \longrightarrow B'$$

is an injective mapping hence $\#(A') \leq \#(B')$. As far as the notation is concerned, it is customary to consider $\#(B) \geq \#(A)$ as an alternative way of writing $\#(A) \leq \#(B)$, and to write $\#(A) < \#(B)$ or $\#(B) > \#(A)$ to indicate that $\#(A) \leq \#(B)$ but $\#(A) \neq \#(B)$. Since any finite set can be represented as a subset of $\mathbb{N}$ but does

[1] This terminology is not universally accepted, so some caution is necessary when comparing discussions of these topics; it is not uncommon to use "at most countable" in place of countable and "countable" in place of countably infinite.

not admit a bijective mapping to $\mathbb{N}$ it follows that $n < \aleph_0$ for any $n \in \mathbb{N}$. For any infinite subset $E \subset \mathbb{N}$ it is the case that $\#(E) = \#(\mathbb{N}) = \aleph_0$, even if E is a proper subset of $\mathbb{N}$; indeed each element of E is in particular a natural number n, so when these natural numbers are arranged in increasing order $n_1 < n_2 < n_3$ then the mapping $f : E \longrightarrow \mathbb{N}$ that associates to $n_i \in E$ the natural number $i \in \mathbb{N}$ is a bijective mapping and consequently $\#(E) = \#(\mathbb{N})$.[2] It is somewhat counterintuitive that any infinite proper subset of $\mathbb{N}$ has the same cardinality, or the same number of elements, as $\mathbb{N}$; if $A \subset B$ is a proper subset and both A and B are finite then $\#(A) < \#(B)$, but that is not necessarily the case for infinite sets. The further examination of inequalities among cardinalities of sets rests upon the following result, which is evident for finite sets but not obvious for infinite sets.

Theorem 1.1 (Cantor-Bernstein Theorem). *If there are injective mappings $f : A \longrightarrow B$ and $g : B \longrightarrow A$ between two sets A and B then there is a bijective mapping $h : A \longrightarrow B$.*

Proof: The mapping $f : A \longrightarrow B$ is injective, but its image is not necessarily all of B; the mapping g is injective so its inverse $g^{-1} : g(B) \longrightarrow B$ is an injective mapping onto B, but it is defined only on the subset $g(B) \subset A$. That suggests considering the mapping $h_0 : A \longrightarrow B$ defined by setting $h_0(a) = f(a)$ for all points $a \in A \sim g(B)$ and $h_0(a) = g^{-1}(a)$ for all points $a \in g(B)$. This is indeed a well-defined surjective mapping $h_0 : A \longrightarrow B$ that restricts to injective mappings on $A \sim g(B)$ and $g(B)$. However there may be, indeed actually are, points $a_1 \in A \sim g(B)$ and $a_2 \in g(B)$ for which $(g \circ f)(a_1) = a_2$ so $f(a_1) = g^{-1}(a_2)$ and the mapping h_0 fails to be injective; the problem thus lies in points

$$a_1 \in (g \circ f)\big(A \sim g(B)\big) \subset g(B).$$

If all such points are moved into the domain of the mapping f rather than the domain of the mapping g^{-1} that solves that problem for such points; but the mapping h_0 may still fail to be injective for a similar reason. The solution is to introduce the set

$$C = (A \sim g(B)) \cup \bigcup_{n \in \mathbb{N}} (g \circ f)^n (A \sim g(B)) \subset A;$$

[2] Actually $\mathbb{N}$ is the smallest infinite set, in the sense that if S is any infinite set then $\#(\mathbb{N}) \leq \#(S)$. The demonstration though does require the use of the axiom of choice; see the discussion in the *Princeton Companion to Mathematics*. Indeed by the axiom of choice it is possible to choose one of the points of S and to label it x_1; since S is infinite there are points in S other than x_1, so by the axiom of choice again it is possible to choose a point in S other than x_1 and to label it x_2; and inductively if $x_1, \ldots, x_\nu$ are labeled there remain other points in S, since S is infinite, so by the axiom of choice again choose another point and label it $x_{\nu+1} = x_{\nu'}$. That establishes the existence of a subset of S indexed by a subset $X = \{\nu\} \subset \mathbb{N}$ of the natural numbers, where X includes 1 and includes ν' for any $\nu \in X$ so by the induction axiom $X = \mathbb{N}$, as desired.

since $(A \sim g(B)) \subset C$ then $(A \sim C) \subset g(B)$. In terms of this set introduce the mapping $h : A \longrightarrow B$ defined by

$$h(x) = \begin{cases} f(x) & \text{if } x \in C \subset A, \\ g^{-1}(x) & \text{if } x \in (A \sim C) \subset A, \end{cases}$$

where $g^{-1}(x)$ is well defined on the subset $(A \sim C) \subset g(B)$ since g is injective. The theorem will be demonstrated by showing that the mapping $h : A \longrightarrow B$ just defined is bijective.

To demonstrate first that h is injective consider any two distinct points $x_1, x_2 \in A$ and suppose to the contrary that $h(x_1) = h(x_2)$. If $x_1, x_2 \in C$ then by definition $h(x_1) = f(x_1)$ and $h(x_2) = f(x_2)$ so $f(x_1) = f(x_2)$, a contradiction since f is injective. If $x_1, x_2 \in \big(A \sim C\big) \subset g(B)$ then $x_1 = g(y_1)$ and $x_2 = g(y_2)$ for uniquely determined distinct points $y_1, y_2 \in B$ since g is injective, and by definition $h(x_1) = y_1$ and $h(x_2) = y_2$ so $y_1 = y_2$, a contradiction. If $x_1 \in C$ and $x_2 \in \big(A \sim C\big)$ then by definition $h(x_1) = f(x_1)$ and $h(x_2) = y_2$ where $y_2 \in B$ is the uniquely determined point for which $g(y_2) = x_2$. Since $h(x_1) = h(x_2)$ it follows that $f(x_1) = y_2$ hence $x_2 = g(y_2) = (g \circ f)(x_1)$. Since $x_1 \in C$ then by the definition of the set C either $x_1 \in (A \sim g(B))$ or there is some $n \in \mathbb{N}$ for which $x_1 \in (g \circ f)^n (A \sim g(B))$; and in either case $x_2 = (g \circ f)(x_1) \in (g \circ f)^k (A \sim g(B)) \subset C$ for some k, a contradiction since $x_2 \in (A \sim C)$. That shows that h is injective.

To demonstrate next that h is surjective, consider a point $y \in B$ and let $x = g(y) \in A$. If $x \in (A \sim C)$ then by definition $h(x) = g^{-1}(x) = y$. If $x \in C$ it cannot be the case that $x \in (A \sim g(B))$ since $x = g(y) \in g(B)$, so it must be the case that there is some $n \in \mathbb{N}$ for which $x \in (g \circ f)^n (A \sim g(B))$; consequently $x = (g \circ f)(x_1)$ for some point $x_1 \in C$. Since $g(f(x_1)) = x = g(y)$ and g is injective it must be the case that $y = f(x_1)$; and since $x_1 \in C$ then by definition $h(x_1) = f(x_1) = y$. That shows that h is surjective and thereby concludes the proof.

The Cantor-Bernstein Theorem shows that the inequality (1.15) among cardinalities of sets satisfies some simple natural conditions; these conditions arise in other contexts as well, so they will be described here a bit more generally. A **partial order** on a set S is defined as a relation $x \leq y$ among elements $x, y \in S$ of that set with the following properties:

(i) *reflexivity*, $x \leq x$ for any $x \in S$;
(ii) *antisymmetry*, if $x \leq y$ and $y \leq x$ then $x = y$ for any $x, y \in S$; and
(iii) *transitivity*, if $x \leq y$ and $y \leq z$ then $x \leq z$ for any $x, y, z \in S$.

For example, the set of all subsets of a set S is clearly a partially ordered set if $A \leq B$ means that $A \subset B$ for any subsets $A, B \subset S$; and the set $\mathbb{N}$ of natural numbers clearly also is a partially ordered set if $a \leq b$ is defined as in (1.15). As in

the case of equivalence relations, the notion of a partial ordering can be applied to relations among sets as well as to relation among elements of a particular set, so in particular it can be applied to the relation $\#(A) \leq \#(B)$.

Corollary 1.2. *The relation* (1.15) *is a partial order on the cardinalities of sets.*

Proof: The reflexivity of the relation (1.15) is clear since the identity mapping $\iota : A \longrightarrow A$ that associates to any $a \in A$ the same element $\iota(a) = a \in A$ is clearly injective; and transitivity is also clear since if $f : A \longrightarrow B$ and $g : B \longrightarrow C$ are injective mappings then the composition $g \circ f : A \longrightarrow C$ is also injective. Antisymmetry on the other hand is far from clear, but is an immediate consequence of the Cantor-Bernstein Theorem; and that is sufficient for the proof.

Theorem 1.3 (Cantor's Theorem). *The power set $\mathfrak{P}(A)$ of any set A has strictly greater cardinality than A, that is,*

$$\#(\mathfrak{P}(A)) > \#(A) \quad \textit{for any set } A. \tag{1.16}$$

Proof: The mapping that sends each element $a \in A$ to the subset $\{a\} \subset A$ is an injective mapping $g : A \longrightarrow \mathfrak{P}(A)$, hence $\#(A) \leq \#(\mathfrak{P}(A))$. Suppose though that there is a bijective mapping $f : A \longrightarrow \mathfrak{P}(A)$. The subset

$$E = \Big\{ x \in A \,\Big|\, x \notin f(x) \Big\}$$

is a well-defined set, possibly the empty set; so since the mapping f is assumed to be surjective the subset E must be the image $E = f(a)$ of some $a \in A$. However if $a \in E$ then from the definition of the set E it follows that $a \notin f(a) = E$, while if $a \notin E = f(a)$ then from the definition of the set E it also follows that $a \in E$; this contradictory situation shows that there cannot be a bijective mapping $f : A \longrightarrow \mathfrak{P}(A)$ and thereby concludes the proof.

In particular $\#(\mathbb{N}) < \#(\mathfrak{P}(\mathbb{N}))$, so the set of all subsets of the set $\mathbb{N}$ of natural numbers is not a countable set but is a strictly larger set, a set with a larger cardinality. The set of all subsets of that set is properly larger still, so there is no end to the size of possible sets. Nonetheless there are many sets that appear to be considerably larger than $\mathbb{N}$ but nonetheless are still countable.

Theorem 1.4 (Cantor's Diagonalization Theorem). *If to each natural number $n \in \mathbb{N}$ there is associated a countable set E_n, then the union $E = \bigcup_{n\in\mathbb{N}} E_n$ is a countable set.*

Proof: Since each set E_n is countable the elements of E_n can be labeled $x_{n,1}, x_{n,2}$ and so on; the elements of the union E then can be arranged in an array

$$\begin{array}{ccccc} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & \cdots \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & \cdots \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array},$$

where row n consists of all the elements $x_{n,i}$ ordered by the natural number $i \in \mathbb{N}$. Imagine these elements now rearranged in order by starting with $x_{1,1}$ then proceeding along the increasing diagonal with $x_{2,1},\ x_{1,2}$, then in the next increasing diagonal $x_{3,1},\ x_{2,2},\ x_{1,3}$ and so on; if row n is finite the places that would be filled by some terms $x_{n,i}$ are blank so are just ignored in this ordering. When all of the elements in E are written out in order as

$$x_{1,1},\ x_{2,1},\ x_{1,2},\ x_{3,1},\ x_{2,2},\ x_{1,3}, \cdots$$

and all the duplicated elements are eliminated there is the obvious injection $E \longrightarrow \mathbb{N}$, showing that E is countable.

Corollary 1.5. *The set of all finite subsets of* $\mathbb{N}$ *is countable.*

Proof: First it follows by induction on n that for any natural number $n \in \mathbb{N}$ the set E_n of all subsets $A \subset \mathbb{N}$ for which $\#(A) = n$ is countable; that is clearly the case for $n = 1$, and if E_n is countable then since

$$E_{n+1} \subset \bigcup_{i \in \mathbb{N}} \left\{ \{i\} \cup A \middle| A \in E_n \right\}$$

it follows that E_{n+1} is countable by Cantor's Diagonalization Theorem and the observation that a subset of a countable set is countable. Then since all the sets E_n are countable it follows again from Cantor's Diagonalization Theorem that the set $E = \bigcup_{n \in \mathbb{N}} E_n$ is countable, which suffices for the proof.

Problems, Group I

1. Write out the proofs of equations (1.4) and (1.8).

2. Show that $A \cup B = A \cap B$ if and only if $A = B$. Is it true that $A \cup B = A$ and $A \cap B = B$ if and only if $A = B$? Why?

3. The formula $A \cup B \cap C \cup D$ does not have a well-defined meaning unless parentheses are added. What are all possible sets this formula can describe for possible placements of parentheses? Illustrate your assertions with suitable Venn diagrams.

4. For which pairs of sets A, B is $A \Delta B = A \cup B$? For which pairs of sets is $A \Delta B = A \cap B$? For which pairs of sets is $A \Delta B = A$?

5. If $f \in B^A$, if $X \subset A$ and if $Y \subset B$ show that $f(X \cap f^{-1}(Y)) = f(X) \cap Y$.

6. If $f \in B^A$ is injective and if X_1, X_2 are subsets of A show that $f(X_1 \cap X_2) = f(X_1) \cap f(X_2)$.

7. Show that the set of all polynomials with rational coefficients is countable.

8. Is the set of all monotonically increasing sequences of natural numbers (sequences of natural numbers a_ν such that $a_\nu \leq a_{\nu+1}$) countable or not? Why?

Problems, Group II

9. Show that the set consisting of all countably long sequences $(a_1, a_2, a_3, \ldots)$, where $a_n = 0$ or 1, is not a countable set. Show though that the set consisting of all countably long sequences $(a_1, a_2, a_3, \ldots)$, where $a_n = 0$ or 1 but $a_n = 0$ for all but finitely many values of n, is countable. Is the set of all sequences $(a_1, a_2, a_3, \ldots)$, where $a_n = 0$ or 1 and where the sequences are periodic in the sense that $a_{n+N} = a_n$ for all n and for some finite number N depending on the sequence, countable or not? Why?

10. Is there an infinite set S for which the power set $\mathfrak{P}(S)$ is countable? Why?

11. For any mapping $f \in B^A$ set $a_1 \asymp a_2$ if $a_1, a_2 \in A$ and $f(a_1) = f(a_2)$. Show that this defines an equivalence relation on the set A. Show that the mapping $f : A \longrightarrow B$ induces an injective mapping $F : A/\asymp \longrightarrow B$.

12. For any given countable collection of sets A_m let

$$A' = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m \quad \text{and} \quad A'' = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

Describe the sets A' and A'', and show that $A' \subset A''$.

13. A partition of a set A is a representation of A as a union $A = \bigcup_\alpha A_\alpha$ of pairwise disjoint subsets $A_\alpha \subset A$. It was shown that to any equivalence relation $a_1 \asymp a_2$ on A there can be associated a partition of A into equivalence classes. Show that conversely for any partition $A = \bigcup_\alpha A_\alpha$ of a set A there is an equivalence relation on A so that the sets A_α are the equivalence classes. (Your proof shows that an equivalence relation on a set is equivalent to a partiton of the set, a useful observation.)

14. To each mapping $f \in B^A$ associate the mapping $f^* \in \mathfrak{P}(B)^{\mathfrak{P}(A)}$ that associates to any subset $X \subset A$ the image $f(X) \subset B$; the mapping $f \longrightarrow f^*$ thus is a mapping $\phi : B^A \longrightarrow \mathfrak{P}(B)^{\mathfrak{P}(A)}$. Is the mapping ϕ injective? Surjective? Why?

1.2 Groups, Rings, and Fields

It is possible to introduce algebraic operations on the cardinalities of sets, and for this purpose two further constructions on sets are relevant. The union $A \cup B$ of two sets was considered in detail in the preceding section; the **disjoint union** $A \bigsqcup B$ is the union of these two sets when they are viewed as being formally disjoint, so that if $a \in A \cap B$ then in the disjoint union the element a is viewed as an element $a_A \in A$ and a distinct element $a_B \in B$. Of course if the two sets A and B are actually disjoint then $A \bigsqcup B = A \cup B$. The **Cartesian product** of the two sets A and B is the set defined by

$$A \times B = \Big\{ (x,y) \big| x \in A,\ y \in B \Big\}. \tag{1.17}$$

In these terms the **sum** and **product** of the cardinalities of sets A and B are defined by

$$\#(A) + \#(B) = \#(A \bigsqcup B) \quad \text{and} \quad \#(A) \cdot \#(B) = \#(A \times B). \tag{1.18}$$

It is fairly clear that these operations are well defined, in the sense that if $A' \leftrightarrow A$ and $B' \leftrightarrow B$ then $(A' \bigsqcup B') \leftrightarrow (A \bigsqcup B)$ and $(A' \times B') \leftrightarrow (A \times B)$. In particular these operations are defined on the natural numbers if the sets are finite and on the number $\aleph_0$ if the sets are countable. They satisfy the following laws:

(i) the **associative law** for addition: $a + (b + c) = (a + b) + c$;
(ii) the **commutative law** for addition: $a + b = b + a$;
(iii) the **associative law** for multiplication: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$;
(iv) the **commutative law** for multiplication: $a \cdot b = b \cdot a$; and
(v) the **distributive law**: $(a + b) \cdot c = a \cdot c + b \cdot c$.

To verify that these laws are satisfied, if $a = \#(A)$, $b = \#(B)$, and $c = \#(C)$ then $b + c = \#(B \bigsqcup C)$ so $a + (b + c) = \#(A \bigsqcup (B \bigsqcup C)) = \#(A \bigsqcup B \bigsqcup C)$ while $a + b = \#(A \bigsqcup B)$ so $(a + b) + c = \#\big((A \bigsqcup B) \bigsqcup C\big) = \#(A \bigsqcup B \bigsqcup C)$, showing that $a + (b + c) = (a + b) + c$. Moreover $a + b = \#(A \bigsqcup B) = \#(B \bigsqcup A) = b + a$. Then for multiplication $b \cdot c = \#(B \times C)$ so $a \cdot (b \cdot c) = \#(A \times (B \times C))$ where $B \times C = \{(b,c)\}$ so $A \times (B \times C) = \{(a,(b,c))\} = \{(a,b,c)\}$; correspondingly $(a \cdot b) \cdot c = \#((A \times B) \times C)$ where $(A \times B) \times C = \{((a,b),c)\} = \{(a,b,c)\}$, and consequently $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. Moreover $a \cdot b = \#(A \times B)$ and $b \cdot a = \#(B \times A)$; but the mapping that sends $(x,y) \in A \times B$ to $(y,x) \in B \times A$ is a bijective mapping so $\#(A \times B) = \#(B \times A)$ and consequently $a \cdot b = b \cdot a$. Finally $(a + b) \cdot c = \#((A \bigsqcup B) \times C)$ where $(A \bigsqcup B) \times C = (A \times C) \bigsqcup (B \times C)$ so $\#((A \bigsqcup B) \times C) = \#((A \times C) \bigsqcup (B \times C)) = a \cdot c + b \cdot c$ and consequently $(a + b) \cdot c = a \cdot c + b \cdot c$.

The algebraic notation is usually simplified by writing ab in place of $a \cdot b$ and by dropping parentheses when the associative laws indicate that they are not needed to specify the result of the operation uniquely, so by writing $a + b + c$

in place of $(a+b)+c$ and abc in place of $(a\cdot b)\cdot c$. Some care must be taken, though, since there are expressions in which parentheses are necessary. The expression $a\cdot b+c$ could stand for either $a\cdot(b+c)$ or $(a\cdot b)+c$ and these can be quite different. The informal convention is to give the product priority, in the sense that products are grouped together first; so normally $a\cdot b+c$ is interpreted as $(ab)+c$. If there are any doubts, though, it is safer to insert parentheses.

The addition of natural numbers is closely related to other operations on natural numbers. First, the successor to a natural number $a=\#(A)$, where A is a finite collection of strokes, is identified with the natural number $a'=\#(A')$, where A' is derived from A by adding another stroke; thus $A'=A\bigsqcup\{/\}$ hence $\#(A')=\#(A)+\#(\{/\})$ so

$$a'=a+1 \quad \text{for any } a\in\mathbb{N}. \tag{1.19}$$

This provides another interpretation of the successor operation on the natural numbers and thereby fits that operation into the standard algebraic machinery. Second, there is the **cancellation law** of the natural numbers:

$$a+n=b+n \quad \text{for } a,b,n\in\mathbb{N} \text{ if and only if } a=b. \tag{1.20}$$

Indeed it is clear that $a+n=b+n$ if $a=b$, and the converse can be established by induction on n. For that purpose note that $a+1=b+1$ is equivalent to $a'=b'$, which by succession implies that $a=b$; thus if $a+n+1=b+n+1$ for some $n\in\mathbb{N}$ then $(a+n)'=(b+n)'$ so by succession $a+n=b+n$ and then by induction $a=b$. The cancellation law plays a critical role in extending the algebraic operations on the natural numbers. There is no analogue of the cancellation law for infinite cardinals; indeed it is easy to see that $\aleph_0=\aleph_0+1=\aleph_0+2$ and so on, so the cancellation law does not hold in this case. Third, the order relation $a<b$ for natural numbers, where $a=\#(A)$ and $b=\#(B)$ for sets A and B consisting of collections of strokes, indicates that there is an injective mapping $f:A\longrightarrow B$ but that there is not a bijective mapping $g:A\longrightarrow B$; the collection of strokes B can be viewed as a collection of strokes bijective to A together with some additional strokes C, hence $B=A\bigsqcup C$ and therefore $\#(B)=\#(A)+\#(C)$ so

$$\text{if} \quad a,b\in\mathbb{N} \ \text{ then } \ a<b \ \text{ if and only if } \ b=a+c \ \text{ for some } \ c\in\mathbb{N}. \tag{1.21}$$

That expresses the order relation for the natural numbers in terms of the group operations on the natural numbers. There is not a similar characterization of the order relation $a\leq b$ since there is no natural number that expresses the equivalence class of the empty set; that is one reason for seeking an extension of the natural numbers.

The basic reason for extending the natural numbers though is to introduce inverses of the algebraic operations of addition and multiplication as far as possible. To describe the algebraic properties of the extended sets it may be clearest first to describe these algebraic structures more abstractly. A **group** is defined to be a set G with a specified element $1 \in G$ and with an operation that associates to any elements $a, b \in G$ another element $a \cdot b \in G$ satisfying

(i) the **associative law**: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for any $a, b, c \in G$;
(ii) the **identity law**: $1 \cdot a = a \cdot 1 = a$ for any $a \in G$;
(iii) the **inverse law**: to each $a \in G$ there is associated a unique $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = 1$.

The element $1 \in G$ is called the **identity** in the group, and the element $a^{-1} \in G$ is called the **inverse** of the element $a \in G$. A group G is said to be an **abelian group**, or equivalently a **commutative group**, if it also satisfies

(iv) the **commutative law**: $a \cdot b = b \cdot a$ for all $a, b \in G$.

A group may consist of either finitely many or infinitely many elements; for a finite group the cardinality of the group is customarily called the **order** of the group. An element a of a group is said to have **order** n if $a^n = 1$, the identity element of the group. Just as in the case of the natural numbers, it is customary to simplify the notation by writing ab in place of $a \cdot b$ and dropping parentheses when the meaning is clear, so by writing abc in place of $(ab)c$. The **cancellation law** holds in any group: if $a \cdot b = a \cdot c$ for some $a, b, c \in G$ then $b = (a^{-1} \cdot a) \cdot b = a^{-1} \cdot (a \cdot b) = a^{-1} \cdot (a \cdot c) = (a^{-1} \cdot a) \cdot c = c$. Although the multiplicative notation for the group operation is most common in general, there are cases in which the additive notation is used. In the additive notation the group operation associates to any elements $a, b \in G$ another element $a + b \in G$; the associative law takes the form $(a + b) + c = a + (b + c)$; the identity element is denoted by 0 and the identity law takes the form $a + 0 = 0 + a = a$; and the inverse law associates to each $a \in G$ a unique element $-a \in G$ such that $a + (-a) = (-a) + a = 0$.

The simplest group consists of just a single element 1 with the group operation $1 \cdot 1 = 1$. A more interesting example of a group is the **symmetric group** $S(A)$ on a set A, the set of all bijective mappings $f : A \longrightarrow A$, where the product $f \cdot g$ of two bijections f and g is the composition $f \circ g$, the identity element of $S(A)$ is the identity mapping $\iota : A \longrightarrow A$ for which $\iota(a) = a$ for all $a \in A$, and the inverse of a mapping f is the usual inverse mapping f^{-1}. It is a straightforward matter to verify that $S(A)$ does satisfy the group laws. Even more interesting though are the groups of bijective mappings of a set A that preserve some additional structures on the set A, often called the groups of **symmetries** of these sets. For instance, assuming some familiarity with elementary plane geometry, suppose that A is a plane rectangle as in the accompanying Figure 1.2

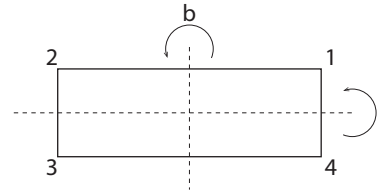

Figure 1.2. Symmetry group of a rectangle

and consider the Euclidean motions of three-space that are bijective mappings of the rectangle to itself. One motion is the flip a of the plane around the horizontal axis of the rectangle, a mapping that interchanges the vertices 1 and 4 and interchanges the vertices 2 and 3. Another motion is the flip b around the vertical axis of the rectangle, a mapping that interchanges the vertices 1 and 2 and interchanges the vertices 3 and 4. Repeating each of these mappings leaves the rectangle unchanged, so $a^2 = b^2 = 1$, the identity mapping. Performing first the flip a and then the flip b interchanges the vertices 1 and 3 and interchanges the vertices 2 and 4, so it amounts to a rotation $c = ba$ of the plane around the center of the rectangle through an angle of π radians or 180 degrees; and repeating this rotation leaves the rectangle unchanged, so $c^2 = 1$ as well. On the other hand performing the two flips in the other order really amounts to the same rotation, so $c = ab$ as well. These symmetries thus form a group of order 4, consisting of the identity mapping 1, the flips a and b and the rotation c. As for other finite groups, the algebraic operations in this group can be written as a multiplication table for the group, often called the **Cayley table** for the group, as in Table 1.1. The elements of the group are listed in the top row and the left-hand column; the entry in the table in the row associated to an element x of the group and the column associated to an element y of the group is the product $x \cdot y$ in the group. Evidently then the group is abelian if and only if the Cayley table is symmetric about the main diagonal, as is the case in Table 1.1. Each row and each column of the Cayley table for any group must be a list of all the elements of the group with no repetitions; for the multiplication of the elements of a group G by a fixed element $a \in G$ is a bijective mapping from the group to itself. In Table 1.1 the entries along the principal diagonal are all 1 reflecting that the elements of the group have order 2, so that $x^2 = 1$ for all elements x of the group. This group $\mathfrak{V}$ of symmetries of a rectangle thus is a group of order 4 in which each element has order 2; it is called **Klein's Vierergruppe**.[3]

Another group of order 4 is the group of rotations that preserve a square; that group consists of the identity mapping, a counterclockwise rotation a of

[3] Alternatively it is called Klein's four group; it was an example examined illustratively by Felix Klein in his classic book *Vorlesungen über das Ikosaeder und die Auflösung der Gleichungen vom fünften Grade* (Lectures on the icosahedron and the solution of equations of the fifth degree) in 1884.

TABLE 1.1.
Cayley table for the group of symmetries of a rectangle

	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

TABLE 1.2.
Cayley table for the group of rotations of a square

	1	a	b	c
1	1	a	b	c
a	a	b	c	1
b	b	c	1	a
c	c	1	a	b

$\pi/2$ radians or 90 degrees, a counterclockwise rotation b of π radians or 180 degrees, and a counterclockwise rotation c of $3\pi/2$ radians or 270 degrees. These operations clearly satisfy $a^2 = b$, $a^3 = c$, $a^4 = 1$, so the group can be viewed as consisting of the products a, a^2, a^3, $a^4 = 1$. This group $\mathbb{Z}_4$ is called the **cyclic** group of order 4, and its Cayley table is Table 1.2. It is clear from comparing Tables 1.1 and 1.2 that the two groups described by these tables are quite distinct; every element of the first group is of order 2 but that is not the case for the second group.

If G is a group and $H \subset G$ is a subset such that $1 \in H$, that $a \cdot b \in H$ whenever $a, b \in H$, and that $a^{-1} \in H$ whenever $a \in H$, then it is clear that the subset H also has the structure of a group with the group operation of G; such a subset is called a **subgroup** of G. A mapping $\phi : G \longrightarrow H$ from a group G to a group H is called a **homomorphism** of groups if $\phi(a \cdot b) = \phi(a) \cdot \phi(b)$ for all elements $a, b \in G$. A group homomorphism preserves other aspects of the group automatically. For instance if 1_G denotes the identity element of G and 1_H denotes the identity element of H then since $\phi(1_G) = \phi(1_G \cdot 1_G) = \phi(1_G) \cdot \phi(1_G)$ it follows from cancellation in H that $\phi(1_G) = 1_H$; moreover since $\phi(a) \cdot \phi(a^{-1}) = \phi(a \cdot a^{-1}) = \phi(1_G) = 1_H$ it follows that $\phi(a^{-1}) = (\phi(a))^{-1}$. A group homomorphism that is a bijective mapping is called a group **isomorphism**; the inverse mapping is clearly also a group isomorphism $\phi^{-1} : H \longrightarrow G$, and the groups G and H can be identified through this isomorphism. For example, it is easy to see that any group of order 4 must be isomorphic either to Klein's Vierergruppe $\mathfrak{V}$ or to the

cyclic group $\mathbb{Z}_4$, so up to isomorphism there are just two groups of order 4, both of which are abelian. The smallest nonabelian group is one of order 6.

Groups involve a single algebraic operation; but the natural numbers and other sets involve two separate algebraic operations. The basic structure of interest in this context is that of a **ring**,[4] defined as a set R with distinct specified elements $0, 1 \in R$ and with operations that associate to any two elements $a, b \in R$ their **sum** $a + b \in R$ and their **product** $a \cdot b \in R$ such that

(i) R is an abelian group under the sum operation with the identity element 0;
(ii) the product operation is associative and commutative and has the identity element $1 \in R$, so that $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ and $a \cdot b = b \cdot a$ and $1 \cdot a = a \cdot 1 = a$ for any elements $a, b, c \in R$; and
(iii) the operations are distributive, in the sense that $a \cdot (b + c) = a \cdot b + a \cdot c$ for any elements $a, b, c \in R$.

The additive inverse of an element $a \in R$ is denoted by $-a \in R$. The operation $a + b$ alternatively is called **addition** and the operation $a \cdot b$ alternatively is called **multiplication**. The additive identity 0 plays a special role in multiplication; for the distributive law implies that $a \cdot 0 = 0$ for any $a \in R$, since $a = a \cdot 1 = a \cdot (1 + 0) = a \cdot 1 + a \cdot 0 = a + a \cdot 0$ hence by cancellation $a \cdot 0 = 0$. Another consequence of the distributive law is that $(-a) \cdot b = -(a \cdot b)$, since $a \cdot b + (-a) \cdot b = (a + (-a)) \cdot b = 0 \cdot b = 0$; therefore of course $(-a) \cdot (-b) = a \cdot b$ as well, and in particular $(-1)^2 = 1$. A **field** is defined to be a ring F for which the set $F^\times$ of nonzero elements of F is a group under multiplication; thus a field is a ring with the additional property that whenever $a \in F$ and $a \neq 0$ then there is an element $a^{-1} \in F$ for which $a\,a^{-1} = a^{-1}\,a = 1$. For both rings and fields the notation is customarily simplified by writing ab in place of $a \cdot b$, by dropping parentheses if the meaning is clear, so by writing $a + b + c$ in place of $a + (b + c)$ and abc in place of $a(bc)$, and by writing $a - b$ in place of $a + (-b)$. If R is a ring a subset $S \subset R$ is called a **subring** if S itself is a ring under the operations of R, and correspondingly a **subfield** for fields. A subring $S \subset R$ thus must contain the elements $0, 1$, the sum $a + b$ and product ab of any elements $a, b \in S$, and the additive inverse $-a$ of any element $a \in S$; and in the case of fields S must contain the multiplicative inverse a^{-1} of any nonzero element $a \in S$. A mapping $\phi : R \longrightarrow S$ between two rings R, S is called a **homomorphism** of rings if

(i) $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(a \cdot b) = \phi(a) \cdot \phi(b)$ for any $a, b \in R$;
(ii) $\phi(1_R) = 1_S$ for the multiplicative identities $1_R, 1_S$ of these two rings.

[4] What is called a ring here sometimes is called a commutative ring with an identity, since there are more general algebraic structures similar to rings but without the assumptions that multiplication is commutative and that there is a multiplicative identity element.

Since a ring is a group under addition it follows as for general groups that $\phi(0_R) = 0_S$ for the additive identities and $\phi(-a) = -\phi(a)$ for any $a \in R$; but since R is not a group under multiplication it is necessary to assume that a homomorphism preserves the multiplicative identity elements. A **homomorphism** of fields has the same definition. A bijective ring or field homomorphism is called a ring or field **isomorphism**.

The natural numbers can be extended to a ring by adding enough elements to provide inverses under addition; however the additional elements must be chosen so that the extension is actually a ring, that is, so that the associative, commutative, and distributive laws continue to hold for the extended elements. A convenient way of ensuring this is to construct the extension directly in terms of the natural numbers and the operations of addition and multiplication of the natural numbers. Thus consider the Cartesian product $\mathbb{N} \times \mathbb{N}$ consisting of all pairs $\{(a, b)\}$ of natural numbers $a, b \in \mathbb{N}$ and define operations[5] on such pairs by

$$(a, b) + (c, d) = (a + c, b + d) \quad \text{and} \quad (a, b) \cdot (c, d) = (ac + bd, ad + bc). \tag{1.22}$$

It is a straightforward matter to verify that these two operations satisfy the same associative, commutative, and distributive laws as the natural numbers, as a simple consequence of those laws for the natural numbers. The only slightly complicated cases are those of the associative law for multiplication and the distributive law, where a calculation leads to

$$\begin{aligned}((a, b)(c, d))(e, f) &= (ace + adf + bcf + bed,\ ade + acf + bce + bdf) \\ &= (a, b)((c, d)(e, f))\end{aligned}$$

and

$$\begin{aligned}((a, b) + (c, d))(e, f) &= (ae + bf + ce + df,\ af + be + cf + de) \\ &= (a, b)(e, f) + (c, d)(e, f).\end{aligned}$$

Next introduce the equivalence relation on the set $\mathbb{N} \times \mathbb{N}$ defined by

$$(a_1, b_1) \asymp (a_2, b_2) \quad \text{if and only if} \quad a_1 + b_2 = a_2 + b_1. \tag{1.23}$$

It is easy to see that this actually is an equivalence relation; indeed reflexivity and symmetry are obvious, and if $(a, b) \asymp (c, d)$ and $(c, d) \asymp (e, f)$ then $a + d = b + c$ and $c + f = d + e$ so $(a + f) + c = a + (f + c) = a + (d + e) = (a + d) + e = (b + c) + e = (b + e) + c$ from which it follows from the cancellation law of the natural numbers, which must be used in the construction of the extension,

[5] The underlying idea is to view pairs (a, b) as differences $a - b$ of natural numbers, to define the ring operations as on these differences, and to introduce an equivalence relation among pairs that describe the same difference $a - b$.

that $a+f=b+e$ hence $(a,b)\asymp(e,f)$. It is an immediate consequence of the definition of equivalence that

$$(a,b)\asymp(a+c,b+c)\quad\text{for any } a,b,c\in\mathbb{N}, \tag{1.24}$$

a very useful observation in dealing with this equivalence relation. It is another straightforward matter to verify that the group operations on $\mathbb{N}\times\mathbb{N}$ preserve equivalence classes, in the sense that if $(a_1,b_1)\asymp(a_2,b_2)$ then

$$\begin{aligned}((a_1,b_1)+(c,d))\asymp((a_2,b_2)+(c,d))\quad\text{and}\\ ((a_1,b_1)\cdot(c,d))\asymp((a_2,b_2)\cdot(c,d)).\end{aligned} \tag{1.25}$$

The quotient $(\mathbb{N}\times\mathbb{N})/\asymp$ of the set $\mathbb{N}\times\mathbb{N}$ by this equivalence relation is then a well-defined set, denoted by $\mathbb{Z}$ and called the set of **integers**; the operations of addition and multiplication are well defined among equivalence classes of elements of $\mathbb{N}\times\mathbb{N}$ so they are well defined on the set $\mathbb{Z}$, and these algebraic operations satisfy the same associative, commutative, and distributive laws as the natural numbers. In view of (1.24) for any $(a,b)\in\mathbb{N}\times\mathbb{N}$

$$(a,b)+(1,1)=(a+1,b+1)\asymp(a,b) \tag{1.26}$$

so the equivalence class of $(1,1)$ in the quotient $\mathbb{Z}$ satisfies the group identity law; and in addition for any $(a,b)\in\mathbb{N}\times\mathbb{N}$

$$(a,b)+(b,a)=(a+b,a+b)\asymp(1,1) \tag{1.27}$$

so the equivalence class of (b,a) in the quotient $\mathbb{Z}$ acts as the additive inverse of the equivalence class of (a,b) and therefore $\mathbb{Z}$ is a well-defined ring.

The mapping $\phi:\mathbb{N}\longrightarrow\mathbb{Z}$ that associates to any $n\in\mathbb{N}$ the equivalence class in $\mathbb{Z}$ of the pair $(n+1,1)\in\mathbb{N}\times\mathbb{N}$ is an injective mapping, since if $(m+1,1)\asymp(n+1,1)$ then $m+2=n+2$ so $m=n$ by the cancellation law of the natural numbers, which is used again here; therefore the mapping ϕ identifies the natural numbers with a subset $\phi(\mathbb{N})\subset\mathbb{Z}$. Furthermore $\phi(m+n)\in\mathbb{Z}$ is the equivalence class of the pair $(m+n+1,1)\asymp(m+n+1+1,1+1)=((m+1)+(n+1),1+1)=(m+1,1)+(n+1,1)$ while $\phi(m)+\phi(n)\in\mathbb{Z}$ is the equivalence class of the pair $(m+1,1)+(n+1,1)$, so $\phi(m+n)=\phi(m)+\phi(n)$; on the other hand since $\phi(m\cdot n)$ is the equivalence class of the pair $(m\cdot n+1,1)$ while $\phi(m)\cdot\phi(n)$ is the equivalence class of the pair $(m+1,1)\cdot(n+1,1)=((m+1)\cdot(n+1)+1\cdot1,(m+1)\cdot1+(n+1)\cdot1)=(mn+1+m+n+1,1+m+n+1)\asymp(mn+1,1)$ in view of (1.24) so $\phi(m\cdot n)=\phi(m)\cdot\phi(n)$. Thus under the imbedding $\phi:\mathbb{N}\longrightarrow\mathbb{Z}$ the algebraic operations on $\mathbb{N}$ are just the restriction of the algebraic operations on $\mathbb{Z}$, so $\mathbb{Z}$ is an extension of the set $\mathbb{N}$ with its algebraic operations to a ring. To see what is contained in the larger set $\mathbb{Z}$ that is not contained in the subset $\mathbb{N}$, it is evident

from (1.24) that any pair $(m,n) \in \mathbb{N} \times \mathbb{N}$ is equivalent to either $(1,1)$ or $(k+1,1)$ or $(1,k+1)$ where $k \in \mathbb{N}$; the equivalence class of $(k+1,1)$ is in the image $\phi(\mathbb{N})$ of the natural number k, the equivalence class of $(1,1)$ is the additive identity in $\mathbb{Z}$ but is not contained in $\phi(\mathbb{N})$, and the equivalence class of $(1,k+1)$ is the inverse $-\phi(k)$ but is not contained in $\phi(\mathbb{N})$; so the only elements in $\mathbb{Z}$ not contained in the image $\phi(\mathbb{N})$ are precisely the elements needed to extend the natural numbers to a ring, in the sense that $\mathbb{Z} = \mathbb{N} \sqcup \{0\} \sqcup -\mathbb{N}$. Whenever $(m,n) \in \mathbb{N} \times \mathbb{N}$ then $(m,n) \asymp (m+1+1, n+1+1) = (m+1,1) + (1,n+1) = \phi(m) - \phi(n)$, so the elements of the ring $\mathbb{Z}$ can be identified with differences of elements in the image $\phi(\mathbb{N})$ of the natural numbers in the ring $\mathbb{Z}$.

A particularly interesting class of rings are the **ordered rings**, defined as rings R with a distinguished subset $P \subset R$ such that

(i) for any element $a \in R$ either $a = 0$ or $a \in P$ or $-a \in P$, and only one of these possibilities can occur, and
(ii) if $a, b \in P$ then $a + b \in P$ and $a \cdot b \in P$.

The set P is called the set of **positive elements** of the ring R; and associated to this set P is the partial order defined by setting $a \leq b$ if either $b = a$ or $b - a \in P$. That this is indeed a partial order is clear: it is reflexive since $a \leq a$; it is transitive since if $a \leq b$ and $b \leq c$ then $b - a \in P$ and $c - b \in P$ so $c - a = (c - b) + (b - a) \in P$; and it is antisymmetric since if $a \leq b$ then either $b = a$ or $b - a \in P$ and if $b \leq a$ then either $b = a$ or $a - b \in P$, and since it cannot be the case that both $b - a \in P$ and $a - b \in P$ it must be the case that $a = b$. Actually it is a rather special partial order, a **linear order**, meaning that for any $a, b \in R$ either $b - a = 0$ or $b - a \in P$ or $a - b \in P$ and only one of these possibilities can arise; and these possibilities correspond to the relations $a = b$, $a < b$ and $b < a$ respectively. The relation of set inclusion $A \subset B$ is an example of a partial order that is not a linear order. As usual for an order $a < b$ is taken to mean that $a \leq b$ but $a \neq b$, and $b \geq a$ is equivalent to $a \leq b$ while $b > a$ is equivalent to $a < b$. Since $a = a - 0$ it is clear that $a > 0$ if and only if $a \in P$, so the positive elements of the ordered ring are precisely those elements $a \in R$ for which $a > 0$; note particularly that the additive identity 0 is not considered a positive element. It may be useful to list a few standard properties of the order in an ordered ring.

(i) If $a \in R$ and $a \neq 0$ then $a^2 > 0$; for if $a \in P$ then $a^2 \in P$, while if $a \notin P$ then $-a \in P$ and $a^2 = (-a)^2 \in P$.
(ii) The multiplicative identity is positive, that is $1 > 0$, since $1 = 1^2$.
(iii) If $a \leq b$ and $c > 0$ then $ca \leq cb$, for if $b - a \in P$ and $c \in P$ then $c \cdot (b - a) \in P$.
(iv) If $a \leq b$ and $c < 0$ then $ca \geq cb$, for if $b - a \in P$ and $c < 0$ then $-c \in P$ and $c(a - b) = (-c)(b - a) \in P$.
(v) If $a \leq b$ then $-a \geq -b$, for $(-a) - (-b) = b - a \in P$.

For the special case of the ring $\mathbb{Z}$ it is clear that the subset $\phi(\mathbb{N})$ satisfies the condition to be the set of positive elements defining the order on $\mathbb{Z}$; these elements are called the **positive integers**, noting again that with this convention 0 is not considered a positive integer.

Some interesting additional examples of rings can be derived from the ring $\mathbb{Z}$ by considering for any $n \in \mathbb{N}$ the equivalence relation defined by

$$a \equiv b \pmod{n} \quad \text{if and only if } a - b = n\,x \text{ for some } x \in \mathbb{Z}. \tag{1.28}$$

It is easy to see that if $a_1 \equiv a_2 \pmod{n}$ and $b_1 \equiv b_2 \pmod{n}$ then $(a_1 + b_1) \equiv (a_2 + b_2) \pmod{n}$ and $(a_1 \cdot b_1) \equiv (a_2 \cdot b_2) \pmod{n}$; hence the operations of addition and multiplication can be defined on equivalence classes, providing the natural structure of a ring on the quotient; that quotient ring is often denoted by $\mathbb{Z}/n\mathbb{Z}$. This ring is not an ordered ring; for if it were an ordered ring then $1 + 1 + \cdots + 1 > 0$ for any such sum since $1 > 0$, but the sum of n copies of 1 is 0. It is amusing and instructive to write out the detailed addition and multiplication tables for some small values of n.

To any ring R it is possible to associate another ring $R[x]$ called the **polynomial ring** over the ring R. The elements $p(x) \in R[x]$ are formal polynomials in a variable x, expressions of the form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{1.29}$$

with coefficients $a_j \in R$; they can be written more succinctly and more conveniently in the form $p(x) = \sum_{j=0}^{n} a_j x^j$. If $a_n \neq 0$ the polynomial is said to have **degree** n, the term $a_n x^n$ is the **leading term** of the polynomial and the coefficient a_n is the **leading coefficient**. The ring R is naturally imbedded as a subset of $R[x]$ by viewing any $a \in R$ as a polynomial of degree 0. A polynomial really is just a string $(a_0, a_1, \ldots, a_n)$ of finitely many elements $a_i \in R$ in the ring, written out as in (1.29) for convenience in defining the algebraic operations. Two polynomials always can be written in the form $p(x) = \sum_{j=0}^{n} a_j x^j$ and $q(x) = \sum_{k=0}^{n} b_k x^k$ for the same integer n, since some of the coefficients a_j or b_k can be taken to be zero. The sum of these polynomials is defined by

$$p(x) + q(x) = \sum_{j=0}^{n} (a_j + b_j) x^j$$

so is just the sum of the corresponding coefficients; and their product is defined by

$$p(x) \cdot q(x) = \sum_{j,k=0}^{n} a_j b_k x^{j+k}$$

so is a collection of products of a coefficient of $p(x)$ and a coefficient of $q(x)$ grouped by powers of the variable x. It is a straightforward matter to verify that $R[x]$ is a ring with these operations, where the additive identity is the identity 0 of the ring R and the multiplicative identity is the multiplicative identity 1 of the ring R.

In rings such as $\mathbb{Z}/6\mathbb{Z}$ there are nonzero elements such as the equivalence class a of the integer 2 and the equivalence class b of the integer 3 for which $a \cdot b = 0$. Such elements cannot possibly have multiplicative inverses; for that would imply that $b = (a^{-1} \cdot a) \cdot b = a^{-1} \cdot (a \cdot b) = a^{-1} \cdot 0 = 0$, a contradiction. A ring for which $a \cdot b = 0$ implies that either $a = 0$ or $b = 0$ is called an **integral domain**. It is clear that any ordered ring R is an integral domain; for if R is an ordered ring, if $a, b \in R$ and neither of these two elements is zero, then $\pm a \in P$ and $\pm b \in P$ for some choice of the signs and consequently $\pm(a \cdot b) = (\pm a) \cdot (\pm b) \in P$ for the appropriate sign so $a \cdot b \neq 0$. In an integral domain the cancellation law holds: if $a \cdot c = b \cdot c$ where $c \neq 0$ then $0 = bc - ac = (b - a) \cdot c$ and since $c \neq 0$ it follows that $b - a = 0$ so $b = a$. Any field obviously is an integral domain; but there are integral domains, such as the ring of integers $\mathbb{Z}$, that are not fields. However any integral domain R actually can be extended to a field. To see this, consider the Cartesian product $R \times R^{\times}$ consisting of all pairs (a, b) where $a, b \in R$ and $b \neq 0$; and define[6] operations of addition and multiplication of elements of $R \times R^{\times}$ by

$$(a, b) + (c, d) = (ad + bc, bd) \quad \text{and} \quad (a, b) \cdot (c, d) = (ac, bd). \tag{1.30}$$

Introduce the equivalence relation on the set $R \times R^{\times}$ defined by

$$(a_1, b_1) \asymp (a_2, b_2) \quad \text{if and only if } a_1 \cdot b_2 = a_2 \cdot b_1. \tag{1.31}$$

It is easy to see that this actually is an equivalence relation. Indeed reflexivity and symmetry are trivial; and for transitivity if $(a_1, b_1) \asymp (a_2, b_2)$ and $(a_2, b_2) \asymp (a_3, b_3)$ then $a_1 b_2 = a_2 b_1$ and $a_2 b_3 = a_3 b_2$ so $a_1 b_2 b_3 = b_1 a_2 b_3 = b_1 a_3 b_2$ and consequently $(a_1 b_3 - a_3 b_1) b_2 = 0$ so $a_1 b_3 - a_3 b_1 = 0$ by the cancellation law. It is an immediate consequence of the definition of equivalence that

$$(a, b) \asymp (a \cdot c, \ b \cdot c) \quad \text{if} \quad c \neq 0, \tag{1.32}$$

a very useful observation in dealing with this equivalence relation. It is a straightforward matter to verify that addition and multiplication on $R \times R^{\times}$

[6]The idea of this construction is to introduce formally the quotients a/b of pairs of elements of the ring integers for which $b \neq 0$; and the definitions of the sum and product of two pairs are defined with this in mind.

preserve equivalence classes, in the sense that if $(a_1, b_1) \asymp (a_2, b_2)$ so that $a_1 b_2 = a_2 b_1$ then

$$\begin{aligned} ((a_1, b_1) + (c, d)) &\asymp ((a_2, b_2) + (c, d)) \quad \text{and} \\ ((a_1, b_1) \cdot (c, d)) &\asymp ((a_2, b_2) \cdot (c, d)). \end{aligned} \tag{1.33}$$

Therefore these operations can be defined on the set of equivalence classes and it is readily verified that this set is a ring under these operations, with the additive identity element the equivalence class of $(0, 1)$, the multiplicative identity element the equivalence class of $(1, 1)$, and the additive inverse $-(a, b)$ the equivalence class of $(-a, b)$. The only slightly complicated case is that of the associative law of addition, where a calculation leads to

$$(a, b) + ((c, d) + (e, f)) = (adf + bcf + bde, bdf) = ((a, b) + (c, d)) + (e, f).$$

The quotient $F = (R \times R^\times)/\asymp$ of the set $R \times R^\times$ by this equivalence relation is then a well-defined ring. If $a \neq 0$ and $b \neq 0$ then $(b, a) \cdot (a, b) = (a \cdot b, a \cdot b) \asymp (1, 1)$, since the product $a \cdot b$ is nonzero in an integral domain; thus (b, a) is the multiplicative inverse of (a, b), and consequently F is a field. This field is called the **field of quotients** of the ring R. The mapping $\phi : R \longrightarrow F$ that sends an element $a \in R$ to the equivalence class $\phi(a) \in F$ of the element $(a, 1) \in R \times R^\times$ is clearly an injective mapping for which $\phi(a_1 + a_2) = \phi(a_1) + \phi(a_2)$ and $\phi(a_1 \cdot a_2) = \phi(a_1) \cdot \phi(a_2)$; in this way R can be realized as a subset of F such that the algebraic operations of R are compatible with those of F.

In particular since the ring $\mathbb{Z}$ is an ordered ring, hence an integral domain, it has a well-defined field of quotients, denoted by $\mathbb{Q}$ and called the field of **rational numbers**. Any rational number in $\mathbb{Q}$ can be represented by the equivalence class of a pair $(a, b) \in \mathbb{Z} \times \mathbb{Z}^\times$; and since $(a, b) = (a, 1) \cdot (1, b)$ then if $\phi(a)$ is the equivalence class of $(a, 1)$ and $\phi(b)^{-1}$ is the equivalence class of $(1, b)$ it follows that $\phi(a) \cdot \phi(b)^{-1}$ is the equivalence class of (a,b), so any rational can be represented by the product $a \cdot b^{-1}$ for some integers a, b, and as in (1.32) that rational also can be represented by the product $(a \cdot c)(b \cdot c)^{-1}$ for any nonzero integer c. It is customary to simplify the notation by setting $a \cdot b^{-1} = a/b$, so that rationals can be represented by quotients a/b for integers a, b. The field $\mathbb{Q}$ is an ordered field, where the set P of positive elements is defined as the set of rationals a/b where $a > 0$ and $b > 0$. To verify that, any rational other than 0 can be represented by a quotient a/b of integers; if the integers are of the same sign then after multiplying both by -1 if necessary both will be positive so $a/b \in P$; but if they are of opposite signs the quotient a/b can never be represented by a quotient of two positive integers so $a/b \notin P$. It is clear that if a/b and c/d are both in P then so are $(a/b) + (c/d)$ and $(a/b) \cdot (c/d)$, so P can serve as the positive elements in the field $\mathbb{Q}$. The natural numbers are a countably infinite

set, as already noted. Since the integers can be decomposed as the union $\mathbb{Z} = \{0\} \sqcup \mathbb{N} \sqcup -\mathbb{N}$ of countable sets, where $\mathbb{N}$ is identified with the positive integers, then the ring $\mathbb{Z}$ is also a countably infinite set. The set of all pairs of integers is also countably infinite, by Cantor's Diagonalization Theorem; and since the rationals can be imbedded in the set of pairs $\mathbb{Z} \times \mathbb{Z}$ by selecting a representative pair for each equivalence class in $\mathbb{Q}$ it follows that the field $\mathbb{Q}$ is also a countably infinite set. This is yet another example of sets $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$ where the inclusions are strict inclusions but nonetheless $\#(\mathbb{N}) = \#(\mathbb{Z}) = \#(\mathbb{Q}) = \aleph_0$.

The rationals $\mathbb{Q}$ form an ordered field, but one that is still incomplete for many purposes. For instance although there are rational numbers x such as $x = 1$ for which $x^2 < 2$ and rational numbers such as $x = 2$ for which $x^2 > 2$, there is no rational number x for which $x^2 = 2$, a very old observation.[7] However the rationals can be extended to a larger field $\mathbb{R}$, the field of real numbers, which does include a number x for which $x^2 = 2$, among many other delightful properties. This is a rather more difficult extension to describe than those from the natural numbers to the integers or from the integers to the rationals; indeed in a sense it is not a purely algebraic extension. Hence the discussion here will begin with an axiomatic definition of the field $\mathbb{R}$ of real numbers; the actual construction of this field as an extension of the rationals, which amounts to a verification that the field $\mathbb{R}$ described axiomatically actually exists, and the demonstration of its uniqueness, will be deferred to the Appendix to Section 2.2. To begin more generally, suppose that F is an ordered field, where the order is described by a subset $P \subset F$ of positive elements. A nonempty subset $E \subset F$ is **bounded above** by $a \in F$ if $x \leq a$ for all $x \in E$, and correspondingly it is **bounded below** by $b \in F$ if $x \geq b$ for all $x \in E$. That E is bounded above by a is indicated by writing $E \leq a$, and that E is bounded below by b is indicated by writing $E \geq b$; more generally $E_1 \leq E_2$ indicates that $x_1 \leq x_2$ for any $x_1 \in E_1$ and $x_2 \in E_2$. A subset is just said to be bounded above if it is bounded above by some element of the field F, and correspondingly for bounded below. If $E \subset F$ is bounded above, an element $a \in F$ is said to be the **least upper bound** or **supremum** of the set E if

(i) $E \leq a$, and
(ii) if $E \leq x$ then $x \geq a$;

[7] This is often attributed to the school of Pythagoras, an almost mythical mathematician who lived in the sixth century BCE in Greece; see for instance the *Princeton Companion to Mathematics*. Suppose that $a, b \in \mathbb{N}$ are any natural numbers for which $(a/b)^2 = 2$, or equivalently $a^2 = 2b^2$, where it can be assumed that not both a and b are multiples of 2. Since a^2 is a multiple of 2 then a itself must be a multiple of 2, so $a = 2c$; but then $4c^2 = (2c)^2 = a^2 = 2b^2$ so b must also be a multiple of 2, a contradiction.

the least upper bound of a set E is denoted by $\sup(E)$. Correspondingly if $E \subset F$ is bounded below, an element $b \in F$ is said to be the **greatest lower bound** or **infimum** of the set E if

(i) $E \geq b$, and
(ii) if $E \geq y$ then $y \leq b$;

the greatest lower upper bound of a set E is denoted by $\inf(E)$. It is evident that if $-E = \{ -x \in F \mid x \in E \}$ then $\sup(-E) = -\inf(E)$.

For a general ordered field it is not necessarily the case that a set that is bounded above has a least upper bound, or that a set that is bounded below has a greatest lower bound; as noted earlier, the set of rational numbers $x \in \mathbb{Q}$ satisfying $x^2 < 2$ is bounded above but has no least upper bound. An ordered field with the property that any nonempty set that is bounded above has a least upper bound is called a **complete ordered field**. For any such field it is automatically the case that any nonempty set that is bounded below has a greatest lower bound, since $\sup(-E) = -\inf(E)$ hence $\inf(E) = -\sup(-E)$.

The set $\mathbb{R}$ of the **real numbers** is defined to be a complete ordered field. That there exists such a field and that it is uniquely defined up to isomorphism remain to be demonstrated; so for the present just assume that $\mathbb{R}$ is a complete ordered field. There is a natural mapping $\phi : \mathbb{N} \longrightarrow \mathbb{R}$ defined inductively by $\phi(1) = 1$ and $\phi(n+1) = \phi(n) + 1$, where 1 denotes either the identity $1 \in \mathbb{N}$ or the multiplicative identity $1 \in \mathbb{R}$ as appropriate and $n + 1 = n'$ is the successor to n in $\mathbb{N}$. Since $\phi(1) = 1 > 0$ in $\mathbb{R}$ then $\phi(2) = \phi(1) + \phi(1) > 0$ in $\mathbb{R}$ as well, and by induction it follows that $\phi(k) > 0$ for every natural number $k \in \mathbb{N}$. Moreover $\phi(n+k) = \phi(n) + \phi(k)$ for all $n, k \in \mathbb{N}$, by induction on k, so $\phi(n+k) > \phi(n)$ for any $n, k \in \mathbb{N}$; consequently the mapping ϕ is injective. By yet another induction on k it follows that $\phi(n \cdot k) = \phi(n) \cdot \phi(k)$, so the mapping ϕ preserves both of the algebraic operations. The ring $\mathbb{Z}$ of integers was constructed in terms of pairs (a, b) of natural numbers. If the mapping ϕ is extended to a mapping $\phi : \mathbb{N}^2 \longrightarrow \mathbb{R}$ by setting $\phi(a, b) = \phi(a) - \phi(b)$ it is a straightforward calculation to verify that $\phi((a, b) + (c, d)) = \phi(a, b) + \phi(c, d)$ and $\phi((a, b) \cdot (c, d)) = \phi(a, b) \cdot \phi(c, d)$ in terms of the algebraic operations on $\mathbb{N}^2$ as defined in (1.22), and that $\phi(a, b) = \phi(c, d)$ whenever $(a, b) \asymp (c, d)$ for the equivalence relation (1.23); this extension thus determines a mapping $\phi : \mathbb{Z} \longrightarrow \mathbb{R}$ that is a homomorphism of rings. If $\phi(m) = \phi(n)$ for some $m, n \in \mathbb{Z}$ then $\phi(m + k) = \phi(n + k)$ for any integer k; there is an integer k so that $m + k$ and $n + k$ are contained in the subset $\mathbb{N} \subset \mathbb{Z}$, and since the mapping ϕ is injective on that subset it must be the case that $m + k = n + k$ hence that $m = n$. That shows that the ring homomorphism $\phi : \mathbb{Z} \longrightarrow \mathbb{R}$ is also injective. The ring $\mathbb{Z}$ is an integral domain, and the field $\mathbb{Q}$ of rational numbers is constructed in terms of pairs (a, b) of integers by the quotient field construction for any integral domain. If the mapping $\phi : \mathbb{Z} \longrightarrow \mathbb{R}$ is extended to a mapping $\phi : \mathbb{Z} \times \mathbb{Z}^\times \longrightarrow \mathbb{R}$ by setting $\phi(a, b) = \phi(a) \cdot \phi(b)^{-1}$ it

is also a straightforward calculation to verify that $\phi((a,b)+(c,d)) = \phi(a,b) + \phi(c,d)$ and $\phi((a,b)\cdot(c,d)) = \phi(a,b)\cdot\phi(c,d)$ in terms of the algebraic operations on $\mathbb{Z}\times\mathbb{Z}^\times$ as defined in (1.30), and that $\phi(a,b) = \phi(c,d)$ whenever $(a,b) \asymp (c,d)$ for the equivalence relation (1.31); this extension thus determines a mapping $\phi : \mathbb{Q} \longrightarrow \mathbb{R}$ which is a homomorphism of fields, and it is readily seen to be an injective homomorphism by the analogue of the argument used to show that the ring homomorphism $\phi : \mathbb{Z} \longrightarrow \mathbb{R}$ is injective. Thus the field $\mathbb{Q}$ of rationals can be viewed as a subfield of the field $\mathbb{R}$ of real numbers, and will be so viewed subsequently; each rational number p/q can be identified with a well-defined real number.

Theorem 1.6. *For any real number $a > 0$ there is an integer n with the property that $1/n < a < n$; and for any real numbers $a < b$ there is a rational number r with the property that $a < r < b$.*

Proof: If it is not true that there is an integer n_1 such that $n_1 > a$ then $n \le a$ for all integers n so the subset $\mathbb{N} \subset \mathbb{R}$ is a nonempty set that is bounded above; and since the real numbers form a complete ordered field the set $\mathbb{N}$ must have a least upper bound b. The real number $b-1$ then is not an upper bound for the set $\mathbb{N}$, so there is some integer $n > b-1$; but in that case $n+1 > b$, so that b cannot be an upper bound of the set $\mathbb{N}$, a contradiction. It follows that the assumption that there is not an integer $n_1 > a$ is false, hence there is an integer $n_1 > a$. In particular there is also an integer $n_2 > \frac{1}{a}$, and then $0 < \frac{1}{n_2} < a$. If $n \ge n_1$ and $n \ge n_2$ then $1/n < a < n$.

If $a < b$ then by what has just been shown there is an integer n_0 such that $\frac{1}{n_0} < (b-a)$. The set $S \subset \mathbb{R}$ of rational numbers $\frac{m}{n_0}$ for all $m \in \mathbb{Z}$ for which $\frac{m}{n_0} \le a$ has the upper bound a, hence it has a least upper bound $r \in \mathbb{R}$. Since $r - \frac{1}{2n_0}$ is not an upper bound of S there must be some rational number $\frac{m_0}{n_0} \in S$ for which $r - \frac{1}{2n_0} < \frac{m_0}{n_0}$, and $\frac{m_0}{n_0} \le a$ since $\frac{m_0}{n_0} \in S$. On the other hand $\frac{m_0+1}{n_0} = \frac{m_0}{n_0} + \frac{1}{n_0} \ge r + \frac{1}{2n_0} > r$ and consequently $\frac{m_0+1}{n_0} \notin S$ so $\frac{m_0+1}{n_0} > a$; and $\frac{m_0+1}{n_0} \le a + \frac{1}{n_0} < b$ so that $a < \frac{m_0+1}{n_0} < b$, which suffices for the proof.

In particular for any real number $r \in \mathbb{R}$ for which $r > 0$ there are integers $n \in \mathbb{Z}$ such that $n > r$ and $\frac{1}{n} < r$; an ordered field with this property is called an **archimedean field**. Furthermore for any two real numbers $r_1 < r_2$ there are rational numbers $\frac{p}{q}$ such that $r_1 < \frac{p}{q} < r_2$; so any real number can be approximated as closely as desired by rational numbers. On the other hand the field $\mathbb{R}$ as defined axiomatically is sufficiently complete that, for example, any positive real number has a unique positive square root. To demonstrate that, for any $r \in \mathbb{R}$ for which $r > 0$ introduce the sets of real numbers

$$X = \left\{ x \in \mathbb{R} \,\middle|\, x > 0, x^2 < r \right\} \quad \text{and} \quad Y = \left\{ y \in \mathbb{R} \,\middle|\, y > 0, y^2 > r \right\}.$$

These sets are clearly nonempty and $x < y$ for any $x \in X$ and $y \in Y$; so by the completeness property of the real numbers there are real numbers x_0, y_0 for which $x_0 = \sup(X)$ and $y_0 = \inf(Y)$. Any $y \in Y$ is an upper bound for X so it must be greater than the least upper bound of X, thus $x_0 \le y$; and x_0 is a lower bound for Y so it must be less than the greatest lower bound of Y, thus $x_0 \le y_0$. If $x_0^2 < r$ then $x_0^2 = r - \epsilon$ for some $\epsilon > 0$, and if $a = 2x_0 + 1$ and h is a real number for which $0 < h < \min(1, \epsilon/a)$ then $2x_0h + h^2 < (2x_0 + 1)h = ah$ so

$$(x_0 + h)^2 = x_0^2 + 2x_0h + h^2 < r - \epsilon + ah < r;$$

thus $(x_0 + h) \in X$, which is a contradiction since $x_0 = \sup(X)$, and consequently $x_0^2 \ge r$. On the other hand if $y_0^2 > r$ then $y_0^2 = r + \epsilon$ for some $\epsilon > 0$, and if $0 < h < \min(y_0, \epsilon/2y_0)$ then

$$(y_0 - h)^2 = y_0^2 - 2y_0h + h^2 > r + \epsilon - 2y_0h > r;$$

thus $(y_0 - h) \in Y$, which is a contradiction since $y_0 = \inf(Y)$, and consequently $y_0^2 \le r$. Altogether then $r \le x_0^2 \le y_0^2 \le r$ hence $x_0^2 = y_0^2 = r$. The uniqueness of the square root of course is clear.

Problems, Group I

1. Verify that the operations on pairs of natural numbers defined in equation (1.22) satisfy the same associative, commutative and distributive laws as do the corresponding operations on the natural numbers, and that these operations are compatible with the equivalence relation on these pairs, in the sense that equation (1.25) holds.

2. i) Show that the power set $\mathfrak{P}(A)$ of a set A is an abelian group, where the group operation is the symmetric difference $A \Delta B$ of sets.
   ii) What is the order of this group?
   iii) Write out the multiplication table for this group in the special case that $\#(A) = 2$.

3. Write out the Cayley table for the group of symmetries of an equilateral triangle (a group of order 6, the smallest nonabelian group).

4. i) Show that if G is a group and if $(ab)^n = 1$ for some elements $a, b \in G$ and some natural number n then $(ba)^n = 1$.
   ii) Show that if G is a group for which $a^2 = 1$ for any elements $a \in G$ then G is an abelian group.

5. Show that in an abelian group the mapping $a \longrightarrow a^n$ for a positive integer n is a group homomorphism. Find an example of a nonabelian group for which this mapping is not a group homomorphism.

6. In an ordered ring R the absolute value $|x|$ of an element $x \in R$ is defined by

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Show that

i) $|x| \geq 0$ for all $x \in R$ and $|x| = 0$ if and only if $x = 0$;
ii) $|xy| = |x||y|$ for any $x, y \in R$;
iii) $|x + y| \leq |x| + |y|$ for any $x, y \in R$.

Problems, Group II

7. The addition and multiplication of cardinal numbers are defined by $\#(A) + \#(B) = \#(A \sqcup B)$ and $\#(A) \cdot \#(B) = \#(A \times B)$ for any sets A, B, finite or not.

i) Show that $n + \aleph_0 = \aleph_0 + \aleph_0 = \aleph_0$ for any $n \in \mathbb{N}$.
ii) Show that $\aleph_0 \cdot \aleph_0 = \aleph_0$.

8. i) Show that the ring $\mathbb{Z}/4\mathbb{Z}$ with 4 elements is not an integral domain.
ii) Construct a field consisting of precisely 4 elements. [Suggestion: Begin with the field $\mathbb{F}_2$ of two elements, and considering the set of pairs (a_1, a_2) of elements $a_1, a_2 \in \mathbb{F}_2$ with the algebraic operations

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2),$$
$$(a_1, a_2) \cdot (b_1, b_2) = (a_1 b_1 + a_2 b_2, a_2 b_1 + a_1 b_2 + a_2 b_2).$$

Remember that $-a = a$ in the field $\mathbb{F}_2$.] (An alternative way of describing the preceding algebraic operations is to write the pairs (a_1, a_2) as $a_1 + \epsilon a_2$ for some entity ϵ and to add and multiply these expressions as polynomials in ϵ where it is assumed that $\epsilon^2 + \epsilon + 1 = 0$; this may remind you of the standard definition of complex numbers.)

9. Show that a finite integral domain is a field.

10. Suppose that $\phi : F \longrightarrow G$ is a mapping from a field F to a field G that preserves the algebraic operations, in the sense that $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in F$.

i) Show that $\phi(a-b) = \phi(a) - \phi(b)$ for all $a, b \in F$.
ii) Show that $\phi(0_F) = 0_G$ where 0_F is the zero element of the field F and 0_G is the zero element in the field G.
iii) Show that either $\phi(a) = 0_G$ for all $a \in F$ or $\phi(a) \neq \phi(b)$ whenever $a, b \in F$ and $a \neq b$.
iv) Show that if $\phi(a) \neq 0$ for some element $a \in F$ then the image of ϕ is a subfield of G.

1.3 Vector Spaces

A **vector space** over a field F is defined to be a set V on which there are two operations, *addition*, which associates to $\mathbf{v}_1, \mathbf{v}_2 \in V$ an element $\mathbf{v}_1 + \mathbf{v}_2 \in V$, and *scalar multiplication*, which associates to $a \in F$ and $\mathbf{v} \in V$ an element $a\mathbf{v} \in V$, such that

(i) V is an abelian group under addition;
(ii) scalar multiplication is associative:

$$(ab)\mathbf{v} = a(b\mathbf{v}) \text{ for any } a, b \in F \text{ and } \mathbf{v} \in V;$$

(iii) the multiplicative identity $1 \in F$ is also the identity for scalar multiplication:

$$1\mathbf{v} = \mathbf{v} \text{ for all } \mathbf{v} \in V;$$

(iv) the distributive law holds for addition and scalar multiplication:

$$\begin{aligned} a(\mathbf{v}_1 + \mathbf{v}_2) &= a\mathbf{v}_1 + a\mathbf{v}_2 \text{ and } (a_1 + a_2)\mathbf{v} \\ &= a_1\mathbf{v} + a_2\mathbf{v} \text{ for all } a, a_1, a_2 \in F \text{ and } \mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in V. \end{aligned}$$

Note that the definition of a vector space involves both the set V and a particular field F. The elements of V are called **vectors**, and usually will be denoted by bold-faced letters; the elements of the field F are called **scalars** in this context. If $0 \in F$ is the additive identity in the field and $\mathbf{v} \in V$ is any vector in the vector space then $0\mathbf{v} = (0+0)\mathbf{v} = 0\mathbf{v} + 0\mathbf{v}$, so since V is a group under addition it follows that $0\mathbf{v}$ is the identity element of that group; that element is denoted by $\mathbf{0} \in V$ and is called the **zero vector**. The simplest vector space over any field F is the vector space consisting of the zero vector $\mathbf{0}$ alone; it is sometimes called the **trivial vector space** or the **zero vector space** and usually it is denoted just by $\mathbf{0}$.

A standard example of a vector space over a field F is the set F^n consisting of n-tuples of elements in the field F. The common practice, which will be followed

systematically here, is to view a vector $\mathbf{v} \in F^n$ as a column vector

$$\mathbf{v} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{1.34}$$

where $x_i \in F$ are the **coordinates** of the vector $\mathbf{v}$. Associated to any vector $\mathbf{v} \in F^n$ though there is also its **transposed vector** ${}^t\mathbf{v}$, the row vector

$${}^t\mathbf{v} = (x_1, \ldots, x_n); \tag{1.35}$$

this alternative version of a vector does play a significant role other than just notational convenience, for instance in matrix multiplication. Sometimes for notational convenience a vector is described merely by listing its coordinates, so as

$$\mathbf{v} = \{x_j\} = \{x_1, \ldots, x_n\}, \tag{1.36}$$

although the vector still will be viewed as a column vector. The sum of two vectors is the vector obtained by adding the coordinates of the two vectors, and the scalar product of a scalar $a \in F$ and a vector $\mathbf{v} \in V$ is the vector obtained by multiplying all the coordinates of the vector by a; thus

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \quad \text{and} \quad a \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} ax_1 \\ \vdots \\ ax_n \end{pmatrix} \quad \text{if } a \in F. \tag{1.37}$$

It is clear that F^n is a vector space over the field F with these operations.

A nonempty subset $V_0 \subset V$ for which $\mathbf{v}_1 + \mathbf{v}_2 \in V_0$ whenever $\mathbf{v}_1, \mathbf{v}_2 \in V_0$ and $a\mathbf{v} \in V_0$ whenever $a \in F$ and $\mathbf{v} \in V_0$ is called a **linear subspace** of V, or sometimes a **vector subspace** of V; clearly a linear subspace is itself a vector space over F. To a linear subspace $V_0 \subset V$ there can be associated an equivalence relation in the vector space V by setting $\mathbf{v}_1 \asymp \mathbf{v}_2 \pmod{V_0}$ whenever $\mathbf{v}_1 - \mathbf{v}_2 \in V_0$. This is clearly an equivalence relation: it is reflexive since $\mathbf{v}_1 - \mathbf{v}_1 = \mathbf{0} \in V_0$; it is symmetric since if $\mathbf{v}_1 - \mathbf{v}_2 \in V_0$ then $(\mathbf{v}_2 - \mathbf{v}_1) = -(\mathbf{v}_1 - \mathbf{v}_2) \in V_0$; and it is transitive since if $\mathbf{v}_1 - \mathbf{v}_2 \in V_0$ and $\mathbf{v}_2 - \mathbf{v}_3 \in V_0$ then $\mathbf{v}_1 - \mathbf{v}_3 = (\mathbf{v}_1 - \mathbf{v}_2) + (\mathbf{v}_2 - \mathbf{v}_3) \in V_0$. The equivalence class of a vector $\mathbf{v} \in V$ is the subset $\mathbf{v} + V_0 = \{\mathbf{v} + \mathbf{w} | \mathbf{w} \in V_0\} \subset V$. If $\mathbf{v}_1 \asymp \mathbf{v}_1'$ and $\mathbf{v}_2 \asymp \mathbf{v}_2'$ then $(\mathbf{v}_1 + \mathbf{v}_2) - (\mathbf{v}_1' + \mathbf{v}_2') = (\mathbf{v}_1 - \mathbf{v}_1') + (\mathbf{v}_2 - \mathbf{v}_2') \in V_0$ since $\mathbf{v}_1 - \mathbf{v}_1' \in V_0$ and $\mathbf{v}_2 - \mathbf{v}_2' \in V_0$, hence $\mathbf{v}_1 + \mathbf{v}_2 \asymp \mathbf{v}_1' + \mathbf{v}_2'$; and similarly $(a\mathbf{v}_1 - a\mathbf{v}_1') = a(\mathbf{v}_1 - \mathbf{v}_1') \in V_0$ so $a\mathbf{v}_1 \asymp a\mathbf{v}_1'$. Thus the algebraic operations on the vector space V extend to the set of equivalence classes, which then also has the structure of a vector space over the field F. This vector space

is denoted by V/V_0 and is called the **quotient space** of V by the subspace V_0, or alternatively the vector space V **modulo** V_0. For example the subset of F^n consisting of vectors for which the first m coordinates are all zero where $0 < m < n$ is a vector subspace $V_0 \subset F^n$. Two vectors in F^n are equivalent modulo V_0 precisely when their first m coordinates are the same, so the equivalence class is described by the first m coordinates; hence the quotient space F^n/V_0 can be identified with the vector space F^m.

If $V_1, V_2 \subset V$ are two linear subspaces of a vector space V over a field F it is clear that their intersection $V_1 \cap V_2$ is a linear subspace of V. If the intersection is the zero subspace, consisting only of the zero vector $\mathbf{0}$, the two linear subspaces are said to be **linearly independent**. The **sum** of any two subspaces is the subset

$$V_1 + V_2 = \left\{ \mathbf{v}_1 + \mathbf{v}_2 \middle| \mathbf{v}_1 \in V_1, \quad \mathbf{v}_2 \in V_2 \right\} \subset V, \tag{1.38}$$

and it is easy to see that it also is a linear subspace of V. The sum of two linear subspaces is said to be a **direct sum** if the two subspaces are linearly independent, and in that case the sum is denoted by $V_1 \oplus V_2$. Any vector $\mathbf{v} \in V_1 + V_2$ can be written as the sum $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ of vectors $\mathbf{v}_1 \in V_1$ and $\mathbf{v}_2 \in V_2$, by the definition of the sum of the subspaces. Any vector $\mathbf{v} \in V_1 \oplus V_2$ also can be written as the sum $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ of vectors $\mathbf{v}_1 \in V_1$ and $\mathbf{v}_2 \in V$, since the direct sum is a special case of the sum of two linear subspaces; but in the case of a direct sum that decomposition is unique. Indeed if $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{w}_1 + \mathbf{w}_2$ for vectors $\mathbf{v}_1, \mathbf{w}_1 \in V_1$ and $\mathbf{v}_2, \mathbf{w}_2 \in V_2$ where $V_1 \cap V_2 = \mathbf{0}$ then $\mathbf{v}_1 - \mathbf{w}_1 = \mathbf{w}_2 - \mathbf{v}_2 \in V_1 \cap V_2 = \mathbf{0}$ hence $\mathbf{v}_1 = \mathbf{w}_1$ and $\mathbf{v}_2 = \mathbf{w}_2$. The uniqueness of this representation of a vector $\mathbf{v} \in V_1 \oplus V_2$ makes the direct sum of two linear subspaces a very useful construction.

On the other hand, to any two vector spaces V_1, V_2 over a field F there can be associated another vector space V over F that is the direct sum of subspaces V_i' isomorphic to V_i; indeed let V be the set of pairs

$$V = \left\{ (\mathbf{v}_1, \mathbf{v}_2) \middle| \mathbf{v}_1 \in V_1, \ \mathbf{v}_2 \in V_2 \right\} \tag{1.39}$$

with the structure of a vector space over F where the sum is defined by

$$(\mathbf{v}_1', \mathbf{v}_2') + (\mathbf{v}_1'', \mathbf{v}_2'') = (\mathbf{v}_1' + \mathbf{v}_1'', \ \mathbf{v}_2' + \mathbf{v}_2'')$$

and the scalar product is defined by

$$a(\mathbf{v}_1, \mathbf{v}_2) = (a\mathbf{v}_1, a\mathbf{v}_2).$$

It is a straightforward matter to verify that $V = V_1' \oplus V_2'$ where

$$V_1' = \left\{ (\mathbf{v}_1', \mathbf{0}) \mid \mathbf{v}_1' \in V_1 \right\} \quad \text{and} \quad V_2' = \left\{ (\mathbf{0}, \mathbf{v}_2') \mid \mathbf{v}_2' \in V_2 \right\},$$

and that V_i' is isomorphic to V_i for $i = 1, 2$. The vector space V so constructed also is called the **direct sum** of the two vector spaces V_1 and V_2, or sometimes their **exterior direct sum**, and also is indicated by writing $V = V_1 \oplus V_2$ as an abbreviation. Sometimes the direct sum of two vector spaces is also called the **product** of these vector spaces and is then denoted by $V_1 \times V_2$. It is of course possible to consider the direct sum, or product, of more than two vector spaces.

The **span** of a finite set $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of vectors in a vector space V over a field F is the set of vectors defined by

$$\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = \left\{ \sum_{j=1}^{n} a_j \mathbf{v}_j \,\middle|\, a_j \in F \right\}; \tag{1.40}$$

clearly the span of any finite set of vectors in V is a linear subspace of V. By convention the span of the empty set of vectors is the trivial vector space $\mathbf{0}$ consisting of the zero vector alone. Vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are said to be **linearly dependent** if $\sum_{j=1}^{n} a_j \mathbf{v}_j = \mathbf{0}$ for some $a_j \in F$ where $a_j \neq 0$ for at least one of the scalars a_j; and these vectors are said to be **linearly independent** if they are not linearly dependent, so if $\sum_{j=1}^{n} a_j \mathbf{v}_j = \mathbf{0}$ implies that $a_j = 0$ for all indices $1 \leq j \leq n$. It is easy to see that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly dependent if and only if $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n)$ for some index $1 \leq j \leq n$. Indeed if $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n)$ then $\mathbf{v}_j \in \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n)$ so $\mathbf{v}_j = \sum_{1 \leq i \leq n, i \neq j} a_i \mathbf{v}_i$, and that shows that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly dependent; conversely if the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly dependent then $\sum_{i=1}^{n} a_i \mathbf{v}_i = \mathbf{0}$ for some scalars a_i, not all of which are zero; so if $a_j \neq 0$ then $\mathbf{v}_j = -\sum_{1 \leq i \leq n, i \neq j} a_i a_j^{-1} \mathbf{v}_i$ so $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n)$.

Theorem 1.7 (Basis Theorem). *If m vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ in a vector space V are contained in the span of n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in V where $m > n$ then the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are linearly dependent.*

Proof: Of course if $\mathbf{w}_i = \mathbf{0}$ for some index i then the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are linearly dependent; so it can be assumed that $\mathbf{w}_i \neq \mathbf{0}$ for all indices i. Since $\mathbf{w}_1 \in \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ it follows that $\mathbf{w}_1 = \sum_{j=1}^{n} a_j \mathbf{v}_j$ for some scalars a_j; not all of the scalars a_j can vanish, so by relabeling the vectors $\mathbf{v}_j$ it can be assumed that $a_1 \neq 0$ hence that $\mathbf{w}_1 = a_1 \mathbf{v}_1 + \sum_{j=2}^{n} a_j \mathbf{v}_j$ or equivalently that $\mathbf{v}_1 = a_1^{-1} \mathbf{w}_1 - \sum_{j=2}^{n} a_1^{-1} a_j \mathbf{v}_j$, and it is clear from this that $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. Then since $\mathbf{w}_2 \in \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \operatorname{span}(\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ it follows that $\mathbf{w}_2 = b_1 \mathbf{w}_1 + \sum_{j=2}^{n} b_j \mathbf{v}_j$ for some scalars b_j, not all of which

vanish. If $b_1 \neq 0$ but $b_j = 0$ for $2 \leq j \leq n$ this equation already shows that the vectors $\mathbf{w}_1, \mathbf{w}_2$ are linearly dependent, and hence of course so are the vectors $\mathbf{w}_1, \cdots, \mathbf{w}_n$. Otherwise $b_j \neq 0$ for some $j > 1$, and by relabeling the vectors $\mathbf{v}_j$ it can be assumed that $b_2 \neq 0$; and by the same argument as before it follows that $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \text{span}(\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n)$. The argument can be continued, so eventually either the set of vectors $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$ is linearly dependent or $\text{span}(\mathbf{w}_1, \ldots, \mathbf{w}_n) = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Since $\mathbf{w}_{n+1} \in \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \text{span}(\mathbf{w}_1, \ldots, \mathbf{w}_n)$ it follows that $\mathbf{w}_{n+1} = \sum_{j=1}^{n} c_j \mathbf{w}_j$ and consequently the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{n+1}$ are linearly dependent and therefore so are the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ and that suffices for the proof.

A vector space V is **finite dimensional** if it is the span of finitely many vectors, that is, if $V = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ for some vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. If these vectors are not linearly independent then as already observed the span is unchanged by deleting at least one of these vectors; and that argument can be repeated until all the remaining vectors are linearly independent. Therefore any finite dimensional vector space can be written as the span of a finite set of linearly independent vectors, called a **basis** for the vector space. It is an immediate consequence of the Basis Theorem that the number of vectors in a basis is the same for all bases. Indeed if $\mathbf{w}_1, \ldots, \mathbf{w}_m$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are two bases for a vector space V then since $\mathbf{w}_1, \ldots, \mathbf{w}_m \in \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ and since the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are linearly independent the Basis Theorem shows that $m \leq n$; reversing the roles of these two bases and applying the same argument shows that $n \leq m$. The number of vectors in a basis for a vector space V is called the **dimension** of that vector space; that V has dimension n is indicated by writing $\dim V = n$.

Theorem 1.8. *If V is a finite dimensional vector space over a field F and $W \subset V$ is a subspace then W and the quotient space V/W are both finite dimensional vector spaces and*

$$\dim V = \dim W + \dim V/W. \tag{1.41}$$

Proof: If $W \subset V$ and if $\mathbf{w}_1, \ldots, \mathbf{w}_m$ is a basis for W then there is a basis for V of the form $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n$ for some vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. Indeed if $W \subset V$ but $W \neq V$ then there is at least one vector $\mathbf{v}_1 \in V \sim W$, and the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_1$ are linearly independent; if these vectors do not span V then the argument can be repeated, providing a vector $\mathbf{v}_2$ such that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_1, \mathbf{v}_2$ are linearly independent, and the argument can be continued. The Basis Theorem shows that this process finally stops; for since V is finite dimensional it has a basis of $m + n$ vectors for some n so there can be no more than $m + n$ linearly independent vectors in V. This argument

starting from the empty set also gives a basis for any subspace of V. Vectors in the quotient space V/W are equivalence classes $\mathbf{v} + W$ for vectors $\mathbf{v} \in V$, and since $\mathbf{w}_i + W = W$ it follows that the equivalence classes $\mathbf{v}_i + W$ for $1 \leq i \leq n$ span V/W. If $\sum_{i=1}^n a_i(\mathbf{v}_i + W) = W$ for some $a_i \in F$ then $\sum_{i=1}^n a_i \mathbf{v}_i \in W$ so that $\sum_{i=1}^n a_i \mathbf{v}_i = \sum_{j=1}^m b_i \mathbf{w}_i$, which can only be the case if $a_i = b_j = 0$ for all i, j, since the vectors $\mathbf{w}_j, \mathbf{v}_i$ are linearly independent; that means that the equivalence classes $\mathbf{v}_i + W$ for $1 \leq i \leq n$ are linearly independent vectors in V/W. These vectors thus are a basis for the vector space so $\dim V/W = n = \dim V - \dim W$ since $\dim V = m + n$ and $\dim W = m$; and that demonstrates (1.41). It follows immediately from this that the vector spaces W and V/W are finite dimensional, which concludes the proof.

For example, the vector space F^n has a canonical basis that can be described in terms of the **Kronecker symbol**

$$\delta_j^i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{1.42}$$

and the associated **Kronecker vector**

$$\boldsymbol{\delta}_j = \begin{pmatrix} \delta_j^1 \\ \delta_j^2 \\ \vdots \\ \delta_j^n \end{pmatrix} \in F^n, \tag{1.43}$$

the column vector for which all coordinates are 0 except for the j-th coordinate which is 1. It is clear that the n vectors $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_n$ are linearly independent and span the vector space F^n hence are a basis for that vector space F^n; indeed any vector (1.34) can be written uniquely in terms of this basis as

$$\mathbf{v} = \sum_{j=1}^n x_j \boldsymbol{\delta}_j. \tag{1.44}$$

A **homomorphism** or **linear transformation** between vector spaces V and W over the same field F is defined to be a mapping $T : V \longrightarrow W$ such that

(i) $T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2)$ for all $\mathbf{v}_1, \mathbf{v}_2 \in V$, and
(ii) $T(a\mathbf{v}) = aT(\mathbf{v})$ for all $a \in F$ and $\mathbf{v} \in V$.

A linear transformation $T : V \longrightarrow W$ is called an **injective linear transformation** if it is an injective mapping when viewed as a mapping between these two sets; and it is called a **surjective linear transformation** if it is a

surjective mapping when viewed as a mapping between these two sets. A linear transformation $T : V \longrightarrow W$ that is both injective and surjective is called an **isomorphism** between the two vector spaces. A particularly simple example of an isomorphism is the **identity** linear transformation $I : V \longrightarrow V$ from a vector space to itself, the mapping defined by $I(\mathbf{v}) = \mathbf{v}$ for every vector $\mathbf{v} \in V$. That two vector spaces V and W are isomorphic is denoted by $V \cong W$, which indicates that there is some isomorphism between the two vector spaces but does not specify the isomorphism. It is clear that this is an equivalence relation between vector spaces over the field F.

The **kernel** or **null space** of a linear transformation $T : V \longrightarrow W$ is the subset

$$\ker(T) = \left\{ \mathbf{v} \in V \middle| T(\mathbf{v}) = \mathbf{0} \right\} \subset V; \tag{1.45}$$

it is clear that $\ker(T) \subset V$ is a linear subspace of V, for if $\mathbf{v}_1, \mathbf{v}_2 \in \ker(T)$ then $T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2) = \mathbf{0} + \mathbf{0} = \mathbf{0}$ so $\mathbf{v}_1 + \mathbf{v}_2 \in \ker(T)$ and $T(a\mathbf{v}_1) = aT(\mathbf{v}_1) = a\mathbf{0} = \mathbf{0}$ so $a\mathbf{v}_1 \in \ker(T)$.

Lemma 1.9. *A linear transformation $T : V \longrightarrow W$ is injective if and only if* $\ker(T) = \mathbf{0}$.

Proof: If $\ker(T) = \mathbf{0}$ and if $T(\mathbf{v}_1) = T(\mathbf{v}_2)$ for two vectors $\mathbf{v}_1, \mathbf{v}_2 \in V$ then $T(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$ so $(\mathbf{v}_1 - \mathbf{v}_2) \in \ker(T) = \mathbf{0}$ hence $\mathbf{v}_1 = \mathbf{v}_2$, so T is an injective mapping. Conversely if a linear transformation $T : V \longrightarrow W$ is an injective mapping then $T^{-1}(\mathbf{0})$ is a single point of V hence $\ker(T) = \mathbf{0}$.

The **image** of a linear transformation $T : V \longrightarrow W$ is the subset

$$T(V) = \left\{ T(\mathbf{v}) \middle| \mathbf{v} \in V \right\} \subset W, \tag{1.46}$$

the image of the mapping T in the usual sense. It is clear that $T(V) \subset W$ is a linear subspace of W; for if $\mathbf{w}_1, \mathbf{w}_2 \in T(V)$ then $\mathbf{w}_1 = T(\mathbf{v}_1)$ and $\mathbf{w}_2 = T(\mathbf{v}_2)$ for some vectors $\mathbf{v}_1, \mathbf{v}_2 \in V$ so $\mathbf{w}_1 + \mathbf{w}_2 = T(\mathbf{v}_1) + T(\mathbf{v}_2) = T(\mathbf{v}_1 + \mathbf{v}_2) \in T(V)$ and $a\mathbf{w}_1 = aT(\mathbf{v}_1) = T(a\mathbf{v}_1) \in T(V)$ for any $a \in F$. Clearly a linear transformation $T : V \longrightarrow W$ is surjective precisely when its image is the full vector space W.

Theorem 1.10. *If $T : V \longrightarrow W$ is a linear transformation between two finite dimensional vector spaces over a field F then T induces an isomorphism*

$$T^* : V/\ker(T) \xrightarrow{\cong} T(V) \quad \textit{where} \quad T(V) \subset W, \tag{1.47}$$

so that

$$\dim V = \dim \ker(T) + \dim T(V). \tag{1.48}$$

In particular T itself is an isomorphism if and only if $\ker(T) = \mathbf{0}$ *and* $T(V) = W$.

Proof: If $V_0 = \ker(T)$ then V_0 is a linear subspace of V. The elements of the quotient vector space V/V_0 are the subsets $\mathbf{v} + V_0 \subset V$; and since $T(V_0) = \mathbf{0}$ the linear transformation T maps each of these subsets to the same element $T(\mathbf{v} + V_0) = T(\mathbf{v}) \in W$ and in that way T determines a mapping $T^* : V/V_0 \longrightarrow W$ which is easily seen to be a linear transformation having as its image the linear subspace $T(V) \subset W$. An element $\mathbf{v} + V_0 \in V/V_0$ is in the kernel of the linear transformation T^* if and only if $T(\mathbf{v} + V_0) = T(\mathbf{v}) = \mathbf{0}$, which is just the condition that $\mathbf{v} \in V_0$ hence that $\mathbf{v} + V_0 = V_0$ is the zero vector in V/V_0. That shows that the linear transformation (1.47) is an isomorphism, and (1.48) follows in view of Theorem 1.8 while it is clear that T itself is an isomorphism if and only if $\ker(T) = \mathbf{0}$ and $T(V) = W$, which suffices for the proof.

The explicit descriptions of linear transformations and the determination of their properties for the special case of mappings between finite dimensional vector spaces rest on their effect on bases of the vector spaces, as in the following observations.

Theorem 1.11 (Mapping Theorem). *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis for a vector space V and $\mathbf{w}_1, \ldots, \mathbf{w}_m$ be a basis for a vector space W over a field F.*
(i) *Any linear transformation $T : V \longrightarrow W$ is determined fully by the images $T(\mathbf{v}_i)$ of the basis vectors $\mathbf{v}_i$; and a linear transformation $T : V \longrightarrow W$ can be defined by specifying as the images $T(\mathbf{v}_j)$ any arbitrary n vectors in W.*
(ii) *The linear transformation defined by the values $T(\mathbf{v}_i) \in W$ is injective if and only if the vectors $T(\mathbf{v}_i)$ are linearly independent.*
(iii) *The linear transformation defined by the values $T(\mathbf{v}_i) \in W$ is surjective if and only if the vectors $T(\mathbf{v}_i)$ span W.*
(iv) *Two vector spaces over a field F are isomorphic if and only if they have the same dimension.*

Proof: (i) If $T : V \longrightarrow W$ is a linear transformation then

$$T\left(\sum_{j=1}^{n} a_j \mathbf{v}_j\right) = \sum_{j=1}^{n} a_j T(\mathbf{v}_j), \tag{1.49}$$

hence the linear transformation T is determined fully by the images $T(\mathbf{v}_i)$; and for any specified vectors $T(\mathbf{v}_i) \in W$ the linear transformation defined by (1.49) is a linear transformation taking these values.
(ii) The linear transformation T determined by the images $T(\mathbf{v}_i)$ fails to be injective if and only if there are scalars $a_j \in F$, not all of which are 0, such that

$$\mathbf{0} = T\left(\sum_{j=1}^{n} a_j \mathbf{v}_j\right) = \sum_{j=1}^{n} a_j T(\mathbf{v}_j);$$

and that is precisely the condition that the vectors $T(\mathbf{v}_j)$ are linearly dependent.
(iii) Since the linear transformation T determined by the images $T(\mathbf{v}_i)$ is given by (1.49) it is evident that the linear transformation is surjective if and only if the vectors $T(\mathbf{v}_j)$ span W.
(iv) If $T : V \longrightarrow W$ is an isomorphism it must be both injective and surjective, and it then follows from the preceding parts of this theorem that the vectors $T(\mathbf{v}_i)$ are a basis for W and consequently $\dim W = \dim V$. On the other hand if $\dim V = \dim W = n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis for V while $\mathbf{w}_1, \ldots, \mathbf{w}_n$ is a basis for W then by the preceding part of this theorem there is a linear transformation $T : V \longrightarrow W$ sending each $\mathbf{v}_i$ to $\mathbf{w}_i$, and it is both injective and surjective so is an isomorphism of vector spaces. That suffices for the proof.

The set of all linear transformations from a vector space V to a vector space W is denoted by $\mathcal{L}(V, W)$; of course this is defined only for two vector spaces over the same field F. The sum of two linear transformations $S, T \in \mathcal{L}(V, W)$ is the mapping defined by $(S + T)(\mathbf{x}) = S(\mathbf{x}) + T(\mathbf{x})$ for any vector $\mathbf{x} \in V$, and is clearly itself a linear transformation $(S + T) \in \mathcal{L}(V, W)$; and the scalar product of an element $c \in F$ and a linear transformation $T \in \mathcal{L}(V, W)$ is the mapping defined by $(cT)(\mathbf{x}) = cT(\mathbf{x})$ for any vector $\mathbf{x} \in V$, and is clearly also a linear transformation. The distributive law obviously holds for addition and scalar multiplications, so the set $\mathcal{L}(V, W)$ is another vector space over the field F. Linear transformations are mappings from one space to another, so as in the case of any mappings the **composition** ST of two mappings S and T is the mapping defined by $(ST)(\mathbf{v}) = S(T(\mathbf{v}))$; the composition also is a linear transformation since $(ST)(\mathbf{v}_1 + \mathbf{v}_2) = S(T(\mathbf{v}_1 + \mathbf{v}_2)) = S(T(\mathbf{v}_1) + T(\mathbf{v}_2)) = S(T(\mathbf{v}_1)) + S(T(\mathbf{v}_2)) = ST(\mathbf{v}_1) + ST(\mathbf{v}_2)$ for any vectors $\mathbf{v}_1, \mathbf{v}_2 \in V$ and $(ST)(a\mathbf{v}) = S(T(a\mathbf{v})) = S(aT(\mathbf{v})) = aS(T(\mathbf{v})) = aST(\mathbf{v})$ for any vectors $\mathbf{v} \in V$ and any scalar $a \in F$.

Theorem 1.12. *A linear transformation $T : V \longrightarrow W$ between vector spaces V and W over a field F is an isomorphism if and only if there is a linear transformation $S : W \longrightarrow V$ such that $ST : V \longrightarrow V$ and $TS : W \longrightarrow W$ are the identity linear transformations, $ST = TS = I$.*

Proof: If $T : V \longrightarrow W$ is an isomorphism of vector spaces then T is surjective so for any vector $\mathbf{w} \in W$ there will be a vector $\mathbf{v} \in V$ such that $T(\mathbf{v}) = \mathbf{w}$; and T is injective so the vector $\mathbf{v}$ is uniquely determined and consequently can be viewed as a function $\mathbf{v} = S(\mathbf{w})$ of the vector $\mathbf{w} \in W$, thus defining a mapping $S : W \longrightarrow V$ for which $TS(\mathbf{w}) = T(\mathbf{v}) = \mathbf{w}$ so that $TS = I$. For any vector $\mathbf{v} \in V$ since $\mathbf{w} = T(\mathbf{v}) \in W$ it follows that $T(\mathbf{v}) = \mathbf{w} = TS(\mathbf{w}) = TST(\mathbf{v})$, so since T is injective necessarily $\mathbf{v} = ST(\mathbf{v})$ and consequently $ST = I$. The mapping S is a linear transformation; for if $\mathbf{w}_1 = T(\mathbf{v}_1)$ and $\mathbf{w}_2 = T(\mathbf{v}_2)$ then by linearity

$\mathbf{w}_1 + \mathbf{w}_2 = T(\mathbf{v}_1 + \mathbf{v}_2)$ hence $S(\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{v}_1 + \mathbf{v}_2 = S(\mathbf{w}_1) + S(\mathbf{w}_2)$ and if $T(\mathbf{v}) = \mathbf{w}$ then also by linearity $T(a\mathbf{v}) = aT(\mathbf{v}) = a\mathbf{w}$ so $S(a\mathbf{w}) = aS(\mathbf{w})$ for any scalar $a \in F$.

Conversely suppose that there is a linear transformation S such that $ST = I$ and $TS = I$. For any vector $\mathbf{w} \in W$ it follows that $\mathbf{w} = TS(\mathbf{w})$ so that $\mathbf{w}$ is in the image of T hence T is surjective. If $T(\mathbf{v}) = \mathbf{0}$ for a vector $\mathbf{v} \in V$ then $\mathbf{v} = ST(\mathbf{v}) = S(\mathbf{0}) = \mathbf{0}$ so that the kernel of T is the zero vector hence T is injective, which suffices for the proof.

Theorem 1.13. *Any n-dimensional vector space V over the field F is isomorphic to the vector space F^n.*

Proof: If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis for V then by the Mapping Theorem there is a linear transformation $T : F^n \longrightarrow V$ defined by $T(\boldsymbol{\delta}_j) = \mathbf{v}_j$. By the Mapping Theorem again the linear transformation T is injective, since the image vectors $T(\boldsymbol{\delta}_j) = \mathbf{v}_j$ are linearly independent, and surjective, since the image vectors $T(\boldsymbol{\delta}_j) = \mathbf{v}_j$ span the vector space V; hence the linear transformation is an isomorphism between the vector spaces F^n and V.

Thus the vector space F^n for a field F is not only a simple example of an n-dimensional vector space over the field F but is also a standard model for any such vector space. The choice of an isomorphism $T : V \longrightarrow F^n$ for a vector space V over a field F can be viewed as a **coordinate system** in V, for it permits any vector $\mathbf{v} \in V$ to be described by its image $T(\mathbf{v}) = \mathbf{x} = \{x_i\} \in F^n$, so to be described by the coordinates $x_i \in F$. A coordinate system for a vector space V is determined by the choice of a basis for V; and since there are a great many bases there are also a great many different coordinate systems for V. It is clear though that if $S, T : V \longrightarrow F^n$ are any two coordinate systems for a vector space V then $T = US$ where $U = TS^{-1} : F^n \longrightarrow F^n$ is an isomorphism of the vector space F^n; and conversely if $S : V \longrightarrow F^n$ is a coordinate system for the vector space V and $U : F^n \longrightarrow F^n$ is an isomorphism of the vector space F^n, then $T = US : V \longrightarrow F^n$ also is a coordinate system for the vector space V. Actually for many purposes it is sufficient just to consider the vector spaces F^n rather than abstract vector spaces over the field F, since some general properties of finite dimensional vector spaces can be proved by a simple calculation for the vector spaces F^n.

A linear transformation $T : F^m \longrightarrow F^n$ can be described quite explicitly in a way that is very convenient for calculation. The image of a basis vector $\boldsymbol{\delta}_j \in F^m$ is a linear combination of the basis vectors $\boldsymbol{\delta}_i \in F^n$, so

$$T\left(\boldsymbol{\delta}_j\right) = \sum_{i=1}^{n} a_{ij}\boldsymbol{\delta}_i \tag{1.50}$$

for some scalars $a_{ij} \in F$; hence by linearity the image of an arbitrary vector $\mathbf{v} = \sum_{j=1}^{m} x_j \boldsymbol{\delta}_j \in F^m$ with the coordinates $\{x_j\}$ is the vector

$$\mathbf{w} = T(\mathbf{v}) = \sum_{j=1}^{m} x_j T(\boldsymbol{\delta}_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} x_j a_{ij} \boldsymbol{\delta}_i = \sum_{i=1}^{n} y_i \boldsymbol{\delta}_i \tag{1.51}$$

with the coordinates $\{y_i\}$ given by

$$y_i = \sum_{j=1}^{m} a_{ij} x_j. \tag{1.52}$$

Thus the linear transformation T is described completely by the set of scalars $a_{ij} \in F$. It is customary and convenient to list these scalars in an array of the form

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\ & & & \cdots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{pmatrix} \tag{1.53}$$

called an $n \times m$ **matrix** over the field F; an $n \times m$ matrix A with entries in the field F thus describes a linear trasformation $T : F^m \longrightarrow F^n$, and that convention should be kept firmly in mind. The horizontal lines are called the **rows** of the matrix and the vertical lines are called the **columns** of the matrix, so an $n \times m$ matrix can be viewed as a collection of n row vectors or alternatively as a collection of m column vectors. The scalars a_{ij} are called the **entries** of the matrix, so $A = \{a_{ij}\}$ where a_{ij} is the entry in row i and column j. The linear relation between the coordinates $\{x_j\}$ of a vector in F^m and the coordinates $\{y_i\}$ of the image vector in F^n is (1.52). The preceding conventions and terminology are rather rigidly followed so should be kept in mind and used carefully and systematically.

Associating to a linear transformation $T : F^m \longrightarrow F^n$ the $n \times m$ matrix describing that linear transformation identifies the vector space $\mathcal{L}(F^m, F^n)$ of all linear transformations with the vector space of all $n \times m$ matrices over the field F, since that association clearly preserves addition of vectors and scalar multiplication; thus there is the natural isomorphism $\mathcal{L}(F^m, F^n) \cong F^{nm}$, since an $n \times m$ matrix is merely a collection of nm scalars, and consequently

$$\dim \mathcal{L}(F^m, F^n) = nm. \tag{1.54}$$

When the vector space F^{nm} is identified in this way with the vector space $\mathcal{L}(F^m, F^n)$ it is often denoted by $F^{n \times m}$; so this stands for the vector space of

matrices over the field F having n rows and m columns. The zero vector in $F^{n\times m}$ is described by the **zero matrix 0**, the matrix having all entries 0; this represents the **trivial** linear transformation that maps all vectors in F^m to the zero vector in F^n. The matrix $I \in F^{n\times n}$ describing the identity linear transformation $F^n \longrightarrow F^n$, the linear transformation that takes the basis vectors $\boldsymbol{\delta}_i$ to themselves, clearly has the entries $I = \{\delta^i_j\}$, the Kronecker symbols; it is called the **identity matrix**. It is often convenient to specify the size of an identity matrix, so to denote the $n \times n$ identity matrix by I_n; thus for instance

$$I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{1.55}$$

The equation (1.52) describing the relations between the coordinate of vectors $\mathbf{v} = \{x_j\}$ and $\mathbf{w} = \{y_i\}$ related by the linear transformation $\mathbf{w} = T\mathbf{v}$ described by a matrix A can be viewed as a collection of linear equations relating the variables x_i and y_j. This system of equations often is viewed as a matrix product, expressing the column vector or $n \times 1$ matrix $\mathbf{y} = \{y_i\}$ as the product of the $n \times m$ matrix A and the column vector or $m \times 1$ matrix $\mathbf{x} = \{x_j\}$ in the form

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\ & & \cdots & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_m \end{pmatrix} \tag{1.56}$$

where the entry y_i in row i of the product is the sum over the index j of the products of the entries a_{ij} in row i of the matrix A and the successive entries x_j of the column vector $\mathbf{x}$, so

$$y_i = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \cdots + a_{im}x_m.$$

More generally the **product** of an $n \times m$ matrix $A = \{a_{ij}\}$ and an $m \times l$ matrix $B = \{b_{ij}\}$ is defined to be the $n \times l$ matrix $AB = C = \{c_{ij}\}$ with entries

$$c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj} \quad \text{for } 1 \le i \le n,\ 1 \le j \le l; \tag{1.57}$$

so column j of the product matrix $AB = C$ is the column vector $\mathbf{c}_j = A\ \mathbf{b}_j$ where $\mathbf{b}_j$ is column j of the matrix B, and row j of the product matrix $AB = C$ is the

row vector ${}^t\mathbf{c}_j = {}^t\mathbf{a}_j B$ where ${}^t\mathbf{a}_j$ is row j of the matrix A. Matrices A and B can be multiplied to yield a product only when the number of columns of the matrix A is equal to the number of rows of the matrix B; and the number of rows of the product AB is the number of rows of A while the number of columns of the product AB is the number of columns of the matrix B. For example in the product

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ a_{31} & a_{32} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

the first column of the product $AB = C$ is the product of the matrix A and the first column of the matrix B; explicitly

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}$$
$$c_{21} = a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31}.$$

It is easy to see from the definition that the matrix product is associative, in the sense that $A(BC) = (AB)C$ and consequently that multiple products such as ABC are well defined, and that it is distributive, in the sense that $A(B + C) = AB + BC$ and $(A + B)C = AC + BC$. It is worth calculating a few matrix products just to become familiar with the technique.

Theorem 1.14. *If a linear transformation $T : F^n \longrightarrow F^m$ is described by an $m \times n$ matrix B and a linear transformation $S : F^m \longrightarrow F^l$ is described by $l \times m$ matrix A then the composition $S \circ T : F^n \longrightarrow F^l$ is described by the product matrix AB.*

Proof: For any index $1 \leq i \leq n$

$$(S \circ T)(\boldsymbol{\delta}_i) = S\big(T(\boldsymbol{\delta}_i)\big) = S\left(\sum_{k=1}^{m} b_{ki}\boldsymbol{\delta}_k\right) = \sum_{k=1}^{m} b_{ki}S(\boldsymbol{\delta}_k)$$
$$= \sum_{k=1}^{m} b_{ki}\sum_{j=1}^{l} a_{jk}\boldsymbol{\delta}_j = \sum_{j=1}^{l}\sum_{k=1}^{m} a_{jk}b_{ki}\boldsymbol{\delta}_j = \sum_{j=1}^{l} c_{ji}\boldsymbol{\delta}_j;$$

thus the composition $S \circ T$ is described by the matrix $C = \{c_{ji}\}$ where $c_{ji} = \sum_{k=1}^{m} a_{jk}b_{ki}$ so that $C = AB$, which suffices for the proof.

In view of the preceding theorem, the linear transformation described by a matrix A is usually also denoted by A, and AB denotes both the matrix product and the composition $A \circ B$ of the linear transformations described by these matrices. This convention will be followed henceforth. It follows from Theorem 1.12 that the linear transformation described by a matrix A is an isomorphism if and only if that linear transformation has an inverse linear transformation, or in matrix terms, if and only if there is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, the identity matrix; such a matrix A is called an **invertible**

matrix or a **nonsingular matrix**. A matrix that is not nonsingular of course is called a **singular** matrix. Not all matrices in $F^{n\times n}$ have inverses; the zero matrix of course has no inverse, nor does the matrix $A = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ since $AA = 0$, so these are examples of **singular** matrices. On the other hand the identity matrix is nonsingular, since it is its own inverse, while if A is nonsingular then so is A^{-1} with $(A^{-1})^{-1} = A$; and if A and B are nonsingular matrices then so is their product since clearly $(AB)(B^{-1}A^{-1}) = I$ hence $B^{-1}A^{-1} = (AB)^{-1}$. Thus the set of nonsingular $n \times n$ matrices is a group under multiplication called the **general linear group** over the field F and denoted by $\mathrm{Gl}(n, F)$.

Some other notational conventions involving matrices are widely used. In extension of the definition (1.35) of the transpose of a vector, the **transpose** of an $n \times m$ matrix $A = \{a_{ij}\}$ is defined to be the $m \times n$ matrix

$$ {}^tA = B = \{b_{ij}\} \quad \text{where} \quad b_{ij} = a_{ji}; \tag{1.58} $$

thus if $\mathbf{a}_i$ is the i-th column vector of the matrix A then ${}^t\mathbf{a}_i$ is the i-th row vector of the matrix ${}^tA$. If the matrix A describes a linear transformation $A : F^n \longrightarrow F^m$ then the transposed matrix ${}^tA$ describes a linear transformation ${}^tA : F^m \longrightarrow F^n$; but note that ${}^tA$ does not describe the inverse linear transformation to that described by A but rather something quite different altogether, since for instance the inverse of a linear transformation between two vector spaces of different dimensions is not even well defined. If $C = AB$ the entries of these matrices are related by $c_{ij} = \sum_k a_{ik}b_{kj}$, and that equation can be interpreted alternatively as ${}^tC = {}^tB\,{}^tA$, so transposition reverses the order of matrix multiplication. Another notational convention is that a matrix can be decomposed into **matrix blocks** by splitting the rectangular array of its entries into subarrays. Thus for example an $n \times m$ matrix $A = \{a_{ij}\}$ can be decomposed into 4 matrix blocks

$$ A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} $$

where A_{11} is a $k \times k$ matrix, A_{12} is a $k \times (m-k)$ matrix, A_{21} is an $(n-k) \times k$ matrix and A_{22} is an $(n-k) \times (m-k)$ matrix. A decomposition of this form in which A_{12} and A_{21} are both zero matrices is called a **direct sum** decomposition, denoted by $A = A_{11} \oplus A_{22}$. An $n \times n$ matrix A which has the direct sum decomposition $A = A_{11} \oplus A_{22} \oplus \cdots \oplus A_{nn}$ in which each of the component matrices A_{ii} is a 1×1 matrix, just a scalar, is called a **diagonal matrix** since all of its terms vanish aside from those terms along the **main diagonal**, as in

$$ \begin{pmatrix} a_1 & 0 & 0 & 0 \\ 0 & a_2 & 0 & 0 \\ 0 & 0 & a_3 & 0 \\ 0 & 0 & 0 & a_4 \end{pmatrix}. \tag{1.59} $$

Another special form of a matrix is a **lower triangular** matrix, a matrix $A = \{a_{ij}\}$ such that $a_{ij} = 0$ whenever $i < j$, as for example

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}; \tag{1.60}$$

an **upper triangular** matrix of course is defined correspondingly as a matrix $A = \{a_{ij}\}$ such that $a_{ij} = 0$ whenever $i > j$.

The direct sum $F^m \oplus F^n$ of vector spaces F^m and F^n was defined in (1.39) to be the set of pairs $(\mathbf{x}, \mathbf{y})$ of vectors $\mathbf{x} \in F^m$ and $\mathbf{y} \in F^n$; it is customary though in this context to write this pair as $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$, thereby naturally identifying the direct sum $F^m \oplus F^n$ with the vector space F^{m+n}. An $r \times m$ matrix A describes a linear transformation $A : F^m \longrightarrow F^r$ that takes the vector $\mathbf{x} \in F^m$ to the vector $A\mathbf{x} \in F^r$; and correspondingly an $s \times n$ matrix B describes a linear transformation $B : F^n \longrightarrow F^s$ that takes the vector $\mathbf{y} \in F^n$ to the vector $B\mathbf{y} \in F^s$. These two linear transformations can be combined into the direct sum $A \oplus B$, an $(r + s) \times (m + n)$ matrix, which acts on the direct sum $F^m \oplus F^n$ through the matrix product

$$\begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & B \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} A\mathbf{x} \\ B\mathbf{y} \end{pmatrix}; \tag{1.61}$$

thus direct sums of matrices act naturally on direct sums of vector spaces indicating the compatibility of the two notions of direct sum.

Two $n \times m$ matrices A and B over a field F are said to be **equivalent matrices** if there are a nonsingular $n \times n$ matrix S and a nonsingular $m \times m$ matrix T such that $B = SAT$. This is a basic equivalence relation among matrices; but there are other equivalence relations of equal importance and even greater interest that will be discussed in Section 4.2. It is clear that this is an equivalence relation among $n \times m$ matrices; indeed reflexivity and symmetry are quite obvious, and if B is equivalent to A so that $B = SAT$ for nonsingular matrices S and T and if C is equivalent to B so that $C = UBV$ for nonsingular matrices U and V then $C = (US)A(TV)$ where US and TV are nonsingular matrices. If a matrix A is viewed as a mapping $A : F^m \longrightarrow F^n$ that takes a vector $\mathbf{x} \in F^m$ to the vector $\mathbf{y} = A\mathbf{x} \in F^n$, it is natural to ask what matrix represents this linear transformation after changes of coordinate in F^m and F^n. Thus an isomorphism $T : F^m \longrightarrow F^m$ takes the vector $\mathbf{x}'$ to the vector $\mathbf{x} = T\mathbf{x}'$, and an isomorphism $S : F^n \longrightarrow F^n$ takes the vector $\mathbf{y}$ to the vector $\mathbf{y}' = S\mathbf{y}$; and making these substitutions in the equation $\mathbf{y} = A\mathbf{x}$ yields the equation $\mathbf{y}' = S\mathbf{y} = SA\mathbf{x} = SAT\mathbf{x}' = B\mathbf{x}'$ where $B = SAT \cong A$. In this sense equivalent matrices describe the same linear transformation but expressed in different coordinates. Alternatively, the equivalence of matrices A and B can be

expressed in terms of the linear transformations defined by these matrices as the assertion that the following diagram is commutative:

$$\begin{array}{ccc} F^m & \xrightarrow{\ A\ } & F^n \\ {\scriptstyle T}\uparrow \text{iso} & & {\scriptstyle S}\downarrow \text{iso} \\ F^m & \xrightarrow{\ B\ } & F^n. \end{array} \tag{1.62}$$

The arrows in the diagram indicate mappings, in this case linear transformations, labeled by the letters A, B, S, T; and the notation "iso" indicates that the two linear transformations S and T are isomorphisms. A diagram such as this is said to be **commutative** if any two sequences of mappings in the diagram that begin and end at the same point are equal; in this case that can only mean that the mapping in the path B is equal to the result of first applying the mapping T, then the mapping A, and finally the mapping S, which is just the assertion that $B = SAT$. Commutativity of diagrams is a very commonly used notion that can simplify considerably statements of the equality of various combinations of mappings. However the notion of equivalence of matrices is viewed, though, it leads to a natural classification problem: does an equivalence class of matrices have a simple standard representative, and are there simple invariants that determine when two matrices are equivalent? That is the motivation for the next segment of the discussion.

Theorem 1.15. *If $A \in F^{n\times m}$ is a matrix describing a linear transformation $A : F^m \longrightarrow F^n$ then there are a nonsingular $n \times n$ matrix S and a nonsingular $m \times m$ matrix T such that the matrix $SAT = B = \{b_{ij}\}$ has the entries*

$$b_{ij} = \begin{cases} \delta^i_j & \text{for } 1 \le i, j \le k, \\ 0 & \text{otherwise}, \end{cases} \tag{1.63}$$

where $m - k$ is the dimension of the kernel of A; thus B is the $n \times m$ matrix

$$B = \begin{pmatrix} I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{1.64}$$

Proof: If the kernel of the linear transformation has dimension $m - k$ choose vectors $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_m$ in F^m that form a basis for the kernel, and extend this set of vectors to a basis $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_m$ for the full vector space F^m. The image vectors $\mathbf{w}_i = A\mathbf{v}_i \in F^n$ for $1 \le i \le k$ are linearly independent; for if $\mathbf{0} = \sum_{i=1}^k c_i\, A\mathbf{v}_i = A(\sum_{i=1}^k c_i\, \mathbf{v}_i)$ where not all the scalars c_i vanish then $\sum_{i=1}^k c_i\, \mathbf{v}_i$ is a nontrivial vector in $\ker(A)$, but that is impossible since this vector is not in the span of the basis $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_m$ for $\ker(A)$. Extend the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ to a basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$ for the vector space F^n. Introduce the isomorphisms $T : F^m \longrightarrow F^m$

for which $T\boldsymbol{\delta}_i = \mathbf{v}_i$ and $S : F^n \longrightarrow F^n$ for which $S\mathbf{w}_i = \boldsymbol{\delta}_i$, and let $B = SAT : F^m \longrightarrow F^n$. If $1 \leq i \leq k$ then $B\boldsymbol{\delta}_i = SAT\boldsymbol{\delta}_i = SA\mathbf{v}_i = S\mathbf{w}_i = \boldsymbol{\delta}_i$, while on the other hand if $k+1 \leq i \leq m$ then $B\boldsymbol{\delta}_i = SAT\boldsymbol{\delta}_i = SA\mathbf{v}_i = \mathbf{0}$; that means that the entries of the matrix B are $b_{ji} = \delta_i^j$ for $1 \leq i \leq k$ and $b_{ji} = 0$ otherwise, so that the matrix B has the form (1.64), and that suffices for the proof.

There remain the questions whether the normal form (1.64) is uniquely determined, and if so whether it can be calculated more directly. To discuss these matters it is useful to introduce a further standard convention in linear algebra. When an $n \times m$ matrix A is viewed as a collection of m column vectors $A = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m)$, where $\mathbf{a}_j \in F^n$, the linear subspace of F^n spanned by these column vectors $\mathbf{a}_j$ is called the **column space** of the matrix A and the dimension of this subspace is called the **column rank** of the matrix A, denoted by $\mathrm{crank}(A)$. The column space of course is just the image $AF^m \subset F^n$, since $A\mathbf{x} = \sum_{i=1}^m \mathbf{a}_j x_j$ for any vector $\mathbf{x} \in F^m$, so the column rank of A is just the dimension of the image AF^m. The **row space** of the matrix A and its **row rank** $\mathrm{rrank}(A)$ are defined correspondingly; or alternatively since the rows of a matrix A are the columns of the transposed matrix ${}^tA$ the row space of A is the transpose of the column space of ${}^tA$ so that $\mathrm{rrank}(A) = \mathrm{crank}({}^tA)$.

Theorem 1.16. *Let A be an $n \times m$ matrix over a field F.*
(i) $\mathrm{crank}(SA) = \mathrm{crank}(A)$ *and* $\mathrm{rrank}(SA) = \mathrm{rrank}(A)$ *for any nonsingular $n \times n$ matrix S.*
(ii) $\mathrm{crank}(AT) = \mathrm{crank}(A)$ *and* $\mathrm{rrank}(AT) = \mathrm{rrank}(A)$ *for any nonsingular $m \times m$ matrix T.*
(iii) $\mathrm{rrank}(A) = \mathrm{crank}(A)$ *for any matrix A.*

Proof: (i) If S is a nonsingular $n \times n$ matrix the mapping $S : F^n \longrightarrow F^n$ that takes a vector $\mathbf{v}$ to $S\mathbf{v}$ is an isomorphism of vector spaces. Therefore the column space of the matrix A, the subspace of F^n spanned by the column vectors $\mathbf{a}_j$ of the matrix A, is isomorphic to the column space of the matrix SA, the subspace of F^n spanned by the column vectors $S\mathbf{a}_j$ of the matrix SA, or equivalently $\mathrm{crank}(SA) = \mathrm{crank}(A)$. On the other hand since the entries of the product matrix $B = SA$ are $b_{ij} = \sum_{k=1}^n s_{ik}a_{kj}$ it follows that the row vectors $\mathbf{b}_i = \{\, b_{ij} \mid 1 \leq j \leq m \,\}$ of the matrix B are the linear combinations $\mathbf{b}_i = \sum_{k=1}^n s_{ik}\mathbf{a}_k$ of the row vectors $\mathbf{a}_k = \{\, a_{kj} \mid 1 \leq j \leq m\}$ of the matrix A and therefore $\mathrm{rrank}B \leq \mathrm{rrank}A$. Conversely since S is nonsingular and $A = S^{-1}B$ it is also the case that $\mathrm{rrank}B \leq \mathrm{rrank}A$, so altogether $\mathrm{rrank}(B) = \mathrm{rrank}(A)$.
(ii) If $B = AT$ then ${}^tB = {}^tT\,{}^tA$ and $\mathrm{rrank}\,{}^tB = \mathrm{crank}B$ and correspondingly for the matrix A; it therefore follows from (i) that $\mathrm{crank}(AT) = \mathrm{crank}(A)$ and $\mathrm{rrank}(AT) = \mathrm{rrank}(A)$ for any nonsingular $m \times m$ matrix T.
(iii) By Theorem 1.15, there are invertible matrices S, T so that SAT has the form (1.64); and it is obvious from that form that

$\operatorname{crank}(SAT) = \operatorname{rrank}(SAT) = k$. Since it was just observed that $\operatorname{crank}(SAT) = \operatorname{crank}(A)$ and correspondingly for the row ranks it follows immediately that $\operatorname{crank}(A) = \operatorname{rrank}(A)$, which suffices for the proof.

Since the row rank and the column rank of any matrix are always the same, the distinction between them is ordinarily ignored and the common value is just called the **rank** of the matrix and denoted by $\operatorname{rank}(A)$.

Corollary 1.17 (Equivalence Theorem). *Two $n \times m$ matrices over a field are equivalent if and only if they have the same rank; hence any $n \times m$ matrix A is equivalent to a unique matrix of the form* $\begin{pmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ *where* $r = \operatorname{rank}(A)$.

Proof: Theorem 1.15 shows that any $n \times m$ matrix is equivalent to one of the form (1.64), which clearly is of rank k; and it follows from the preceding Theorem 1.16 (i) that equivalent matrices have the same rank; consequently the normal form (1.64) is uniquely determined, by the rank, which suffices for the proof.

Corollary 1.18. *An $n \times n$ matrix over a field F is invertible if and only if* $\operatorname{rank}(A) = n$.

Proof: It is clear that a matrix of the form $\begin{pmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ is invertible if and only if $r = n$; any matrix is equivalent to a matrix of this form, by the Equivalence Theorem, and since any two equivalent matrices have the same rank by Theorem 1.16 and since any matrix equivalent to an invertible matrix is invertible that suffices for the proof.

An auxiliary tool that is required for the further discussion deals with rearrangements or permutations of a set of variables, where for example x_4, x_1, x_3, x_2 is a rearrangement or **permutation** of the set of variables x_1, x_2, x_3, x_4; thus a permutation of such a set of variables is just an element in the symmetric group $S(x_1, x_2, x_3, x_4)$. To the usual order of the variables there can be associated the polynomial $P(x) = \prod_{i<j}(x_i - x_j)$, a polynomial of degree $n(n-1)/2$ consisting of the product of the differences of any two distinct variables where the index of the first variable is less than the index of the second variable; the square $P(x)^2$ of this polynomial can be rewritten as the product of all the differences $x_i - x_j$ of distinct variables so is independent of the order of the variables. After a permutation π of the variables the resulting polynomial $\pi^* P(x)$ still consists of the products of the differences of any two distinct variables, but possibly with a reversal of the signs of the differences; for example if $P(x) = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$ then if π is the permutation that interchanges the variables x_2 and x_3 the result of applying this permutation to the polynomial is the polynomial $\pi^* P(x) = (x_1 - x_3)(x_1 - x_2)(x_3 - x_2) = -P(x)$.

In general $\pi^*P(x) = \pm P(x)$ for some sign, which is called the **sign** of the permutation π and is denoted by $\text{sgn}(\pi)$ so that $\pi^*P(x) = \text{sgn}(\pi) \cdot P(x)$. A permutation π is **even** if $\text{sgn}(\pi) = +1$ and is **odd** if $\text{sgn}(\pi) = -1$.

Theorem 1.19. (i) $\text{sgn}(\pi) = -1$ *for the permutation* $\pi \in S(x_1, \ldots, x_n)$ *that interchanges two variables* x_i *and* x_j.
(ii) *Any permutation can be written as a composition of permutations that interchange two variables.*
(iii) $\text{sgn}(\pi_1\pi_2) = \text{sgn}(\pi_1) \cdot \text{sgn}(\pi_2)$ *for any two permutations* $\pi_1, \pi_2 \in S(x_1, \ldots, x_n)$.

Proof: (i) A permutation π that interchanges x_i and x_j does not change the sign of any factor of the polynomial $P(x)$ that does not involve the variable x_i or the variable x_j. If $k < i < j$ or $i < j < k$ the product $(x_i - x_k)(x_j - x_k)$ is unchanged by the permutation; and if $i < k < j$ the product $(x_i - x_k)(x_k - x_j)$ is replaced by $(x_j - x_k)(x_k - x_i)$ so the sign of the product is unchanged. Thus the only effect of the permutation π on the sign of the polynomial $P(x)$ arises from the change in the sign of the factor $(x_i - x_j)$ hence $\pi^*P(x) = -P(x)$ so $\text{sgn}(\pi) = -1$.
(ii) Suppose a permutation π replaces the variables $x_1, x_2, \ldots, x_n$ by a rearranged set $x_{j_1}, x_{j_2}, \ldots, x_{j_n}$ of variables. Composing the permutation π with the permutation π_{1,j_1} that interchanges the variables x_1 and x_{j_1} replaces the variables $x_{j_1}, x_{j_2}, \ldots, x_{j_n}$ by $x_1, x_{k_2}, \ldots, x_{k_n}$; composing this permutation with the permutation π_{2,k_2} replaces the variables $x_1, x_{k_2}, \ldots, x_{k_n}$ by $x_1, x_2, x_{l_3}, \ldots, x_{l_n}$, and so on. Thus the composition of the permutation π with the composition of these permutations which interchange two variables leads to the identity permutation.
(iii) For any two permutations π_1, π_2 of the variables $\pi_2^*P(x) = \text{sgn}(\pi_2)P(x)$ by definition; applying the permutation π_1 to both sides of this equality yields $\pi_1^*(\pi_2^*P(x)) = \text{sgn}(\pi_2) \cdot \pi_1^*P(x) = \text{sgn}(\pi_2)\text{sgn}(\pi_1)P(x)$, by definition. On the other hand $\pi_1^*(\pi_2^*P(x))$ is the result of applying first the permutation π_2 and then the permutation π_1 so is the result of applying the composite permutation $\pi_1 \circ \pi_2$, and consequently $\pi_1^*(\pi_2^*P(x)) = \text{sgn}(\pi_1 \circ \pi_2)P(x)$, and that suffices for the proof.

The preceding theorem does provide a method that can be useful for determining whether a permutation is odd or even. By that theorem any permutation can be written as a composition of **transpositions**, permutations that interchange two variables, while each transposition has sign -1 and the sign of the permutation is the product of the signs of these transpositions. It follows that a permutation is even if it can be written as a product of an even number of transpositions and is odd if it can be written as a product of an odd number of transpositions. A permutation can be written as a product of transpositions in a number of different ways; but the theorem ensures that the parity of the total number of transpositions is an invariant of the permutation.

A linear transformation $T : V \longrightarrow F$ from a vector space V over a field F to the field F itself is often called a **linear function**. Also of interest of course are linear functions of several variables, or **multilinear functions**, mappings $T : V^k \longrightarrow F$ from Cartesian products V^k of a vector space V over the field F to the field F that are linear transformations in each separate variable; thus a multilinear function $T : V^k \longrightarrow F$ associates to any k vectors $\mathbf{v}_i \in V$ a value $T(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) \in F$, where $T(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ is a linear function of the vector $\mathbf{v}_j \in V$ whenever the remaining vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_k$ are held fixed. A multilinear function $T : V^k \longrightarrow F$ is said to be a **symmetric** multilinear function if

$$T(\mathbf{v}_{j_1}, \mathbf{v}_{j_2}, \ldots, \mathbf{v}_{j_k}) = T(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) \tag{1.65}$$

for any permutation $j_1, j_2, \ldots, j_k$ of the indices $1, 2, \cdots, k$; and it is said to be an **alternating** multilinear function if

$$T(\mathbf{v}_{j_1}, \mathbf{v}_{j_2}, \ldots, \mathbf{v}_{j_k}) = \operatorname{sgn} \begin{pmatrix} j_1 & j_2 & \cdots & j_k \\ 1 & 2 & \cdots & k \end{pmatrix} T(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k) \tag{1.66}$$

for any permutation $j_1, j_2, \ldots j_k$ of the indices $1, 2, \cdots, k$, where the factor $\operatorname{sgn}\begin{pmatrix} j_1 & j_2 & \cdots & j_k \\ 1 & 2 & \cdots & k \end{pmatrix}$ is the sign of the permutation $j_1, j_2, \cdots, j_k$ of the indices $1, 2, \cdots, k$, as defined on the preceding page. It is convenient to extend the definition by setting $\operatorname{sgn}\begin{pmatrix} j_1 & j_2 & \cdots & j_k \\ 1 & 2 & \cdots & k \end{pmatrix} = 0$ if the indices $j_1, j_2, \cdots, j_k$ are not a permutation of the indices $1, 2, \cdots, k$, so if for example there are some repeated indices among $j_1, j_2, \cdots, j_k$. Both symmetric and alternating multilinear functions are of considerable interest; but actually for the purposes of the discussion here the alternating multilinear functions are of the greatest interest. The following special case is particularly important; other cases of great interest will be considered in the later discussion of differential forms.

Theorem 1.20. *For any field F and any n there is an alternating multilinear function*

$$\mathbf{\Phi} : \underbrace{F^n \times \cdots \times F^n}_{n} \longrightarrow F \tag{1.67}$$

from n copies of the vector space F^n to F, and it is uniquely determined up to a constant factor.

Proof: If $\mathbf{\Phi}$ is an alternating multilinear function as in (1.67) and the vectors $\mathbf{v}_i \in F^n$ are written in terms of the basis vectors $\boldsymbol{\delta}_j$ as $\mathbf{v}_i = \sum_{j=1}^{n} v_{ji}\boldsymbol{\delta}_j$ then since

$\mathbf{\Phi}$ is multilinear it follows that

$$\begin{aligned}\mathbf{\Phi}(\mathbf{v}_1, \cdots, \mathbf{v}_n) &= \mathbf{\Phi}\left(\sum_{j_1=1}^{n} v_{j_1 1}\boldsymbol{\delta}_{j_1}, \ldots, \sum_{j_n=1}^{n} v_{j_n n}\boldsymbol{\delta}_{j_n}\right) \\ &= \sum_{j_1,\ldots,j_n=1}^{n} v_{j_1 1}\cdots v_{j_n n}\mathbf{\Phi}(\boldsymbol{\delta}_{j_1}, \ldots, \boldsymbol{\delta}_{j_n})\end{aligned}$$

where $\mathbf{\Phi}(\boldsymbol{\delta}_{j_1}, \ldots, \boldsymbol{\delta}_{j_n}) \in F$; and since $\mathbf{\Phi}$ also is assumed to be alternating

$$\mathbf{\Phi}(\boldsymbol{\delta}_{j_1}, \ldots, \boldsymbol{\delta}_{j_n}) = \operatorname{sgn}\begin{pmatrix} j_1 & \ldots & j_n \\ 1 & \ldots & n \end{pmatrix}\mathbf{\Phi}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n).$$

Thus if there is an alternating multilinear function (1.67) then it must have the form

$$\mathbf{\Phi}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = C\sum_{j_1,\ldots,j_n=1}^{n} \operatorname{sgn}\begin{pmatrix} j_1 & \ldots & j_n \\ 1 & \ldots & n \end{pmatrix} v_{j_1 1}\cdots v_{j_n n} \tag{1.68}$$

where $C = \mathbf{\Phi}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)$. Conversely the mapping

$$\mathbf{\Phi} : F^n \times \cdots \times F^n \longrightarrow F$$

defined by (1.68) is an alternating multilinear function, since each term $v_{j_1 1}, \ldots, v_{j_n n}$ has exactly one factor from each column and

$$\begin{aligned}\mathbf{\Phi}(\mathbf{v}_{k_1}, \ldots, \mathbf{v}_{k_n}) &= C\sum_{j_1,\ldots,j_n=1}^{n} \operatorname{sgn}\begin{pmatrix} j_1 & \ldots & j_n \\ 1 & \ldots & n \end{pmatrix} v_{j_1 k_1}\cdots v_{j_n k_n} \\ &= C\sum_{j_1,\ldots,j_n=1}^{n} \operatorname{sgn}\begin{pmatrix} k_1 & \ldots & k_n \\ 1 & \ldots & n \end{pmatrix}\begin{pmatrix} j_1 & \ldots & j_n \\ k_1 & \ldots & k_n \end{pmatrix} v_{j_1 k_1}\cdots v_{j_n k_n} \\ &= \operatorname{sgn}\begin{pmatrix} k_1 & \ldots & k_n \\ 1 & \ldots & n \end{pmatrix}\mathbf{\Phi}(\mathbf{v}_1, \ldots, \mathbf{v}_n).\end{aligned}$$

That suffices for the proof.

The preceding theorem shows that any alternating multilinear function has the form (1.68) for some constant C; for any $n \times n$ matrix A the value of the multilinear function (1.68) where $\mathbf{v}_i = \mathbf{a}_i$ are the columns of the matrix A and

$C = 1$ is called the **determinant** of the $n \times n$ matrix $A = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n)$ and is denoted by $\det A$, so that

$$\det A = \sum_{j_1,\ldots,j_n=1}^{n} \operatorname{sgn} \begin{pmatrix} j_1 & \cdots & j_n \\ 1 & \cdots & n \end{pmatrix} a_{j_1 1} \cdots a_{j_n n}; \tag{1.69}$$

the summation is formally extended over all choices of the integers $j_1, \ldots, j_n$, with the understanding that $\operatorname{sgn}\begin{pmatrix} j_1 & \cdots & j_n \\ 1 & \cdots & n \end{pmatrix} = 0$ unless the indices $j_1, \ldots, j_n$ are a permutation of the indices $1, \ldots, n$. The basic properties of the determinant follow quite directly from its definition.

Theorem 1.21. *The determinant mapping over a field F satisfies*
(i) $\det I = 1$ *for the identity matrix I.*
(ii) $\det {}^tA = \det A$.
(iii) $\det(AB) = \det A \cdot \det B$ *for any $n \times n$ matrices A and B, and*
(iv) $\det A \neq 0$ *if and only if A is a nonsingular matrix.*

Proof: (i) By definition the determinant mapping is normalized so that $\det I = \boldsymbol{\Phi}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n) = 1$.
(ii) The products $a_{1j_1} \cdots a_{nj_n}$ and $a_{j_1 1} \cdots a_{j_n n}$ run over the same set of values as the indices $j_1, \ldots, j_n$ run through all permutations of the integers $1, \ldots, n$ so the basic formula (1.69) has the same value when the terms a_{ij} are replaced by a_{ji}.
(iii) If $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ are $n \times n$ matrices, the j-th column $\mathbf{v}_j$ of their product $AB = \{\sum_{k=1}^{n} a_{ik} b_{kj}\}$ is the linear combination $\mathbf{v}_j = \sum_{k=1}^{n} \mathbf{a}_k b_{kj}$ of the column vectors $\mathbf{a}_k$ of the matrix A; so for any linear function $\phi : F^n \longrightarrow F$ it follows that

$$\phi(\mathbf{v}_j) = \phi\left(\sum_{k=1}^{n} \mathbf{a}_k b_{kj}\right) = \sum_{k=1}^{n} b_{kj} \phi(\mathbf{a}_k).$$

The determinant is defined by (1.69), and since the function $\boldsymbol{\Phi}$ is multilinear and alternating

$$\begin{aligned} \det AB = \boldsymbol{\Phi}(\mathbf{v}_1, \ldots, \mathbf{v}_n) &= \sum_{k_1,\ldots,k_n=1}^{n} b_{k_1 1} \ldots b_{k_n n} \boldsymbol{\Phi}\left(\mathbf{a}_{k_1}, \ldots, \mathbf{a}_{k_n}\right) \\ &= \sum_{k_1,\ldots,k_n=1}^{n} b_{k_1 1} \cdots b_{k_n n} \operatorname{sgn} \begin{pmatrix} k_1 & \cdots & k_n \\ 1 & \cdots & n \end{pmatrix} \boldsymbol{\Phi}(\mathbf{a}_1, \ldots, \mathbf{a}_n) \\ &= \det B \cdot \det A. \end{aligned}$$

(iv) If an $n \times n$ matrix A is nonsingular then it has a well-defined inverse matrix A^{-1} for which $AA^{-1} = I$, and it then follows from parts (i) and (iii)

of this theorem that $\det A \det A^{-1} = \det(A \cdot A^{-1}) = \det I = 1$, and consequently $\det A \neq 0$. Conversely if A is an $n \times n$ matrix for which $\det A \neq 0$ then by the Equivalence Theorem, Theorem 1.15, there are nonsingular matrices S, T such that $SAT = B$ where $B = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ for some number k with $0 \leq k \leq n$; and since the matrices S and T are nonsingular $\det S \neq 0$ and $\det T \neq 0$ by what has just been proved, while $\det A \neq 0$ by assumption so $\det B \neq 0$ in view of (iii). That can be the case only if $k = n$, since if one of the columns of B is zero then $\det B = 0$ as a consequence of the defining formula (1.69); and if $k = n$ then $B = I$ is a nonsingular matrix, and therefore A is a product of nonsingular matrices so it is also nonsingular. That suffices for the proof.

Some further properties of determinants are useful in calculating their actual values.

Corollary 1.22. *The determinant mapping over a field F satisfies the following.*
(i) $\det A$ *changes signs if any two columns or rows are permuted.*
(ii) $\det A = 0$ *if the columns or rows are linearly dependent.*
(iii) *If a matrix A' arises from a matrix A by multiplying any row or column by $a \in F$ then* $\det A' = a \det A$.
(iv) *If a matrix A' arises from a matrix A by adding to one column a constant multiple of another column, or the corresponding operation on rows, then* $\det A' = \det A$.

Proof: (i) The determinant is defined as $\det A = \Phi(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ in terms of the columns of the matrix A; and since the function Φ is alternating and the sign of a transposition is -1 it follows that $\det A$ changes sign whenever two columns of A are interchanged. Since $\det {}^tA = \det A$ by Theorem 1.21 (ii) it follows from what has just been demonstrated that $\det A$ also changes sign whenever any two rows are interchanged.
(ii) If the columns of A are linearly dependent then a multiple of one of the columns is in the span of the remaining columns; if it is assumed for simplicity of notation that $\mathbf{a}_1 = \sum_{j=2}^{n} c_j \mathbf{a}_j$ then by linearity

$$\mathbf{\Phi}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \sum_{j=2}^{n} c_j \mathbf{\Phi}(\mathbf{a}_j, \mathbf{a}_2, \ldots, \mathbf{a}_n).$$

By (i) the function F changes sign when columns 1 and j are interchanged; but if the two columns are equal it also must be the case that the function does not change signs, and that can be the case only when the function F is zero.

(iii) Since the function $\mathbf{\Phi}$ in (1.67) is multilinear it is multiplied by a if any column is multiplied by a; and since $\det {}^tA = \det A$ by Theorem 1.21 (ii) that also is the case when any row is multiplied by a.
(iv) Since the function $\mathbf{\Phi}$ is multilinear

$$\mathbf{\Phi}(\mathbf{a}_1 + c\mathbf{a}_2, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathbf{\Phi}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) + c\mathbf{\Phi}(\mathbf{a}_2, \mathbf{a}_2, \ldots, \mathbf{a}_n)$$

and $\mathbf{\Phi}(\mathbf{a}_2, \mathbf{a}_2, \ldots, \mathbf{a}_n) = 0$ by (ii).

One rather straightforward way to calculate the explicit value of a determinant is to reduce the calculation to that of the determinants of smaller matrices. The basic formula (1.69) can be written

$$\begin{aligned}
\det A &= \sum_{j_1=1}^{n} \sum_{j_2,\ldots,j_n=1}^{n} \operatorname{sgn} \begin{pmatrix} j_1 & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix} a_{j_1 1} a_{j_2 2} \cdots a_{j_n n} \\
&= a_{11} \sum_{j_2,\ldots,j_n=1}^{n} \operatorname{sgn} \begin{pmatrix} 1 & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix} a_{j_2 2} \cdots a_{j_n n} \\
&\quad + a_{21} \sum_{j_2,\ldots,j_n=1}^{n} \operatorname{sgn} \begin{pmatrix} 2 & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix} a_{j_2 2} \cdots a_{j_n n} \\
&\quad + \cdots + \\
&\quad + a_{n1} \sum_{j_2,\ldots,j_n=1}^{n} \operatorname{sgn} \begin{pmatrix} n & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix} a_{j_2 2} \cdots a_{j_n n}
\end{aligned}$$

so

$$\det A = a_{11} \det A_{11} - a_{21} \det A_{21} + \cdots + (-1)^{n+1} a_{n1} \det A_{n1} \tag{1.70}$$

where A_{ij} is the matrix arising from the matrix A by deleting both row i and column j. Thus for instance

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a \det d - c \det b = ad - cb \tag{1.71}$$

and

$$\det \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 7 & 0 & 8 \end{pmatrix} = 1 \det \begin{pmatrix} 5 & 4 \\ 0 & 8 \end{pmatrix} - 6 \det \begin{pmatrix} 2 & 3 \\ 0 & 8 \end{pmatrix} + 7 \det \begin{pmatrix} 2 & 3 \\ 5 & 4 \end{pmatrix} = -105.$$

The columns of the matrix A can be permuted, with a change of sign, and the same formula can be applied to the rows, which provides a good deal of flexibility in this approach to calculating determinants. Note that if the entries a_{j1} in (1.70) are replaced by a_{ji} for $i \neq 1$ the result would be the determinant of the matrix obtained from A by replacing column 1 by column i, so the result would be 0, the determinant of a matrix with two identical columns. Consequently if $A^\dagger$ is the matrix with the entries $A^\dagger_{ij} = (-1)^{i+j} \det A_{ji}$, a matrix called the **adjugate** of the matrix A, it follows from (1.70) that

$$A \cdot A^\dagger = A^\dagger \cdot A = (\det A) \cdot I. \tag{1.72}$$

For example

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^\dagger = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \text{where} \quad \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc. \tag{1.73}$$

This can be applied to solve a system of linear equations $A\mathbf{x} = \mathbf{a}$ for a square matrix A explicitly as $\mathbf{x} = \frac{1}{\det A} A^\dagger \mathbf{a}$, an application customarily called **Cramer's rule**.

Matrices and vector spaces arise very commonly in other algebraic structures. A permutation of a set of variables can be described by a matrix operation; for example

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ x_1 \end{pmatrix}. \tag{1.74}$$

The matrices describing permutations in this way, called **permutation matrices**, are those matrices such that each row and each column contain a single entry of 1 but all other entries are 0. Multiplication in a group has the effect of permuting the elements of the group, so any finite group can be viewed as a subgroup of the set of all permutations of the elements of the group, and in that way any group is isomorphic to a group of matrices. This is actually an immensely useful tool in the investigation of abstract groups.

The algebraic discussion here will be concluded by examining yet another general algebraic structure. An **algebra** over field F is defined to be a vector space $\mathcal{A}$ over the field F with an additional operation that associates to any two vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{A}$ their product $\mathbf{v}_1 \cdot \mathbf{v}_2 \in \mathcal{A}$ such that

(i) multiplication of vectors is associative and has an identity element $I \in \mathcal{A}$;

(ii) the distributive laws $\mathbf{v} \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{v} \cdot \mathbf{v}_1 + \mathbf{v} \cdot \mathbf{v}_2$ and $(\mathbf{v}_1 + \mathbf{v}_2) \cdot \mathbf{v} = \mathbf{v}_1 \cdot \mathbf{v} + \mathbf{v}_2 \cdot \mathbf{v}$ hold for any vectors $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in \mathcal{A}$;

(iii) scalar multiplication and the multiplication of vectors are related by the laws $c(\mathbf{v}_1 \cdot \mathbf{v}_2) = (c\mathbf{v}_1) \cdot \mathbf{v}_2 = \mathbf{v}_1 \cdot (c\mathbf{v}_2)$ for all $c \in F$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{A}$.

Note particularly that it is not assumed that the multiplication of vectors is commutative; if the multiplication of vectors is commutative the algebra is called a **commutative algebra**. The **dimension** of an algebra $\mathcal{A}$ is defined to be its dimension as a vector space. Just as in the case of rings, the zero element $\mathbf{0}$ plays a special role in multiplication in an algebra, so that $\mathbf{0} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{0} = \mathbf{0}$ for any vector $\mathbf{v} \in \mathcal{A}$, with the same proof as in the preceding examination of rings. The matrix space $F^{n\times n}$ under multiplication of matrices is an example of an algebra, which of course is not necessarily commutative. The field F itself is a commutative one-dimensional algebra over F, indeed can readily be seen to be the unique one-dimensional algebra over F. The set $F[X]$ of all polynomials with coefficients in F is another commutative algebra, although not a finite dimensional algebra; and other infinite dimensional algebras will arise in the subsequent discussion.

It is not difficult to describe all two-dimensional commutative real algebras; indeed that is a good exercise in the study of algebras. If $\mathcal{A}$ is a commutative real algebra and $\dim \mathcal{A} = 2$, choose a basis $\mathbf{v}_1, \mathbf{v}_2$ for the vector space $\mathcal{A}$ so that $\mathbf{v}_1$ is the identity for multiplication in the algebra; the multiplication table in terms of this basis then must be of the form

$$\mathcal{A}_{x_1,x_2}: \quad \mathbf{v}_1 \cdot \mathbf{v}_1 = \mathbf{v}_1, \ \ \mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_2 \cdot \mathbf{v}_1 = \mathbf{v}_2, \ \ \mathbf{v}_2 \cdot \mathbf{v}_2 = x_1\mathbf{v}_1 + x_2\mathbf{v}_2$$

for some $x_1, x_2 \in \mathbb{R}$. This multiplication table can be simplified by replacing $\mathbf{v}_2$ by $\mathbf{v}_2' = \mathbf{v}_2 - t\mathbf{v}_1$ for a suitable real number $t \in \mathbb{R}$; for any choice of $t \in \mathbb{R}$ the vectors $\mathbf{v}_1, \mathbf{v}_2'$ are another basis for V, and by the distributive law

$$\begin{aligned}\mathbf{v}_2' \cdot \mathbf{v}_2' &= (\mathbf{v}_2 - t\mathbf{v}_1) \cdot (\mathbf{v}_2 - t\mathbf{v}_1) = \mathbf{v}_2 \cdot \mathbf{v}_2 - 2t\mathbf{v}_1 \cdot \mathbf{v}_2 + t^2\mathbf{v}_1 \cdot \mathbf{v}_1 \\ &= (x_1\mathbf{v}_1 + x_2\mathbf{v}_2) - 2t\mathbf{v}_2 + t^2\mathbf{v}_1 = (x_1 + t^2)\mathbf{v}_1 + (x_2 - 2t)\mathbf{v}_2 \\ &= (x_1 + t^2)\mathbf{v}_1 + (x_2 - 2t)(\mathbf{v}_2' + t\mathbf{v}_1) = (x_1 + tx_2 - t^2)\mathbf{v}_1 + (x_2 - 2t)\mathbf{v}_2'\end{aligned}$$

so if $2t = x_2$ it follows that

$$\mathbf{v}_2' \cdot \mathbf{v}_2' = y\mathbf{v}_1 \quad \text{for some } y \in \mathbb{R}.$$

For a further simplification replace the vector $\mathbf{v}_2'$ by $\mathbf{v}_2'' = s\mathbf{v}_2'$ for another real number $s \in \mathbb{R}$, so that

$$\mathbf{v}_2'' \cdot \mathbf{v}_2'' = s^2 y\mathbf{v}_1.$$

The real number y appearing in this formula can be either 0 or positive or negative. Recalling that any positive real number can be written as the square of a positive real number, which was demonstrated at the end of Section 1.2, it is possible to choose s so that $s^2y = 1$ if $y > 0$ and $s^2y = -1$ if $y < 0$. Therefore after relabeling the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ the multiplication table for the algebra takes the simpler form

$$\mathcal{A}_\epsilon : \quad \mathbf{v}_1 \cdot \mathbf{v}_1 = \mathbf{v}_1, \ \ \mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_2 \cdot \mathbf{v}_1 = \mathbf{v}_2, \ \ \mathbf{v}_2 \cdot \mathbf{v}_2 = \epsilon \mathbf{v}_1 \tag{1.75}$$

where ϵ is either 0 or 1 or -1. For the case $\epsilon = 0$ the 2×2 real matrices

$$\mathbf{v}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{v}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{1.76}$$

obviously satisfy (1.75); the vector space spanned by these two matrices is

$$\mathcal{A}_0 = \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \,\middle|\, a, b \in \mathbb{R} \right\}, \tag{1.77}$$

and a straightforward calculation shows that this set of matrices is closed under multiplication and form a commutative real algebra. For the case $\epsilon = 1$ the 2×2 real matrices

$$\mathbf{v}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{v}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{1.78}$$

clearly satisfy (1.75); the vector space spanned by these two matrices is

$$\mathcal{A}_1 = \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} \,\middle|\, a, b \in \mathbb{R} \right\}, \tag{1.79}$$

and a straightforward calculation shows that this set of matrices is closed under multiplication and form a commutative real algebra. For the case $\epsilon = -1$ the 2×2 real matrices

$$\mathbf{v}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{v}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{1.80}$$

obviously satisfy (1.75); the vector space spanned by these two matrices is

$$\mathcal{A}_{-1} = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \,\middle|\, a, b \in \mathbb{R} \right\}, \tag{1.81}$$

and a straightforward calculation shows that this set of matrices also is closed under multiplication and form a commutative real algebra. It is also easy to see that these three real commutative algebras are distinct algebras, in the sense that no two are isomorphic. Indeed there are elements in $\mathcal{A}_0$ the squares of which are zero, as for instance $\mathbf{v}_2 \cdot \mathbf{v}_2 = \mathbf{0}$; in $\mathcal{A}_1$ however $(x_1\mathbf{v}_1 + x_2\mathbf{v}_2) \cdot (x_1\mathbf{v}_1 + x_2\mathbf{v}_2) = (x_1^2 + x_2^2)\mathbf{v}_1 + 2x_1x_2\mathbf{v}_2$ for any real x_1, x_2, so nonzero elements always have nonzero squares, but $(\mathbf{v}_1 + \mathbf{v}_2) \cdot (\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$ so there are nonzero elements having products $\mathbf{0}$; finally if $(x_1\mathbf{v}_1 + x_2\mathbf{v}_2) \in \mathcal{A}_{-1}$ then $(x_1\mathbf{v}_1 + x_2\mathbf{v}_2) \cdot (x_1\mathbf{v}_1 - x_2\mathbf{v}_2) = (x_1^2 + x_2^2)\mathbf{v}_1$ for any real x_1, x_2, so any nonzero element $(x_1\mathbf{v}_1 + x_2\mathbf{v}_2) \in \mathcal{A}_{-1}$ has the multiplicative inverse $(x_1\mathbf{v}_1 + x_2\mathbf{v}_2)^{-1} = (x_1^2 + x_2^2)^{-1}(x_1\mathbf{v}_1 - x_2\mathbf{v}_2)$, showing that $\mathcal{A}_{-1}$ actually is a field.

The field $\mathcal{A}_{-1}$ is called the field of **complex numbers** and is denoted by $\mathbb{C}$. The basis vector $\mathbf{v}_1$ is customarily identified with the real identity element 1, and a product $x\mathbf{v}_1 \in \mathbb{C}$ is identified with the real number x viewed as being imbedded in $\mathbb{C}$; since $(x\mathbf{v}_1) \cdot (y\mathbf{v}_1) = (xy)\mathbf{v}_1$ the imbedding $\mathbb{R} \subset \mathbb{C}$ is compatible with the algebraic operations in the fields $\mathbb{R}$ and $\mathbb{C}$. The basis vector $\mathbf{v}_2$ is usually denoted by i, and the products $y\mathbf{v}_2 = yi$ for $y \in \mathbb{R}$ are known for historical reasons as **imaginary numbers**. The elements of the field $\mathbb{C}$, the complex numbers, thus are written $z = x + iy$, where $x = \Re(z) \in \mathbb{R}$ is called the **real part** of the complex number z and $y = \Im(z) \in \mathbb{R}$ is called the **imaginary part** of the complex number z. In (1.80) there is a choice whether to use $\mathbf{v}_2$ or $-\mathbf{v}_2$ as the second basis vector; whichever is chosen clearly leads to the same algebra (1.81) and the same algebraic operations. Thus it is possible to associate to any complex number $z = x + iy \in \mathbb{C}$ another complex number $\overline{z} = x - iy$ called the **complex conjugate** of z; and the mapping $z \longrightarrow \overline{z}$ is a field isomorphism since it is readily verified that $\overline{(z_1 + z_2)} = \overline{z_1} + \overline{z_2}$ and $\overline{(z_1 \cdot z_2)} = \overline{z_1} \cdot \overline{z_2}$. In terms of the matrices (1.81) complex conjugation corresponds to taking the transpose of the matrix. It is worth noting that obviously there is no nontrivial isomorphism of the field $\mathbb{R}$ with itself, exhibiting an interesting difference between the two fields $\mathbb{R}$ and $\mathbb{C}$. It is an interesting exercise to go through the preceding classification of two-dimensional algebras for the case of algebras over the complex numbers. The argument goes through as before; but any complex number is the square of another complex number, so there are only the two distinct algebras (1.75) over the complex numbers corresponding to the values $\epsilon = 0, 1$ and neither is a field. The world is in some ways not as interesting as it might be; alas there are no other finite dimensional real algebras that are fields.[8]

[8] See the *Princeton Companion to Mathematics*.

Problems, Group I

1. If W_1, W_2 are linear subspaces of a vector space V over a field F, which of the following subsets of V are also linear subspaces and why:
(i) $W_1 \cap W_2$ (ii) $W_1 \cup W_2$ (iii) $W_1 \sim W_2$ (iv) $W_1 + W_2$.

2. Find a basis for the subspace of $\mathbb{R}^4$ spanned by the vectors $(1, 2, -1, 0)$, $(4, 8, -4, -3)$, $(0, 1, 3, 4)$, and $(2, 5, 1, 4)$.

3. Find an example of a vector space V and subspaces $M, N_1, N_2 \subset V$ such that $M \oplus N_1 = M \oplus N_2 = V$ but $N_1 \neq N_2$.

4. If A is an $m \times n$ matrix over a field F show that the vector space V consisting of those vectors $\mathbf{x} \in F^n$ such that $A\mathbf{x} = \mathbf{0}$ has dimension at least $n - m$. Find an example in which $n > m$ and the dimension of V is strictly greater than $n - m$.

5. If A is an $m \times n$ matrix over a field F with $n < m$ show that there do not exist any $n \times m$ matrices B over that field for which $AB = I$.

6. Find 2×2 matrices A, B for which $AB = 0$ but $BA \neq 0$.

7. Show that $\operatorname{rank} AB \leq \min(\operatorname{rank} A,\ \operatorname{rank} B)$ for any two matrices A, B for which the product is defined; and find an example for which this is an equality and another example for which this is a strict inequality.

8. Consider the set of $n \times n$ matrices

$$I_{i_1, i_2, \ldots, i_n} = (\boldsymbol{\delta}_{i_1}, \boldsymbol{\delta}_{i_2}, \ldots, \boldsymbol{\delta}_{i_n})$$

where $\boldsymbol{\delta}_i \in \mathbb{R}^n$ is the Kronecker vector having the entry 1 in row i and entries 0 in the other rows and $i_1, i_2, \ldots, i_n$ are any n integers in the range $[1, n]$, not necessarily distinct integers.

   i) What is the rank of the matrix $I_{i_1, i_2, \ldots, i_n}$?
   ii) What is $\det I_{i_1, i_2, \ldots, i_n}$?
   iii) What is the vector $I_{i_1, i_2, \ldots, i_n}\mathbf{x}$ for a column vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_n \end{pmatrix}$ in $\mathbb{R}^n$?

9. Let X be the real matrix

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ x_2 & x_3 & x_1 \\ x_3 & x_1 & x_2 \end{pmatrix}.$$

i) For what values x_i is rank$X = 2$?
ii) For what values x_i is rank$X = 1$?

10. A reasonably effective technique for calculating the rank and determinant of an $m \times n$ matrix A over a field F and examining explicitly the system of linear equations $A\mathbf{x} = \mathbf{y}$ is through the **elementary row operations** on A:

i) add c times row i to row j for some scalar $c \in F$;
ii) interchange rows i and j;
iii) multiply row i by a nonzero scalar $c \in F$.

Show that these operations do not change the rank of a matrix, and that each can be realized by multiplying the matrix A on the left by a suitable $m \times m$ matrix (an **elementary matrix**). These operations can be applied to simplify the form of a matrix A by reducing it to a matrix A' that is as close to an identity matrix as possible, so that it is easy to calculate its rank and determinant. When the same product E of elementary matrices is applied to both sides the equation $A\mathbf{x} = \mathbf{y}$ becomes $A'\mathbf{x} = EA\mathbf{x} = E\mathbf{y}$, which has the same solution $\mathbf{x}$; so finding the solution $\mathbf{x}$ and determining the conditions on $\mathbf{y}$ under which there exists a solution $\mathbf{x}$ (or equivalently finding the image of the mapping defined by the matrix A) are simplified as well. To ensure that the same elementary operations are applied to the matrix A as to the vector $\mathbf{y}$ it is customary to apply the operations to the extended matrix $(A\ \mathbf{y})$ where the vector $\mathbf{y}$ is added as an additional last column of the matrix A. Apply this procedure to show that A can be reduced to A' in the following example:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & y_1 \\ 1 & 0 & 1 & 0 & 1 & y_2 \\ 1 & 2 & 2 & 2 & 1 & y_3 \end{pmatrix} \quad A' = \begin{pmatrix} 1 & 0 & 0 & -2 & -3 & y'_1 \\ 0 & 1 & 0 & 0 & -2 & y'_2 \\ 0 & 0 & 1 & 2 & 4 & y'_3 \end{pmatrix}.$$

Determine the vector $\mathbf{y}'$ explicitly in terms of the vector $\mathbf{y}$. Determine the rank of the matrix A. Show that the equation $A\mathbf{x} = \mathbf{y}$ can be solved for $\mathbf{x}$ whenever $\mathbf{y}$ is specified arbitrarily; and in particular find a solution $\mathbf{x}$ when $\mathbf{y} = \{1, 2, 3\}$. Is your solution unique?

Problems, Group II

11. If V is an n-dimensional real vector space show that the kernel of any nonzero linear transformation $T : V \longrightarrow \mathbb{R}$ is a linear subspace of dimension $n - 1$ and conversely that any such linear subspace is the kernel of some linear transformation $T : V \longrightarrow \mathbb{R}$.

12. If A is an $m \times n$ and B is an $n \times m$ matrix show that $I - AB$ is nonsingular if and only if $I - BA$ is nonsingular.

13. Show that an $m \times n$ matrix A over a field F has rank 1 if and only if $A = \mathbf{b}\,{}^t\mathbf{c}$ for some nonzero (column) vectors $\mathbf{b} \in F^m$ and $\mathbf{c} \in F^n$.

14. If A, B, C, D are square real matrices where $\det A \neq 0$ and $AC = CA$ show that $\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(AD - CB)$. Find an example to show that this is not necessarily the case if the hypotheses are not assumed.

15. An $n \times n$ matrix A is **symmetric** if $A = {}^tA$ and is **skew symmetric** if $A = -{}^tA$.

    i) Show that the set of symmetric matrices is a linear subspace of the vector space of all $n \times n$ matrices, as is the set of skew-symmetric matrices, and find the dimensions of these linear subspaces.
    ii) Show that the vector space of $n \times n$ matrices is the direct sum of the subspace of symmetric and the subspace of skew symmetric matrices, provided that $2 \neq 0$ in the field F. What happens for the field $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ of two elements in which $2 = 0$?

16. The **trace** $\mathrm{tr}(A)$ of an $n \times n$ matrix is defined to be the sum of its diagonal terms.

    i) Show that $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$, that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and that $\mathrm{tr}(ABA^{-1}) = \mathrm{tr}(B)$ if the matrix A is invertible.
    ii) Is $\mathrm{tr}(AB) = \mathrm{tr}(A)\mathrm{tr}(B)$? Why?
    iii) Show that the equation $AB - BA = I$ has no solutions in $n \times n$ matrices A and B over a field F provided that $n \neq 0$ in F.

17. If $T : F^n \longrightarrow F^n$ is a linear transformation such that the kernel of T is equal to the image of T show that n must be an even number. Find an example of such a mapping.

18. If $T : V \longrightarrow V$ is a linear transformation on a finite dimensional vector space show that there is some number n for which $T^n(V) \cap \ker T^n = \mathbf{0}$.

19. i) For any elements $a_1, \ldots, a_n$ in a field F show that

$$\det \begin{pmatrix} 1 & a_1 & \cdots & a_1^{n-1} \\ 1 & a_2 & \cdots & a_2^{n-1} \\ \cdots & \cdots & & \cdots \\ 1 & a_n & \cdots & a_n^{n-1} \end{pmatrix} = \prod_{1 \leq i < j \leq n} (a_i - a_j).$$

    This determinant is called the **Vandermonde** determinant.

ii) Using the Vandermonde determinant show that for any n pairs

$$(a_1, b_1), \ldots, (a_n, b_n)$$

of elements in the field F, where a_i are distinct elements, there is a polynomial $p(x)$ of degree at most $n-1$ such that $p(a_i) = b_i$ for $1 \leq i \leq n$.

20. A diagram of vector spaces V_i and linear transformations T_i of the form

$$V_1 \xrightarrow{T_1} V_2 \xrightarrow{T_2} V_3 \xrightarrow{T_3} V_4 \xrightarrow{T_4} \ldots$$

is said to be an **exact sequence** of vector spaces if in any subsegment of the form $V_{i-1} \xrightarrow{T_{i-1}} V_i \xrightarrow{T_i} V_{i+1}$ the image of the linear transformation T_{i-1} is precisely the kernel of the linear transformation T_i. For example, the assertion that the sequence $\mathbf{0} \xrightarrow{T_0} V_1 \xrightarrow{T_1} V_2$ is exact is equivalent to the assertion that the linear transformation T_1 is injective while on the other hand the assertion that the sequence $V_1 \xrightarrow{T_1} V_2 \xrightarrow{T_2} \mathbf{0}$ is exact is equivalent to the assertion that the linear transformation T_1 is surjective.

i) What is the interpretation of the exactness of the sequence

$$\mathbf{0} \xrightarrow{T_0} V_1 \xrightarrow{T_1} V_2 \xrightarrow{T_2} \mathbf{0}?$$

ii) What is the interpretation of the exactness of the sequence

$$\mathbf{0} \xrightarrow{T_0} V_1 \xrightarrow{T_1} V_2 \xrightarrow{T_2} V_3 \xrightarrow{T_3} \mathbf{0},$$

traditionally called a **short exact sequence**?

iii) Show that $\sum_{i=1}^{n} (-1)^i \dim V_i = 0$ for any exact sequence of the form

$$\mathbf{0} \xrightarrow{T_0} V_1 \xrightarrow{T_1} V_2 \xrightarrow{T_2} \cdots \xrightarrow{T_{n-1}} V_n \xrightarrow{T_n} \mathbf{0},$$

a sequence beginning and ending with the trivial vector space $\mathbf{0}$.