The Zakon Series on Mathematical Analysis

Basic Concepts of Mathematics

Mathematical Analysis I

Mathematical Analysis II

The Zakon Series on Mathematical Analysis

Basic Concepts of

Mathematics

Elias Zakon
University of Windsor

The Trillia Group

West Lafayette, IN

Basic Concepts of Mathematics

ISBN 978-1-931705-00-3
Published by The Trillia Group, West Lafayette, Indiana, USA
First published: May 26, 2001. This version released: February 3, 2014.
Technical Typist: Judy Mitchell. Copy Editor: John Spiegelman. Logo: Miriam Bogdanic.

This book was prepared by Bradley J. Lucier and Tamara Zakon from a manuscript prepared by Elias Zakon. We intend to correct and update this work as needed. If you notice any mistakes in this work, please send e-mail to lucier@math.purdue.edu and they will be corrected in a later version.

Half the proceeds from the sale of this book go to the *Elias Zakon Memorial Scholarship* fund at the University of Windsor, Canada, funding scholarships for undergraduate students majoring in Mathematics and Statistics.

Contents*

* "Starred" sections may be omitted by beginners.

Preface

This text helps the student complete the transition from purely manipulative to rigorous mathematics. It spells out in all detail what is often treated too briefly or vaguely because of lack of time or space. It can be used either for supplementary reading or as a half-year course. It is self-contained, though usually the student will have had elementary calculus before starting it. Without the "starred" sections and problems, it can be (and *was*) taught even to freshmen. The three chapters are fairly independent and, with small adjustments, may be taught in arbitrary order. The chapter on n-space "imitates" the geometry of lines and planes in 3-space, and ensures a thorough review of the latter, for students who may not have had it. A wealth of problems, some simple, some challenging, follow almost every section.

Several years' class testing led the author to these conclusions:

(1) The earlier such a course is given, the more time is gained in the follow-up courses, be it algebra, analysis or geometry. The longer students are taught "vague analysis", the harder it becomes to get them used to rigorous proofs and formulations and the harder it is for them to get rid of the misconception that mathematics is just memorizing and manipulating some formulas.

(2) When teaching the course to freshmen, it is advisable to start with Sections 1–7 of Chapter 2, then pass to Chapter 3, leaving Chapter 1 and Sections 8–10 of Chapter 2 for the end. The students should be urged to *preread* the material to be taught next. (Freshmen must *learn* to read mathematics by *rereading* what initially seems "foggy" to them.) The teacher then may confine himself to a brief summary, and *devote most of his time to solving as many problems* (*similar to those assigned*) *as possible*. This is absolutely necessary.

(3) An early and constant use of logical quantifiers (even in the text) is extremely useful. Quantifiers are there to stay in mathematics.

(4) Motivations are necessary and good, provided they are brief and do not use terms that are not yet clear to students.

About the Author

Elias Zakon was born in Russia under the czar in 1908, and he was swept along in the turbulence of the great events of twentieth-century Europe.

Zakon studied mathematics and law in Germany and Poland, and later he joined his father's law practice in Poland. Fleeing the approach of the German Army in 1941, he took his family to Barnaul, Siberia, where, with the rest of the populace, they endured five years of hardship. The Leningrad Institute of Technology was also evacuated to Barnaul upon the siege of Leningrad, and there he met the mathematician I. P. Natanson; with Natanson's encouragement, Zakon again took up his studies and research in mathematics.

Zakon and his family spent the years from 1946 to 1949 in a refugee camp in Salzburg, Austria, where he taught himself Hebrew, one of the six or seven languages in which he became fluent. In 1949, he took his family to the newly created state of Israel and he taught at the Technion in Haifa until 1956. In Israel he published his first research papers in logic and analysis.

Throughout his life, Zakon maintained a love of music, art, politics, history, law, and especially chess; it was in Israel that he achieved the rank of chess master.

In 1956, Zakon moved to Canada. As a research fellow at the University of Toronto, he worked with Abraham Robinson. In 1957, he joined the mathematics faculty at the University of Windsor, where the first degrees in the newly established Honours program in Mathematics were awarded in 1960. While at Windsor, he continued publishing his research results in logic and analysis. In this post-McCarthy era, he often had as his house-guest the prolific and eccentric mathematician Paul Erdős, who was then banned from the United States for his political views. Erdős would speak at the University of Windsor, where mathematicians from the University of Michigan and other American universities would gather to hear him and to discuss mathematics.

While at Windsor, Zakon developed three volumes on mathematical analysis, which were bound and distributed to students. His goal was to introduce rigorous material as early as possible; later courses could then rely on this material. We are publishing here the latest complete version of the first of these volumes, which was used in a one-semester class required of all first-year Science students at Windsor.

Chapter 1

Some Set Theoretical Notions

§1. Introduction. Sets and Their Elements

The theory of sets, initiated by the German mathematician G. Cantor (1842–1918), constitutes the basis of almost all modern mathematics. The set concept itself cannot be defined in simpler terms. A set is often described as a collection ("aggregate", "class", "totality", "family") of objects of any specified kind. However, such descriptions are no definitions, as they merely replace the term "set" by other undefined terms. Thus the term "set" must be accepted as a *primitive notion*, without definition. Examples of sets are as follows: the set of all men; the set of all letters appearing on this page; the set of all straight lines in a given plane; the set of all positive integers; the set of all English songs; the set of all books in a library; the set consisting of the three numbers 1, 4, 17. Sets will usually be denoted by capital letters, A, B, C, $\ldots$, X, Y, Z.

The objects belonging to a set A are called its *elements* or *members*. We write $x \in A$ if x is an element of the set A, and $x \notin A$ if it is not.

Example.

If N is the set of all positive integers, then $1 \in N$, $3 \in N$, $+\sqrt{9} \in N$, but $\sqrt{7} \notin N$, $0 \notin N$, $-1 \notin N$, $\frac{1}{2} \notin N$.

It is also convenient to introduce the so-called *empty* ("*void*", "*vacuous*") set, denoted by $\emptyset$, i.e., a set that contains no elements at all. Instead of saying that there are no objects of some specific kind, we shall say that the set of these elements is empty; *however*, *this set itself*, *though empty*, *will be regarded as an existing thing*.

Once a set has been formed, it is regarded as a new entity, that is, a new object, different from any of its elements. This object may, in its turn, be an element of some other set. In fact, we can consider whole collections of sets (also called "families of sets", "classes of sets", etc.), i.e., sets whose elements are other sets. Thus, if $\mathcal{M}$ is a collection of certain sets A, B, C, $\ldots$, then these sets are elements of $\mathcal{M}$, i.e., we have $A \in \mathcal{M}$, $B \in \mathcal{M}$, $C \in \mathcal{M}$, $\ldots$;

but the single elements of A need not be members of $\mathcal{M}$, and the same applies to single elements of B, C, ... Briefly, *from $p \in A$ and $A \in \mathcal{M}$, it does not follow that $p \in \mathcal{M}$*. This may be illustrated by the following examples. Let a "nation" be defined as a certain set of individuals, and let the United Nations (U.N.) be regarded as a certain set of nations. Then single persons are elements of the nations, and the nations are members of U.N., but individuals are not members of U.N. Similarly, the Big Ten consists of ten universities, each university contains thousands of students, but no student is one of the Big Ten. Families of sets will usually be denoted by *script* letters: $\mathcal{M}$, $\mathcal{N}$, $\mathcal{P}$, etc.

If all elements of a set A are also elements of a set B, we say that A is a *subset* of B, and write $A \subseteq B$. In this instance, we also say that B is a *superset* of A, and we can write $B \supseteq A$. The set B is *equal* to A if $A \subseteq B$ and $B \subseteq A$, i.e., the two sets consist of exactly the same elements. If, however, $A \subseteq B$ but $B \neq A$ (i.e., B contains some elements not in A), then A is referred to as a *proper subset* of B; in this case we shall use the notation $A \subset B$. The empty set $\emptyset$ is considered a subset of *any* set; it is a proper subset of any nonempty set. The equality of two sets A and B is expressed by the formula $A = B$.[1] Instead of $A \subseteq B$ we shall also write $B \supseteq A$; similarly, we write $B \supset A$ instead of $A \subset B$. The relation "$\subseteq$" is called the *inclusion relation*.[2] Summing up, for any sets A, B, C, the following are true:

(a) $A \subseteq A$.

(b) *If* $A \subseteq B$ *and* $B \subseteq C$, *then* $A \subseteq C$.

(c) *If* $A \subseteq B$ *and* $B \subseteq A$, *then* $A = B$.

(d) $\emptyset \subseteq A$.

(e) *If* $A \subseteq \emptyset$, *then* $A = \emptyset$.

The properties (a), (b), (c) are usually referred to as the *reflexivity*, *transitivity*, and *anti-symmetry* of the inclusion relation, respectively; (c) is also called the *axiom of extensionality*.[3]

A set A may consist of a single element p; in this case we write $A = \{p\}$. This set must not be confused with the element p itself, especially if p itself is a set consisting of some elements a, b, c, ..., (recall that these elements are *not* regarded as elements of A; thus A consists of a *single* element p, whereas p may have *many* elements; A and p then are not identical). Similarly, the empty set

[1] The equality sign, here and in the sequel, is tantamount to *logical identity*. A formula like "$A = B$" means that the letters A and B denote *one and the same thing*.

[2] Some authors write $A \subset B$ for $A \subseteq B$. We prefer, however, to reserve the sign $\subset$ for *proper* inclusion.

[3] The statement that $A = B$ if A and B have the same elements shall be treated as an *axiom*, not a definition.

$\emptyset$ has no elements, while $\{\emptyset\}$ *has* an element, namely $\emptyset$. Thus $\emptyset \neq \{\emptyset\}$ and, in general, $p \neq \{p\}$.

If A contains the elements a, b, c, ..., we write

$$A = \{a, b, c, \dots\}$$

(the dots in this symbol imply that A may contain some other elements). If A consists of a small number of elements, it may be convenient to list them *all* in braces. In particular, if A consists of two elements a, b, we write $A = \{a, b\}$. Similarly for a set of three elements, $A = \{a, b, c\}$, etc. If confusion is unlikely, a finite set may be indicated by the use of dots and a terminal member, as with $\{1, 2, 3, \dots, 10\}$, or $\{2, 4, 6, \dots, 100\}$, or $\{1, 3, 5, \dots, 2n-1\}$.

It should be noted that the order in which the elements of a set follow each other does not affect the equality of sets as stated above. For instance, we have $\{a, b\} = \{b, a\}$ because the two sets consist of the same elements. Also, if some element is mentioned several times, it still counts as one element only. Thus we have $\{a, a\} = \{a\}$. In this respect, a set consisting of two elements a and b must be distinguished from the *ordered pair* (a, b); and, more generally, a set consisting of n elements, $\{x_1, x_2, \dots, x_n\}$, should not be confused with the *ordered n-tuple* $(x_1, \dots, x_n)$. Two ordered pairs (a, b) and (x, y) are considered equal iff[4] $a = x$ and $b = y$, whereas the sets $\{a, b\}$ and $\{x, y\}$ are also equal if $a = y$ and $b = x$. A similar distinction applies to ordered n-tuples.[5]

If $P(x)$ is some proposition or formula involving a variable x, we shall use the symbol

$$\{x \mid P(x)\}$$

to denote *the set of all objects x for which the formula $P(x)$ is true*. For instance, the set of all men can be denoted by $\{x \mid x \text{ is a man}\}$. Similarly, $\{x \mid x \text{ is a number}, x < 5\}$ stands for "the set of all numbers less than 5." We write $\{x \in A \mid P(x)\}$ for "the set of all elements of A for which $P(x)$ is true." The variable x in such symbols may be replaced by any other variable; $\{x \mid P(x)\}$ is the same as $\{y \mid P(y)\}$.

Thus the set of all positive integers less than 5 can be denoted either by $\{1, 2, 3, 4\}$, or by $\{x \mid x \text{ is an integer}, 0 < x < 5\}$. **Note:** The comma in such symbols stands for the word "and".

§2. Operations on Sets

We now proceed to define some operations on sets.

[4] "iff" means "if and only if".

[5] We shall not attempt at this stage to give a *definition* of an ordered pair or n-tuple, though this can be done (cf. Problem 6 after §2).

Definition 1.

For any two sets A and B, we define as follows:

(a) The *union*, or *join*, of A and B, denoted by $A \cup B$, is the set of all elements x such that $x \in A$ or $x \in B$ (i.e., the set of all elements of A and B taken together).[1]

(b) The *intersection*, or *meet*, of A and B, denoted by $A \cap B$, is the set of all elements x such that $x \in A$ and $x \in B$ simultaneously (it is the set of all *common* elements of A and B).

(c) The *difference* $A - B$ is the set of all elements that are in A but not in B (B may, but need not, be a subset of A).

In symbols,

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\},\ A \cap B = \{x \mid x \in A,\ x \in B\},\ \text{and}$$
$$A - B = \{x \mid x \in A,\ x \notin B\}.$$

The sets A and B are said to be *disjoint* iff $A \cap B = \emptyset$, i.e., iff they have no elements in common. The symbols $\cup$ and $\cap$ are called "cup" and "cap", respectively; sometimes the symbols $+$ and $\cdot$ are used instead. Note that, if A and B have some elements in common, these elements need not be mentioned *twice* when forming the union $A \cup B$. The difference $A - B$ is also called the *complement* of B relative to A (briefly, "*in* A").[2]

Examples.

(1) If $A = \{1, 2, 3, 4, 5\}$ and $B = \{2, 4, 6\}$, then

$$A \cup B = \{1, 2, 3, 4, 5, 6\}, \quad A \cap B = \{2, 4\},$$
$$A - B = \{1, 3, 5\}, \quad B - A = \{6\}.$$

(2) If A is the set of all soldiers and B the set of all students, then $A \cup B$ consists of all persons who are either soldiers or students or both; $A \cap B$ is the set of all studying soldiers; $A - B$ is the set of all soldiers who do not study; and $B - A$ consists of those students who are not soldiers.

When speaking of sets, we shall always tacitly assume that we are given some "master set", called *the space*, from which our initial elements are selected. From these elements we then form the various sets (subsets of the space); then we proceed to form *families of sets*, etc. The space will often remain unspecified, so that we retain the possibility of changing it if required. If S is

[1] The word "or" is used in mathematics in the *inclusive* sense; that is, "$x \in A$ or $x \in B$" means "$x \in A$ or $x \in B$ or *both*".

[2] Some authors write $A \setminus B$ for $A - B$; some use this notation only if $B \subseteq A$. Others use the terms "sum" and "product" for "union" and "intersection", respectively. We shall not follow this practice.

the space, and E is its subset (i.e., $E \subseteq S$), we call the difference $S - E$ simply the *complement of* E and denote it briefly by $-E$; thus $-E = S - E$ (provided that S is the space and $E \subseteq S$).[3]

The notions of union, intersection, and difference can be graphically illustrated by means of so-called "Venn diagrams"[4] on which they appear as the shaded areas of two or more intersecting circles or other suitable areas. In Figures 1, 2, and 3, we provide Venn diagrams illustrating the union, intersection, and difference of two sets A and B.

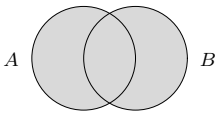


FIGURE 1: $A \cup B$

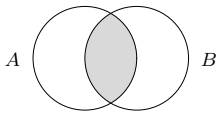


FIGURE 2: $A \cap B$

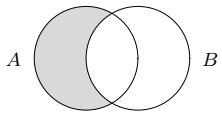


FIGURE 3: $A - B$

Theorem 1. *For any sets* A, B, *and* C, *we have the following*:

(a) $A \cup A = A$; $A \cap A = A$ (idempotent laws).

(b) $A \cup B = B \cup A$; $A \cap B = B \cap A$ (commutative laws).

(c) $(A \cup B) \cup C = A \cup (B \cup C)$

(d) $(A \cap B) \cap C = A \cap (B \cap C)$ } (associative laws).

(e) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

(f) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ } (distributive laws).

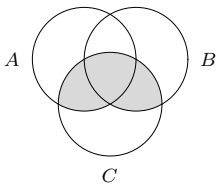
(g) $A \cup \emptyset = A$; $A \cap \emptyset = \emptyset$; $A - \emptyset = A$; $A - A = \emptyset$.

To verify these formulas, we have to check, each time, that every element contained in the set occurring on the left-hand side of the equation also belongs to the right-hand side, and conversely. For example, we shall verify formula (e), leaving the proof of the remaining formulas to the reader. Suppose then that some element x belongs to the set $(A \cup B) \cap C$; this means that $x \in (A \cup B)$ and, simultaneously, $x \in C$; in other words, we have $x \in A$ or $x \in B$ and, simultaneously, $x \in C$. It follows that we have ($x \in A$ and $x \in C$) or ($x \in B$ and $x \in C$); that is, $x \in (A \cap C)$ or $x \in (B \cap C)$, whence $x \in [(A \cap C) \cup (B \cap C)]$. Thus we see that every element x contained in the left-hand side of (e) is also contained in the right-hand side. The converse assertion is proved in the same way by simply reversing the order of the steps of the proof.

In Figures 4 and 5, we illustrate the distributive laws (e) and (f) by Venn diagrams; the shaded area represents the set resulting from the operations involved in each case.

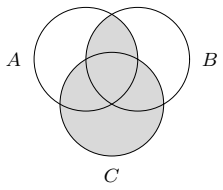
[3] Other notations in use for complement are as follows: $\sim E$, $\widetilde{E}$, $E^{\sim}$, $\complement E$, E', etc.

[4] After the English logician John Venn (1834–1883).



$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

FIGURE 4



$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

FIGURE 5

Because of the associative laws, we may omit the brackets in expressions occurring in formulas (c) and (d). Thus we may write $A \cup B \cup C$ and $A \cap B \cap C$ instead of $(A \cup B) \cup C$ and $(A \cap B) \cap C$, respectively.[5] Similarly, unions and intersections of four or more sets may be written in various ways:

$$A \cup B \cup C \cup D = (A \cup B) \cup (C \cup D) = A \cup (B \cup C \cup D) = (A \cup B \cup C) \cup D;$$
$$A \cap B \cap C \cap D = (A \cap B \cap C) \cap D = (A \cap B) \cap (C \cap D), \text{ etc.}$$

As we noted in §1, we may consider not just one or two, but a whole family of sets, even infinitely many of them. Sometimes we can *number* the sets under consideration: X_1, X_2, X_3, …, X_n, … (compare this to the numbering of buildings in a street, or books in a library). More generally, we may denote all sets of a family $\mathcal{M}$ by one and the same letter (say, X), with some indices (subscripts or superscripts) attached to it: X_i or X^i, where i runs over a suitable (sufficiently large) set I of indices, called the *index set*. The indices may, but need not, be *numbers*. They are just "labels" of arbitrary nature, used solely to distinguish the sets from each other, in the same way that a good cook uses labels to distinguish the jars in the kitchen. The whole family $\mathcal{M}$ then is denoted by $\{X_i \mid i \in I\}$, briefly $\{X_i\}$. Here i is a variable ranging over the index set I. This is called *index notation*.

The notions of union and intersection can easily be extended to arbitrary families of sets. If $\mathcal{M}$ is such a family, we define its *union*, $\bigcup \mathcal{M}$, to be the set of all elements x, each belonging to *at least one* set of the family. The *intersection*, $\bigcap \mathcal{M}$, consists of those elements x that belong to *all* sets of the family *simultaneously*. Instead of $\bigcup \mathcal{M}$ and $\bigcap \mathcal{M}$, we also use

$$\bigcup \{X \mid X \in \mathcal{M}\} \quad \text{and} \quad \bigcap \{X \mid X \in \mathcal{M}\}, \quad \text{respectively.}$$

Here X is a variable denoting any arbitrary set of the family. **Note:** $x \in \bigcup \mathcal{M}$ iff x is in *at least one* set X of the family; $x \in \bigcap \mathcal{M}$ iff x belongs to *every* set X of the family.

[5] As will be seen, unions and intersections of three or more sets can be defined independently. Thus, in set theory, such formulas as $A \cap B \cap C = (A \cap B) \cap C$ or $A \cup B \cup C = (A \cup B) \cup C$ are *theorems*, not definitions.

Thus $\bigcap \mathcal{M}$ is the *common part* of all sets X from $\mathcal{M}$ (possibly $\bigcap \mathcal{M} = \emptyset$), while $\bigcup \mathcal{M}$ comprises all elements of all these sets combined.

If $\mathcal{M} = \{X_i \mid i \in I\}$ (index notation), we also use symbols like

$$\bigcup \{X_i \mid i \in I\} = \bigcup_{i \in I} X_i = \bigcup_i X_i = \bigcup X_i \text{ and}$$
$$\bigcap \{X_i \mid i \in I\} = \bigcap_{i \in I} X_i = \bigcap_i X_i = \bigcap X_i$$

for $\bigcup \mathcal{M}$ and $\bigcap \mathcal{M}$, respectively. Finally, if the indices are *integers*, we use symbols like

$$\bigcup_{n=1}^{\infty} X_n, \ \bigcap_{n=1}^{q} X_n, \ \bigcup_{n=k}^{\infty} X_n, \ X_1 \cup X_2 \cup \cdots \cup X_n \cup \cdots,$$

or the same with $\bigcup$ and $\bigcap$ interchanged, imitating a similar notation known from elementary algebra for sums and products of numbers.

The following theorem has many important applications.

Theorem 2 (de Morgan's duality laws[6])**.** *Given a set E and any family of sets $\{A_i\}$ (where i ranges over some index set I), we always have*

$$\text{(i)} \ \ E - \bigcup_i A_i = \bigcap_i (E - A_i); \qquad \text{(ii)} \ \ E - \bigcap_i A_i = \bigcup_i (E - A_i).$$

Verbally, this reads as follows:

(i) *The complement (in E) of the union of a family of sets equals the intersection of their complements (in E).*

(ii) *The complement (in E) of the intersection of a family of sets equals the union of their complements (in E).*

Proof of (i). We have to show that the set $E - \bigcup_i A_i$ consists of exactly the same elements as the set $\bigcap_i (E - A_i)$, i.e., that we have

$$x \in E - \bigcup_i A_i \text{ iff } x \in \bigcap_i (E - A_i).$$

This follows from the equivalence of the following statements (we indicate log-

[6] Augustus de Morgan, Indian-born English mathematician and logician (1806–1871).

ical inference by arrows):[7]

$$\left\downarrow \begin{array}{l} x \in E - \bigcup_i A_i, \\ x \in E \text{ but } x \notin \bigcup_i A_i, \\ x \in E \text{ but } x \text{ is not in any of the sets } A_i, \\ x \text{ is in each of the sets } E - A_i, \\ x \in \bigcap_i (E - A_i). \end{array} \right\uparrow$$

Similarly for part (ii), which we leave to the reader. □

Note: In the special case where E is the entire space, the duality laws can be written more simply:

$$\text{(i)} \quad -\bigcup_i A_i = \bigcap_i (-A_i); \qquad \text{(ii)} \quad -\bigcap_i A_i = \bigcup_i (-A_i).$$

Note: The duality laws (Theorem 2) hold also when the sets A_i are not subsets of E.

The importance of the duality laws consists in that they make it possible to derive from each general set identity its so-called "dual", i.e., a new identity *that arises from the first by interchanging all "cap" and "cup" signs*. For example, the two associative laws, Theorem 1(c) and (d), are each other's duals, and so are the two distributive laws, (e) and (f).

To illustrate this fact, we shall show how the second distributive law, (f), can be deduced from the first, (e), which has already been proved. Since Theorem 1(e) holds for *any* sets, it also holds for their complements. Thus we have, for any sets A, B, C,

$$(-A) \cap (-B \cup -C) = (-A \cap -B) \cup (-A \cap -C).$$

But, by the duality laws, $-B \cup -C = -(B \cap C)$; similarly,

$$-A \cap -B = -(A \cup B) \text{ and } -A \cap -C = -(A \cup C).$$

Therefore, we obtain

$$-A \cap -(B \cap C) = -(A \cup B) \cup -(A \cup C),$$

or, applying again the duality laws to both sides,

$$-[A \cup (B \cap C)] = -[(A \cup B) \cap (A \cup C)],$$

[7] Sometimes horizontal arrows are used instead of the vertical ones (to be explained in §3).

whence $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, as required. This procedure is quite general and leads to the following *duality rule*: *Whenever an identity holds for all sets, so also does its dual.*[8]

As an exercise, the reader may repeat the same procedure for the two associative laws (prove one of them in the ordinary way and then derive the second by using the duality laws), as well as for the following theorem.

Theorem 3 (Generalized distributive laws)**.** *If E is a set and $\{A_i\}$ is any set family, then*

$$\text{(i)} \quad E \cap \bigcup_i A_i = \bigcup_i (E \cap A_i); \qquad \text{(ii)} \quad E \cup \bigcap_i A_i = \bigcap_i (E \cup A_i).$$

Problems in Set Theory

1. Verify the formulas (c), (d), (f), and (g) of Theorem 1.

2. Prove that $-(-A) = A$.

3. Verify the following formulas (distributive laws with respect to the subtraction of sets), and illustrate by Venn diagrams:

(a) $A \cap (B - C) = (A \cap B) - (A \cap C)$;

(b) $(A - C) \cap (B - C) = (A \cap B) - C$.

4. Show that the relations $(A \cup C) \subset (A \cup B)$ and $(A \cap C) \subset (A \cap B)$, when combined, imply $C \subset B$. Disprove the converse by an example.

5. Describe geometrically the following sets on the real line:

(i) $\{x \mid x < 0\}$; (ii) $\{x \mid |x| < 1\}$;
(iii) $\{x \mid |x - a| < \varepsilon\}$; (iv) $\{x \mid |x| < 0\}$;
(v) $\{x \mid a < x < b\}$; (vi) $\{x \mid a \le x \le b\}$.

6. If (x, y) denotes the set $\{\{x\}, \{x, y\}\}$, prove that, for any x, y, v, u, we have $(x, y) = (u, v)$ iff $x = u$ and $y = v$. Treat this as a definition of an *ordered pair*.
[Hint: Consider separately the two cases $x = y$ and $x \neq y$, noting that $\{x, x\} = \{x\}$.]

7. Let $A = \{x_1, x_2, \ldots, x_n\}$ be a set consisting of n distinct elements. How many subsets does it have? How many proper subsets?

8. Prove that

$$(A \cup B) \cap (B \cup C) \cap (C \cup A) = (A \cap B) \cup (B \cap C) \cup (C \cap A)$$

[8] More precisely, this applies to set identities involving no operations other than $\cap$ and $\cup$; cf. also Problem 10 (iii) below.

in two ways:

(i) using definitions only;

(ii) using the commutative, associative, and distributive laws.

(In the second case, write AB for $A \cap B$ and $A + B$ for $A \cup B$, etc., and proceed to remove brackets, noting that $A + A = A = AA$.)

9. Show that the following relations hold iff $A \subseteq E$:

(i) $(E - A) \cup A = E$;

(ii) $E - (E - A) = A$;

(iii) $A \cup E = E$;

(iv) $A \cap E = A$;

(v) $A - E = \emptyset$.

10. Prove de Morgan's duality laws:

(i) $E - \bigcap X_i = \bigcup (E - X_i)$;

(ii) $E - \bigcup X_i = \bigcap (E - X_i)$;

(iii) if $A \subseteq B$, then $(E - B) \subseteq (E - A)$.

11. Prove the generalized distributive laws:

(i) $A \cap \bigcup X_i = \bigcup (A \cap X_i)$;

(ii) $A \cup \bigcap X_i = \bigcap (A \cup X_i)$;

(iii) $\bigcap X_i \cup \bigcap Y_j = \bigcap_{i,j} (X_i \cup Y_j)$;

(iv) $\bigcup X_i \cap \bigcup Y_j = \bigcup_{i,j} (X_i \cap Y_j)$.

12. In Problem 11, show that (i) and (ii) are *duals* (i.e., follow from each other by de Morgan's duality laws) and so are (iii) and (iv).

13. Prove the following:

$$\text{(i) } \left(\bigcap X_i\right) - A = \bigcap (X_i - A); \quad \text{(ii) } \left(\bigcup X_i\right) - A = \bigcup (X_i - A)$$

(generalized distributive laws with respect to differences).

14. If (x, y) is defined as in Problem 6, which of the following is true?

$$x \in (x, y); \quad \{x\} \in (x, y); \quad y \in (x, y);$$
$$\{y\} \in (x, y); \quad \{x, y\} \in (x, y); \quad \{x\} = (x, x); \quad \{\{x\}\} = (x, x).$$

15. Prove that

(i) $A - B = A \cap -B = (-B) - (-A) = -((-A) \cup B)$ and

(ii) $A \cap B = A - (-B) = B - (-A) = -(-A \cup -B)$.

Give also four various expressions for $A \cup B$.

16. Prove the following:

(i) $(A \cup B) - B = A - B = A - (A \cap B)$;

(ii) $(A - B) - C = A - (B \cup C)$;

(iii) $A - (B - C) = (A - B) \cup (A \cap C)$;

(iv) $(A - B) \cap (C - D) = (A \cap C) - (B \cup D)$.

17. The *symmetric difference* of two sets A and B is

$$A \triangle B = (A - B) \cup (B - A).$$

Prove the following:

(i) $A \triangle B = B \triangle A$;

*(ii) $A \triangle (B \triangle C) = (A \triangle B) \triangle C$;

(iii) $A \triangle \emptyset = A$;

(iv) If $A \cap B = \emptyset$, $A \triangle B = A \cup B$;

(v) If $A \supseteq B$, $A \triangle B = A - B$;

(vi) $A \triangle B = (A \cup B) - (A \cap B) = (A \cup B) \cap (-A \cup -B)$;

(vii) $A \triangle A = \emptyset$;

(viii) $A \triangle B = (-A) \triangle (-B)$;

(ix) $-(A \triangle B) = A \triangle (-B) = (-A) \triangle B = (A \cap B) \cup (-A \cap -B)$;

(x) $(A \triangle B) \cap C = (A \cap C) \triangle (B \cap C)$.

***18.** For $n = 2, 3, \ldots$ define the following:

$$\overset{n}{\underset{i=1}{\triangle}} A_i = A_1 \triangle A_2 \triangle \cdots \triangle A_n = (A_1 \triangle A_2 \triangle \cdots \triangle A_{n-1}) \triangle A_n.$$

Prove that $x \in \triangle_{i=i}^{n} A_i$ iff $x \in A_i$ for an *odd* number of values of i.

19. Use Venn diagrams to check the consistency of this report: Of 100 patients, 47 were inoculated against smallpox, 43 against polio, 51 against tetanus, 21 against both smallpox and polio, and 19 against tetanus and polio, while 7 had to obtain all three shots.

***20.** (Russell paradox.) A set M is said to be *abnormal* iff $M \in M$, i.e., iff it contains itself as one of its members (such as, e.g., the family of "all possible" sets); and *normal* iff $M \notin M$. Let $\mathcal{N}$ be the class of all normal sets, i.e., $\mathcal{N} = \{X \mid X \notin X\}$. Is $\mathcal{N}$ itself normal? Verify that any answer to this question implies its own negation, and thus the very definition of $\mathcal{N}$ is contradictory, i.e., $\mathcal{N}$ is an impossible ("contradictory") set. (To exclude this and other paradoxes, various systems of axioms have been set up, so as to define which sets may, and which may not, be formed.)

§3. Logical Quantifiers

From logic we borrow the following widely-used abbreviations:

"$(\forall x \in A)$ ..." means "For each member x of A, it is true that ...."

"$(\exists x \in A)$ ..." means "There is at least one x in A such that ...."

"$(\exists! x \in A)$ ..." means "There is a *unique* x in A such that ...."

The symbols "$(\forall x \in A)$" and "$(\exists x \in A)$" are called the *universal* and *existential quantifiers*, respectively. If confusion is ruled out, we simply write "$(\forall x)$", "$(\exists x)$", and "$(\exists! x)$" instead. For example, if N is the set of all *naturals* (positive integers), then the formula

$$\text{"}(\forall n \in N)\ (\exists m \in N)\ m > n\text{"}$$

means "For each natural n there is a natural m such that $m > n$." If we agree that m, n denote *naturals*, we may write "$(\forall n)\ (\exists m)\ m > n$" instead. Some more examples follow:

Let $\mathcal{M} = \{A_i \mid i \in I\}$ be an indexed set family (see §2). By definition, $x \in \bigcup_i A_i$ means that x is in *at least one* of the sets A_i. In other words, *there is at least one index $i \in I$ for which $x \in A_i$*; in symbols, $(\exists i \in I)\ x \in A_i$. Thus

$$x \in \bigcup_{i \in I} A_i \text{ iff } (\exists i \in I)\ x \in A_i; \quad \text{similarly,} \quad x \in \bigcap_i A_i \text{ iff } (\forall i)\ x \in A_i.$$

Also note that $x \notin \bigcup_i A_i$ iff x is in *none* of the A_i, i.e., $(\forall i)\ x \notin A_i$. Similarly, $x \notin \bigcap_i A_i$ iff x fails to be in *some* A_i, i.e., $(\exists i)\ x \notin A_i$. Thus

$$x \notin \bigcap_i A_i \text{ iff } (\exists i)\ x \notin A_i; \qquad x \notin \bigcup_i A_i \text{ iff } (\forall i)\ x \notin A_i.$$

As an application, we now prove Theorem 2 of §2, *using quantifiers*:

(i) $\left\downarrow \begin{array}{l} x \in E - \bigcup A_i, \\ x \in E \text{ but } x \notin \cup A_i, \\ x \in E \text{ and } (\forall i)\ x \notin A_i, \\ (\forall i)\ x \in E - A_i, \\ x \in \bigcap (E - A_i). \end{array} \right\uparrow$ (ii) $\left\downarrow \begin{array}{l} x \in E - \bigcap A_i, \\ x \in E \text{ but } x \notin \cap A_i, \\ x \in E \text{ and } (\exists i)\ x \notin A_i, \\ (\exists i)\ x \in E - A_i, \\ x \in \bigcup (E - A_i). \end{array} \right\uparrow$. □

The reader should practice such examples thoroughly. Quantifiers not only shorten formulations but often make them more precise. We shall therefore briefly dwell on their properties.

Order. The order in which quantifiers follow each other is *essential*; e.g., the formula

$$\text{"}(\forall n \in N)\ (\exists m \in N)\ m > n\text{"}$$

(each natural n is exceeded by some $m \in N$) is true; but

$$\text{"}(\exists m \in N)\ (\forall n \in N)\ m > n\text{"}$$

is *false* since it states that some natural m exceeds *all* naturals. However, two *consecutive* universal quantifiers (or two *consecutive* existential ones) may be interchanged. Instead of "$(\forall x \in A)\ (\forall y \in A)$" we briefly write "$(\forall x, y \in A)$". Similarly, we write "$(\exists x, y \in A)$" for "$(\exists x \in A)\ (\exists y \in A)$", "$(\forall x, y, z \in A)$" for "$(\forall x \in A)\ (\forall y \in A)\ (\forall z \in A)$", and so on.

Qualifications. Sometimes a formula $P(x)$ holds not for all $x \in A$, but only for those with some additional property $Q(x)$. This will be written as "$(\forall x \in A \mid Q(x))\ P(x)$," where the vertical stroke $\mid$ stands for "*such that*". For example, if N is again the naturals, then the formula

$$\text{"}(\forall x \in N \mid x > 3) \quad x \geq 4\text{"} \tag{1}$$

means "For each natural x *such that* $x > 3$, it is true that $x \geq 4$." In other words, for *naturals*, $x > 3$ *implies* $x \geq 4$; this is also written

$$\text{"}(\forall x \in N) \quad [x > 3 \implies x \geq 4]\text{"}$$

(the arrow $\implies$ stands for "*implies*"). The symbol $\iff$ is used for "iff" ("if and only if"). For instance,

$$\text{"}(\forall x \in N) \quad [x > 3 \iff x \geq 4]\text{"}$$

means "For natural numbers x, we have $x > 3$ *if and only if* $x \geq 4$."

Negations. In mathematics, we often have to form the *negation* of a formula that starts with one or several quantifiers. Then it is noteworthy that *each universal quantifier is replaced by an existential one* (*and vice versa*), followed by the negation of the subsequent part of the original formula. For example, in calculus, a real number p is called the *limit* of a sequence $x_1, x_2, \ldots, x_n, \ldots$ iff the following is true: "*For every real $\varepsilon > 0$, there is a natural k (depending on ε) such that for all integers $n > k$, we have $|x_n - p| < \varepsilon$*." If we agree that lower-case letters (possibly with subscripts) denote real numbers, and that n, k denote naturals, this sentence can be written thus:

$$(\forall \varepsilon > 0)\ (\exists k)\ (\forall n > k) \quad |x_n - p| < \varepsilon. \tag{2}$$

Here "$(\forall \varepsilon > 0)$" and "$(\forall n > k)$" stand for "$(\forall \varepsilon \mid \varepsilon > 0)$" and "$(\forall n \mid n > k)$". Such self-explanatory abbreviations will also be used in other similar cases.

Now let us form the negation of (2). As (2) states that "*for all $\varepsilon > 0$*" something (i.e., the rest of the formula) is true, the negation of (2) starts with "*there is an $\varepsilon > 0$*" (for which the rest of the formula *fails*). Thus we start with "$(\exists \varepsilon > 0)$" and form the negation of the rest of the formula, i.e., of "$(\exists k)\ (\forall n > k)\ |x_n - p| < \varepsilon$". This negation, in turn, starts with "$(\forall k)$" (why?), and

so on. Step by step, we finally arrive at

$$(\exists \varepsilon > 0)\ (\forall k)\ (\exists n > k) \quad |x_n - p| \geq \varepsilon,$$

i.e., "there is at least one $\varepsilon > 0$ such that, for every natural k, one can find an integer $n > k$, with $|x_n - p| \geq \varepsilon$". Note that here the choice of n may depend on k. To stress it, we write n_k for n. Thus the negation of (2) emerges as

$$(\exists \varepsilon > 0)\ (\forall k)\ (\exists n_k > k) \quad |x_{n_k} - p| \geq \varepsilon. \tag{3}$$

Rule: *To form the negation of a quantified formula, replace all universal quantifiers by existential ones, and conversely; finally, replace the remaining (unquantified) formula by its negation.* Thus, in (2), "$|x_n - p| < \varepsilon$" must be replaced by "$|x_n - p| \geq \varepsilon$", or rather by "$|x_{n_k} - p| \geq \varepsilon$", as explained.

Note 1. Formula (3) is also the negation of (2) when (2) is written as

$$\text{"}(\forall \varepsilon > 0)\ (\exists k)\ (\forall n) \quad [n > k \implies |x_n - p| < \varepsilon]\text{"}.$$

In general, to form the negation of a formula containing the implication sign $\implies$, it is advisable first to re-write all *without* that sign, using the notation "$(\forall x \mid \dots)$" (here: "$(\forall n \mid n > k)$").

Note 2. The *universal* quantifier in a formula $(\forall x \in A)\ P(x)$ *does not* imply the existence of an x for which $P(x)$ is true. It is only meant to imply that *there is no x in A for which $P(x)$ fails*. This remains true even if $A = \emptyset$; we then say that "$(\forall x \in A)\ P(x)$" is *vacuously* true. For example, the statement "all witches are beautiful" is vacuously true because there are no witches at all; but so also is the statement "all witches are ugly". Similarly, the formula $\emptyset \subseteq B$, i.e., $(\forall x \in \emptyset)\ x \in B$, is vacuously true.

Problem. Redo Problems 11 and 13 of §2 using *quantifiers*.

§4. Relations (Correspondences)

We already have occasionally used terms like "relation", "operation", etc., but they did not constitute part of our theory. In this and the next sections, we shall give a precise definition of these concepts and dwell on them more closely.

Our definition will be based on the concept of an *ordered pair*. As has already been mentioned, by an ordered pair (briefly "pair") (x, y), we mean two (possibly equal) objects x and y given in a *definite order*, so that one of them, x, becomes the *first* (or *left*) and the other, y, is the *second* (or *right*) part of the pair.[1] We recall that two pairs (a, b) and (x, y) are equal iff their *corresponding* members are the same, that is, iff $a = x$ and $b = y$. The pair

[1] For a more precise definition (avoiding the undefined term "order"), see Problem 6 after §2.

(y, x) should be distinguished from (x, y); it is called the *inverse* to (x, y). Once a pair (x, y) has been formed, it is treated as a new thing (i.e., as *one* object, different from x and y taken separately); x and y are called the *coordinates* of the pair (x, y).

Nothing prevents us, of course, from considering also *sets* of ordered pairs, i.e., *sets whose elements are pairs*, (each pair being regarded as *one* element of the set). If the pair (x, y) is an element of such a set R, we write $(x, y) \in R$. **Note:** This *does not* imply that x and y *taken separately*, are elements of R; (then we write $x, y \in R$).

Definition 1.

By a *relation*, or *correspondence*, we mean any set of ordered pairs.[2]

If R is a relation, and $(x, y) \in R$, then y is called an *R-relative* of x (but x is not called an R-relative of y unless $(y, x) \in R$); we also say in this case that y is *R-related to x* or that the *relation R holds between x and y*. Instead of $(x, y) \in R$, we also write xRy. The letter R, designating a relation, may be replaced by other letters; it is often replaced by special symbols like $<$, $>$, $\sim$, $\equiv$, etc.

Examples.

(1) Let R be the set of all pairs (x, y) of integers x and y such that x is less than y.[3] Then R is a relation (called "*inequality relation between integers*"). The formula xRy means in this case that x and y are integers, with x less than y. Usually the letter R is here replaced by the special symbol $<$, so that "xRy" turns into "$x < y$".

(2) The inclusion relation $\subseteq$ introduced in §1 may be interpreted as the set of all pairs (X, Y) where X and Y are subsets of a given space, with X a subset of Y. Similarly, the $\in$-relation is the set of all pairs (x, A) where A is a subset of the space and x is an element of A.

(3) *$\emptyset$ is a relation* ("an empty set of pairs").

If $P(x, y)$ is a proposition or formula involving the variables x and y, we denote by $\{(x, y) \mid P(x, y)\}$ the set of all ordered pairs for which the formula $P(x, y)$ is true. For example, the set of all married couples could be denoted by $\{(x, y) \mid x \text{ is the wife of } y\}$.[4] Any such set is a relation.

[2] This use of the term "relation" may seem rather strange to a reader unfamiliar with exact mathematical terminology. The justification of this definition is in that it fits exactly all mathematical purposes, as will be seen later, and makes the notion of relation precise, reducing it to that of a "*set*".

[3] Though the theory of integers and real numbers will be formally introduced only in Chapter 2, we feel free to use them in illustrative examples.

[4] This set could be called "the relation of being married".

Since relations are *sets*, the equality of two relations, R and S, means that they consist of exactly the same elements (ordered pairs); that is, *we have $R = S$ iff xRy always implies xSy, and vice versa*. Similarly, $R \subseteq S$ means that xRy always implies xSy (but the converse need not be true).

By replacing all pairs (x, y) belonging to a relation R by their inverses (y, x) we obtain a new relation, called the *inverse* of R and denoted by R^{-1}. Clearly, we have $xR^{-1}y$ iff yRx; thus

$$R^{-1} = \{(x, y) \mid yRx\} = \{(y, x) \mid xRy\}.$$

This shows that R, in its turn, is the inverse of R^{-1}; i.e., $(R^{-1})^{-1} = R$. For example, the relations $<$ and $>$ between numbers are inverse to each other; so also are the relations $\subseteq$ and $\supseteq$ between sets.

If a correspondence R contains the ordered pairs (x, x'), (y, y'), (z, z'), ..., we shall write

$$R = \begin{pmatrix} x & y & z & \dots \\ x' & y' & z' & \dots \end{pmatrix}, \tag{1}$$

i.e., the pairs will be written in *vertical* notation, so that each left coordinate of a pair is written *above* the corresponding right coordinate (i.e., above its R-relative). Thus, e.g., the symbol

$$\begin{pmatrix} 1 & 4 & 1 & 3 \\ 2 & 2 & 1 & 1 \end{pmatrix} \tag{2}$$

denotes the relation consisting of the four pairs $(1, 2)$, $(4, 2)$, $(1, 1)$, and $(3, 1)$. The inverse relation is obtained by simply interchanging the upper and the lower rows.

Definition 2.

The set of all left coordinates of pairs contained in a relation R is called the *domain* of R, denoted D_R. The set of all right coordinates of these pairs is called the *range* or *co-domain* of R, denoted D'_R. Clearly, $x \in D_R$ iff xRy for some y. Thus (note these formulas)

$$D_R = \{x \mid xRy \text{ for some } y\}; \qquad D'_R = \{y \mid xRy \text{ for some } x\};$$

or, using quantifiers,

$$D_R = \{x \mid (\exists y)\ xRy\}; \qquad D'_R = \{y \mid (\exists x)\ xRy\}.$$

In symbols of the form (1), the domain and range appear as the upper and the lower row, respectively; thus, e.g., in (2) the domain is $\{1, 4, 3\}$ and the range is $\{2, 1\}$. Clearly, if all pairs of a relation R are replaced by their inverses, then the left coordinates turn into the right ones, and conversely. Therefore,

the domain of the inverse relation R^{-1} coincides with the range of R, and the range of R^{-1} is the domain of R; that is,

$$D_{R^{-1}} = D'_R, \qquad D'_{R^{-1}} = D_R. \tag{3}$$

Definition 3.

Given a relation R and any set A we say that R is

(i) *reflexive* on A iff we have xRx for all elements x of A;

(ii) *symmetric* on A iff xRy implies yRx for any x and y in A;

(iii) *transitive* on A iff xRy combined with yRz implies xRz for all x, y, and z in A;

(iv) *trichotomic* on A iff, for any x and y in A, we always have either xRy, or yRx, or $x = y$, but never two of these together.

Examples.

(a) The inequality relation $<$ between real numbers is transitive and trichotomic because $x < y$ and $y < z$ always implies $x < z$ (transitivity); and we always have either $x < y$, or $y < x$, or $x = y$ (trichotomy); we shall dwell on these properties more closely in Chapter 2.

(b) The inclusion relation $\subseteq$ between sets is reflexive (because $A \subseteq A$) and transitive (because $A \subseteq B$ and $B \subseteq C$ implies $A \subseteq C$); but it is neither symmetric nor trichotomic, the latter because it may well happen that neither of two sets contains the other, and because $A \subseteq B$ and $A = B$ may *both* hold.

(c) The relation of *proper* inclusion, $\subset$, is only transitive.

(d) The equality relation, $=$, is reflexive, symmetric, and transitive because we always have $x = x$, $x = y$ always implies $y = x$, and $x = y = z$ implies $x = z$. It is, however, not trichotomic. (Why?)

(e) The $\in$ relation between an element and a set is neither reflexive nor symmetric, nor transitive, nor trichotomic (on the set $\mathcal{A}$ consisting of all elements and all subsets of a given space).

Definition 4.

The *image* of a set A under a relation R (briefly, the *R-image of A*) is the set of all R-relatives of elements of A; it is denoted by $R[A]$ (*square* brackets always!). The *inverse image* (the R^{-1}-image) of A, denoted $R^{-1}[A]$, is the image of A under the inverse relation, R^{-1}. The R-image of a *single element* x (or of the *set* $\{x\}$) is simply the set of all R-relatives of x. It is customary to denote it by $R[x]$ instead of the more precise notation $R[\{x\}]$. **Note:** $R[A]$ may be empty!

To form $R[A]$, we first find the R-relatives of every element x of A (if any), thus obtaining $R[x]$ for each $x \in A$. The union of all these $R[x]$ combined is the desired image $R[A]$.

Example.

Let

$$R = \begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 3 & 4 \\ 1 & 3 & 4 & 5 & 3 & 4 & 1 & 3 & 5 & 1 \end{pmatrix}.$$

Then $R[1] = \{1, 3, 4\}$; $R[2] = \{3, 5\}$; $R[3] = \{1, 3, 4, 5\}$; $R[5] = \emptyset$; $R^{-1}[1] = \{1, 3, 4\}$; $R^{-1}[2] = \emptyset$; $R^{-1}[3] = \{1, 2, 3\}$; $R^{-1}[4] = \{1, 3\}$.

If, further, $A = \{1, 2\}$ and $B = \{2, 4\}$, then $R[A] = \{1, 3, 4, 5\}$; $R[B] = \{1, 3, 5\}$; $R^{-1}[A] = \{1, 3, 4\}$; and $R^{-1}[B] = \{1, 3\}$.

By definition, $R[x]$ is the set of all R-relatives of x. Hence $y \in R[x]$ means that y is an R-relative of x, i.e., that $(x, y) \in R$, which can also be written as xRy. Thus *the formulas*

$$(x, y) \in R, \quad xRy \quad \textit{and} \quad y \in R[x]$$

are equivalent. More generally, $y \in R[A]$ means that y is an R-relative of some element $x \in A$; i.e., *there is* $x \in A$ *such that* $(x, y) \in R$. In symbols, $y \in R[A]$ *is equivalent to* $(\exists x \in A)\ (x, y) \in R$, *or* $(\exists x \in A)\ xRy$.

Note that *the expressions* $R[A]$, $R^{-1}[A]$, $R[x]$ *and* $R^{-1}[x]$ *are defined even if* A *or* x *are not contained in the domain* (*or range*) *of* R. These images may, however, be empty. In particular, $R[x] = \emptyset$ iff $x \notin D_R$.

We conclude this section with an important example of a relation. Given any two sets A and B, we can consider the set of *all* ordered pairs (x, y) with $x \in A$ and $y \in B$. This set is called the *Cartesian product*, or *cross product*, of A and B, denoted $A \times B$. Thus

$$A \times B = \{(x, y) \mid x \in A,\ y \in B\}.$$

In particular, $A \times A$ is the set of all ordered pairs that can be formed of elements of A. **Note:** $A \times \emptyset = \emptyset \times A = \emptyset$. (Why?)

The Cartesian product $A \times B$ is a *relation* since it is a set of ordered pairs. Its domain is A and its range is B (provided that A and B are not empty). Moreover, it is the "largest" possible relation with this domain and this range, because any other relation with the same domain and range is a subset of $A \times B$, i.e., it contains only *some* of the ordered pairs contained in $A \times B$. Thus, to form a relation with domain A and range B means to select certain pairs from $A \times B$. The inverse of $A \times B$ is $B \times A$ (the set of all inverse pairs).

On the other hand, the formation of Cartesian products may also be treated as a new operation on sets (called *cross multiplication*). This operation is not commutative since, in general, the inverse relation $B \times A$ is different from $A \times B$, so that $A \times B \neq B \times A$. It is also not associative; i.e., we have, in general, $(A \times B) \times C \neq A \times (B \times C)$. (Why?) Nevertheless, we can speak of cross products of more than two sets if we agree to write $A \times B \times C$ for $(A \times B) \times C$ (but not for $A \times (B \times C)$). Similarly, we define

$$A \times B \times C \times D = (A \times B \times C) \times D, \quad A \times B \times C \times D \times E = (A \times B \times C \times D) \times E,$$

etc. Instead of $A \times A$, we also write A^2. Similarly, $A^3 = A \times A \times A$, $A^4 = A \times A \times A \times A$, etc.

There is a simple and suggestive graphic representation of the Cartesian product $A \times B$. Take two perpendicular straight lines OX and OY. Represent A and B symbolically as line segments on OX and OY, respectively. Then the rectangle $PQRS$ (see diagram) represents $A \times B$. Of course, this representation is *symbolic* only since the sets A and B need not actually be line segments, and $A \times B$ need not actually be a rectangle in the xy-plane. This is similar to Venn diagrams, where sets are *symbolically* represented by discs or other areas.

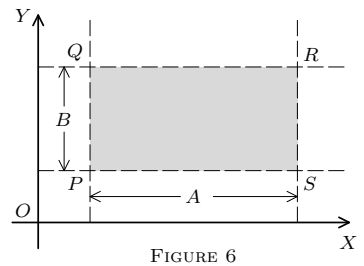


FIGURE 6

Problems in the Theory of Relations

1. For each of the following relations R, find its domain D_R, its range D'_R, and the inverse relation R^{-1}. Specify some values (if any) of x and y such that xRy is true, and some for which it is false; similarly for $xR^{-1}y$.

(i) $R = \begin{pmatrix} 1 & 1 & 2 & 3 & 7 \\ 3 & 1 & 4 & 4 & 0 \end{pmatrix}$; (ii) $R = \begin{pmatrix} 3 & 7 & 1 & -15 & 2 \\ 1 & 8 & 2 & -20 & 9 \end{pmatrix}$;

(iii) $R = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix}$; (iv) $R = \begin{pmatrix} 3 & 5 & 7 & 9 & 11 & 2 \\ 2 & 4 & 0 & 1 & 1 & 5 \end{pmatrix}$;

(v) $R = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$; (vi) $R = \emptyset$.

1′. In Problem 1(i)–(vi), find $R[A]$ and $R^{-1}[A]$, given that

(a) $A = \left\{\frac{1}{2}\right\}$; (b) $A = \{1\}$;

(c) $A = \{7\}$; (d) $A = \{0\}$;

(e) $A = \emptyset$; (f) $A = \{0, 3, -15\}$;

(g) $A = \{3, 4, 7, 0, -1, 6\}$; (h) $A = \{3, 8, 2, 4, 5\}$;

(i) $A = E^1$ (= the entire real axis); (j) $A = \{x \in E^1 \mid -20 < x < 5\}$.

2. Describe the following sets in the xy-plane:

(i) $\{(x,y) \mid x < y\}$; (ii) $\{(x,y) \mid x^2 + y^2 < 1\}$;

(iii) $\{(x,y) \mid \max(|x|,|y|) < 1\}$; (iv) $\{(x,y) \mid |x| + |y| \leq 4\}$;

(v) $\{(x,y) \mid (x-2)^2 + (y+5)^2 > 9\}$; (vi) $\{(x,y) \mid y^2 \geq x\}$;

(vii) $\{(x,y) \mid x^2 + y < 1\}$; (viii) $\{(x,y) \mid x^2 - 2xy + y^2 < 0\}$;

(ix) $\{(x,y) \mid x^2 - 2xy + y^2 = 0\}$.

Treating each of these sets as a relation R, answer the same questions as in Problem 1. Then find $R[A]$ and $R^{-1}[A]$ as in Problem 1′.

3. Prove the following: If $A \subseteq B$, then $R[A] \subseteq R[B]$. Disprove the converse by giving an example in which $R[A] \subseteq R[B]$ but $A \nsubseteq B$.

4. Prove the following:

(i) $R[A \cup B] = R[A] \cup R[B]$;

(ii) $R[A \cap B] \subseteq R[A] \cap R[B]$;

(iii) $R[A - B] \supseteq R[A] - R[B]$.

Generalize formulas (i) and (ii) by proving them with A, B replaced by an arbitrary family of sets $\{A_i\}$ $(i \in I)$. Disprove the reverse inclusions in (ii) and (iii) by counterexamples (thus showing that equality may fail). Also, try to prove them and explain where and why the proof fails.

5. State and prove necessary and sufficient conditions for the following:

(i) $R[x] = \emptyset$; (ii) $R^{-1}[x] = \emptyset$; (iii) $R[A] = \emptyset$; (iv) $R^{-1}[A] = \emptyset$.

6. In what case does $R[x] \subseteq A$ imply $x \in R^{-1}[A]$? Give a proof.

7. Which of the relations specified in Problems 1 and 2 are transitive, reflexive, symmetric, or trichotomic on A if

(i) $A = D_R \cup D'_R$? (ii) $A = \{1\}$? (iii) $A = \emptyset$?

8. In Problem 1, add (as few as possible) new pairs to each of the relations R, so as to make them reflexive, symmetric, and transitive. Try to achieve the same results by *dropping* some pairs.

8′. Solve (as far as possible) Problem 8 for *trichotomy*.

9. Is R^{-1} reflexive, symmetric, transitive, or trichotomic on a set A if R is? (Give a proof or a counterexample.) Consider the general case and the case $A = D_R \cup D'_R$.

10. Let R be a relation with $D_R = D'_R = A$. Show that

(i) R is symmetric on A iff $R = R^{-1}$;

(ii) R is reflexive on A iff $R \supseteq I_A$, where $I_A = \{(x,x) \mid x \in A\}$ is the identity relation on A;

(iii) R is trichotomic on A if $R \cap R^{-1} = \emptyset = R \cap I_A$ and $A \times A \subseteq R \cup R^{-1} \cup I_A$.

11. Let R be a transitive relation on $A \neq \emptyset$ with $D_R = D'_R = A$, and let $S = \{(x,y) \in R \mid (y,x) \notin R\}$. Prove that S is transitive and show by example that it may or may not be trichotomic.

***12.** Show by examples that a relation R may have any two of the properties "reflexive", "symmetric", and "transitive" on a set A, without possessing the third one (i.e., the three properties are *independent* of each other).

13. Which of the properties "reflexive", "symmetric", "transitive", and "trichotomic" (on $A = D_R \cup D'_R$) does the relation R possess if xRy means

(i) x is a brother of y;

(ii) x is an ancestor of y;

(iii) x is the father of y;

(iv) x and y are integers, such that x divides y;

(v) x and y are concentric disks in a plane such that $x \subset y$;

(vi) $x \in A$ and $y \in A$.

14. Treat $A \times B$ as a relation. What are its inverse, domain, and range? What if $A = \emptyset$ or $B = \emptyset$? How many elements (ordered pairs) does $A \times B$ contain if A has m elements and B has n elements (both finite)? How many subsets?[5]

15. Prove the following identities, and illustrate by diagrams. (In each case show that a *pair* (x,y) is in one set iff it is in the other.)

(i) $(A \cup B) \times C = (A \times C) \cup (B \times C)$;

(ii) $(A \cap B) \times (C \cap D) = (A \times C) \cap (B \times D)$;

*(iii) $(X \times Y) - (X' \times Y') = [(X \cap X') \times (Y - Y')] \cup [(X - X') \times Y]$.

16. Prove the following:

(i) $(A \times B) \cap (C \times D) = \emptyset$ iff $A \cap C = \emptyset$ or $B \cap D = \emptyset$;

(ii) $A \times B = C \times D$ iff each product has $\emptyset$ as one of the factors or $A = C$ and $B = D$;

[5] In this and the following problems, we shall be satisfied with the intuitive notion of a *finite set* and the number of its elements. A precise definition of a finite set will be given in §8.

(iii) If $A \times B = (A' \times B') \cup (A'' \times B'')$, with all three products not void, then we have $A = A' \cup A''$, $B = B' \cup B''$, and at least one of $A' = A''$, $B' = B''$, $A' \times B' \subseteq A'' \times B''$, or $A'' \times B'' \subseteq A' \times B'$.

(iv) If $A \neq \emptyset \neq B$ and $(A \times B) \cup (B \times A) = C \times C$, then $A = B = C$.

(v) If A has at least two elements p and q, then $(A \times \{p\}) \cup (\{q\} \times A) \neq A \times A$.

17. Prove the following:

(i) $(\bigcup A_i) \times B = \bigcup (A_i \times B)$;

(ii) $(\bigcap A_i) \times B = \bigcap (A_i \times B)$;

(iii) $(\bigcup_i A_i) \times (\bigcup_j B_j) = \bigcup_{i,j} (A_i \times B_j)$;

(iv) $\bigcap_i (A_i \times B_i) = (\bigcap_i) A_i) \times (\bigcap B_i)$;

(v) $\bigcap_i (A_i \times B_i \times C_i) = (\bigcap_i A_i) \times (\bigcap_i B_i) \times (\bigcap_i C_i)$.

***18.** We say that a family $\mathcal{M}$ of sets is *closed under intersections* iff $\mathcal{M}$ contains the intersection of any two of its members, i.e., iff

$$(\forall X, Y \in \mathcal{M}) \quad X \cap Y \in \mathcal{M}.$$

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two such set families, and let $\mathcal{P}$ be the family of all cross products $X \times Y$, with $X \in \mathcal{M}_1$, $Y \in \mathcal{M}_2$. Show that $\mathcal{P}$ is likewise closed under intersections.

[Hint: Use Problem 15(ii).]

***19.** In Problem 18 assume that the families $\mathcal{M}_1$ and $\mathcal{M}_2$ also have the following property: The difference $X - Y$ of any two sets $X, Y \in \mathcal{M}_i$ can always be represented as a union of finitely many disjoint members of $\mathcal{M}_i$ $(i = 1, 2)$. Show that, then, the family $\mathcal{P}$ also has this property.

[Hint: First, verify the following identity (see Problem 15 (iii)):

$$(X \times Y) - (X' \times Y') = [(X - X') \times Y] \cup [(X \cap X') \times (Y - Y')].$$

Note that the union on the right side is disjoint. (Why?) Now, if $X, X' \in \mathcal{M}_1$ and $Y, Y' \in \mathcal{M}_2$, then $X - X'$ and $Y - Y'$ can be represented as finite disjoint unions, say $X - X' = \bigcup_{i=1}^{m} X_i$, $Y - Y' = \bigcup_{k=1}^{n} Y_k$, with $X_i \in \mathcal{M}_1$, $Y_k \in \mathcal{M}_2$, and the required decomposition of $(X \times Y) - (X' \times Y')$ is obtained by Problem 17 (iii).]

§5. Mappings

We shall now consider an especially important class of relations, called *mappings* or *functions*. The mapping concept is a generalization of that of a function as usually given in calculus.

Definition 1.

A relation R is a *mapping*, a *map*, or a *function* iff the image $R[x]$ for every element $x \in D_R$ consists of a single element (in other words, every element $x \in D_R$ has a *unique* relative under R). This unique element is denoted by $R(x)$ and is called the *function value* at x. (Thus $R(x)$ is the unique element of $R[x]$.)[1] Equivalently, R is a mapping iff no two pairs belonging to R have the same *first* coordinate. (Explain!)

If, in addition, different elements of D_R have *different* images, R is called a *one-to-one-mapping* or a *one-to-one correspondence*. In this case,

$$x \neq y \text{ implies } R(x) \neq R(y),$$

provided that $x, y \in D_R$. Equivalently,

$$R(x) = R(y) \text{ implies } x = y \text{ for } x, y \in D_R.$$

Mappings will usually be denoted by the letters f, g, h, F, φ, ψ, etc.

A mapping f is said to be "*from A to B*" if $D_f = A$ and $D'_f \subseteq B$. In this case we write $f\colon A \to B$. If, in particular, $D_f = A$ and $D'_f = B$, we say that f is a mapping *of A onto B* and write $f\colon A \underset{\text{onto}}{\to} B$. If f is both onto and one-to-one, we write $f\colon A \underset{\text{onto}}{\leftrightarrow} B$. We shall also use expressions like "f maps A into B" and "f maps A onto B" instead of $f\colon A \to B$ and $f\colon A \underset{\text{onto}}{\to} B$, respectively.

Since every element $x \in D_f$ has a *unique* f-relative, $f(x)$, under a mapping f, all pairs belonging to f have the form $(x, f(x))$, where $f(x)$ is the function value at x. Therefore, *in order to define a function f, it suffices to define its domain D_f and to indicate the function value $f(x)$ for every $x \in D_f$.*[2] We shall often use such definitions.

It is customary to say that a function f is *defined on a set A* if $A = D_f$.[3]

Examples.

(1) The relation $R = \{(x, y) \mid x \text{ is the wife of } y\}$ is a one-to-one map of the set of all wives onto the set of all husbands. Under this map, every husband is the (unique) R-relative of his wife. The inverse relation, R^{-1}, is a one-to-one map of the set of all husbands onto the set of all wives.

(2) The relation $f = \{(x, y) \mid y \text{ is the father of } x\}$ is a mapping of the set of all people onto the set of their fathers. It is not one-to-one since several

[1] $R(x)$ is often called the *image of x under R* if confusion with $R[x]$ is irrelevant. Note that $R(x)$ is defined only if $x \in D_R$, whereas $R[x]$ is *always* defined. If $x \notin D_R$, $R[x] = \emptyset$.

[2] Note, however, that it does not suffice to give a formula for $f(x)$ only, without indicating the domain D_f.

[3] In this connection, D_f is often referred to as the *domain of definition* of the function, while D'_f is called its *range of values*.

persons may have one and the same father, and thus $x \neq x'$ does not imply $f(x) \neq f(x')$.

(3) Let g be the set of the four pairs $(1,2)$, $(2,2)$, $(3,3)$, $(4,8)$. Then g is a mapping from $D_g = \{1, 2, 3, 4\}$ onto $D'_g = \{2, 3, 8\}$, with $g(1) = 2$, $g(2) = 2$, $g(3) = 3$, $g(4) = 8$. (These formulas could serve as the definition of g.)[4] It is not one-to-one since $g(1) = g(2)$, i.e., two distinct elements of the domain have one and the same image.

(4) Let the domain of a mapping f be the set of all integers, J, with $f(x) = 2x$ for every integer x. By what has been said above, f is well defined. f is one-to-one since $x \neq y$ implies $2x \neq 2y$. The domain of f is J; its range, however, consists of *even* integers only. Thus $f\colon J \to J$, but it is not *onto* J. This example shows that *a mapping may be one-to-one without being onto*.[5]

(5) The *identity map* (denoted I) is the set of all pairs of the form (x, x) where x ranges over some given space (i.e., it is the set of all pairs with *equal* left and right coordinates). It can also be defined by the formula $I(x) = x$ for each x; that is, the function value at x is x itself. This map is clearly one-to-one and onto.[6]

If f is a mapping, its inverse, f^{-1}, is always a certain *relation* (namely, the set of all ordered pairs inverse to those contained in f). However, this relation may fail to be a mapping. For example, let

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 3 & 8 \end{pmatrix}; \quad \text{then} \quad f^{-1} = \begin{pmatrix} 2 & 3 & 3 & 8 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Here f is a mapping (see Example (3)), but f^{-1} is not, because $f^{-1}[3] = \{2, 3\}$ consists of *two* elements (not of one). On the other hand, as is easily seen, the mappings given in Examples (1), (4), and (5) yields inverse relations that are mappings likewise. This justifies the following definition.

Definition 2.

A mapping f is said to be *invertible* iff its inverse, f^{-1}, is a map itself. In this case f^{-1} is called the *inverse map* or *inverse function*.

Equivalently,

a mapping (function) is invertible iff it is one-to-one.

[4] As we have noted, such a definition suffices provided that the domain of the function is known.

[5] Note, however, that we may also regard it as a map of J *onto* the smaller set E of all *even* integers: $f\colon J \underset{\text{onto}}{\leftrightarrow} E$.

[6] We may also consider the relation $\{(x, x) \mid x \in A\}$, denoted I_A, where A is a proper *subset* of the given space S. Then $I_A\colon A \to S$ is one-to-one but not *onto* S (it is onto A only). I_A is called the *identity map on* A.

For, if f is one-to-one, then no distinct elements of its domain can have one and the same function value y. But this very fact means that $f^{-1}[y]$ cannot consist of more than one element, i.e., that f^{-1} is a *function*.

The function value $f(x)$ is also sometimes denoted by fx, xf, or f_x. In the latter case (called "*index notation*"), the domain of f is also referred to as an *index set*, and the range of f is denoted by $\{f_x\}$. It is convenient to regard x in such symbols as a *variable* ranging over the domain of f (index set). Then also the function value $f(x)$ (respectively, fx, xf, or f_x) becomes a variable depending on x; we call it then the *variable function value*. If, in particular, D_f and D'_f are sets of real numbers, we obtain what is called *a real-valued function of a real variable*. Such functions are considered in the elementary calculus. Our function concept is, however, much more general since we consider maps with *arbitrary* domains and ranges (not necessarily sets of numbers).

Note 1. We shall strictly distinguish between the *function value* $f(x)$ and the *function* f itself. The latter is a set of ordered pairs while the former, $f(x)$, is only a single (though possibly variable) element of the range of values of f. These two notions are very often confused in elementary calculus, e.g., in such expressions as "the function $f(x) = 2x$." What is actually meant is "the function f defined by the formula $f(x) = 2x$." Another correct way of expressing this is by saying that "*f is the function that carries (or transforms) each $x \in D_f$ into* $2x$" or, briefly, that "*f is the map $x \to 2x$*" or "*f assigns to x the value* $2x$," etc.

Note 2. Mappings are also often referred to as *transformations*.

Note 3. If index notation is used, the range of function values D'_f, also written as $\{f_x\}$, can be regarded as a certain set of objects $\{f_x\}$ that are distinguished from each other by the various values of the variable index x. We have already encountered this notation in §2, with respect to families of sets.

As we have already mentioned, the domain and range of a function f may be quite arbitrary sets.[7] In particular, we can consider functions in which each element of the domain is itself an ordered pair, (x, y). Such mappings are called *functions of two variables*. Similarly, we speak of a *function of n variables* if the domain D_f of that function is a set of ordered n-tuples. To any such n-tuple, $(x_1, x_2, \ldots, x_n)$, the function f assigns a unique function value, denoted by $f(x_1, x_2, \ldots, x_n)$, provided that the n-tuple belongs to D_f. Note that each n-tuple $(x_1, \ldots, x_n)$ is treated as *one* element of D_f and is assigned only *one* function value. Usually (but not always) the domain D_f consists of all n-tuples that can be formed from elements of a given set A; that is, $D_f = A^n$ (the Cartesian product of n sets, each equal to A). The range may be any arbitrary set. The formula $f\colon A^n \to B$ is used to denote such a

[7] These sets may even be *empty*. Then also $f = \emptyset$ ("an empty set of ordered pairs"). Thus $\emptyset$ is a mapping, with $D_f = D'_f = \emptyset$.

function. Similarly, we write $f\colon (A \times B) \to C$ for a function of two variables, with $D_f = A \times B$ and $D'_f \subseteq C$, etc.

Note 4. Functions of two variables are also called (binary) *operations*. When this terminology is used, we usually replace the function symbols f, g, F, ... by special symbols $+$, $\cdot$, $\cup$, $\cap$, etc., and write $x + y$, $x \cdot y$, etc., instead of $f(x, y)$. The function value $f(x, y)$ then is called the *sum* (*product*, *composite*, etc.) of x and y.

Problems on Mappings

1. Which of the following relations, or their inverses, are mappings?

$\{(x, y) \mid y \text{ is the mother of } x\}$; $\{(x, y) \mid x \text{ is the father of } y\}$;
$\{(x, y) \mid y \text{ is a child of } x\}$; $\{(x, y) \mid x \text{ is a friend of } y\}$;
$\{(x, y) \mid y \text{ is the oldest son of } x\}$; $\{(x, y) \mid x \text{ is the oldest cousin of } y\}$;
$\{(x, y) \mid x \text{ real}, y = x^2\}$; $\{(x, y) \mid y \text{ real}, x = y^3\}$.

2. Are there any mappings among the relations specified in Problems 1 and 2 of §4? Which, if any, are one-to-one? Why or why not?

3. Let $f\colon N \to N$, where N is the set of all positive integers (naturals). Specify $f[N]$ (i.e., D'_f) and determine whether f is one-to-one and onto given that, for all $x \in N$,

(i) $f(x) = |x| + 2$; (ii) $f(x) = x^3$; (ii) $f(x) = 4x + 5$;
(iv) $f(x) = x^2$; (v) $f(x) = 1$;
(vi) $f(x)$ is the greatest common divisor of x and 15.

4. Do Problem 3 assuming that N is the set of *all* integers. Do cases (i)–(v) also with $N =$ set of all real numbers.

5. In Problems 3 and 4, find (in all cases) $f^{-1}[A]$ and $f[A]$ given that

(a) $A = \{x \in N \mid x \geq 0\}$; (b) $A = \{x \in N \mid -1 \leq x \leq 0\}$;
(c) $A = \{x \in N \mid -1 \leq x \leq 4\}$.

6. Prove that, for any mapping f, any set A, and any x, we have $x \in f^{-1}[A]$ iff $x \in D_f$ and $f(x) \in A$.

7. Using the result of Problem 6, prove for any *mapping* f that

(i) $f^{-1}[A \cup B] = f^{-1}[A] \cup f^{-1}[B]$;
(ii) $f^{-1}[A \cap B] = f^{-1}[A] \cap f^{-1}[B]$;
(iii) $f^{-1}[A - B] = f^{-1}[A] - f^{-1}[B]$.

Compare this with Problem 4 of §4. In what case do these formulas hold with "f^{-1}" replaced by "f"? In what case are they true for *both* f and f^{-1}?

8. Generalize formulas (i) and (ii) of Problem 7 by proving them with A, B replaced by an arbitrary family of sets, $\{A_i\}$; i.e., prove that

$$\text{(i) } f^{-1}\left[\bigcup A_i\right] = \bigcup f^{-1}[A_i]; \quad \text{(ii) } f^{-1}\left[\bigcap A_i\right] = \bigcap f^{-1}[A_i].$$

9. If f is a mapping, show that $f[f^{-1}[A]] \subseteq A$ and that if $A \subseteq D'_f$, then $f[f^{-1}[A]] = A$. In what case do we have $f^{-1}[f[A]] = A$? Give a proof.

10. Which (if any) of the relations $\subseteq$ and $\supseteq$ holds between the sets $f[A] \cap B$ and $f[A \cap f^{-1}[B]]$? Give a proof.

11. The *characteristic* function C_A of a set A in a space S is defined on S by setting $C_A(x) = 1$ if $x \in A$, and $C_A(x) = 0$ if $x \notin A$. Given $A \subseteq S$, $B \subseteq S$, prove the following:

(i) If $A \subseteq B$, then $C_{B-A}(x) = C_B(x) - C_A(x)$ for $x \in S$. (Briefly: $C_{B-A} = C_B - C_A$.)

(ii) With a similar notation, we have $C_{A\cap B} = C_A \cdot C_B$, and if $A \cap B = \emptyset$, then $C_{A\cup B} = C_A + C_B$.

(iii) $C_{A\cup B} = \max(C_A, C_B)$, the larger of C_A and C_B.

(iv) $C_A + C_B = C_{A\cup B} + C_{A\cap B}$.

(v) $A \subseteq B$ iff $C_A \leq C_B$.

(vi) $A = B$ iff $C_A = C_B$.

12. Use Problem 11(vi) to give another proof of the set identities specified in the following problems of §2: 1, 2, 3, 8, 9, 14, 15.
[Hint: Use the results of Problem 11 to show that the characteristic functions of the left and right sides of the required identities coincide.]

***13.** An *ordered triple* (x, y, z) can be defined as an ordered pair $((x, y), z)$ in which the first coordinate is itself an ordered pair, (x, y). Accordingly, every function f of two variables is a set of ordered triples $((x, y), z)$ in which the pairs (x, y) form the domain D_f of f; and, for each such pair, $z = f(x, y)$, so that z is *uniquely* determined by (x, y). Is every set T of ordered triples a function of two variables? If not, what condition must T satisfy? Give a proof.
[Hint: T must not contain two *different* triples (x, y, z) and (x', y', z') with $x = x'$ and $y = y'$.]

***14.** Using Problem 13, investigate which of the following sets of ordered triples are functions of two variables. If they are, specify the function

value $f(x,y)$, as well as D_f and D'_f. (Below, x, y, and z denote real numbers.)

(i) $f=\{(x,y,z)\mid x<y<z\}$; (ii) $f=\{(x,y,z)\mid x<y=z\}$;
(iii) $f=\{(x,y,z)\mid x=y+z\}$; (iv) $f=\{(x,y,z)\mid z=xy\}$;
(v) $f=\{(x,y,z)\mid z=1\}$; (vi) $f=\{(x,y,z)\mid x^2+y^2=z^2\}$.

15. Let N be the set of all positive integers. Define a function of two variables $f\colon (N\times N)\to N$ by setting, for x, $y\in N$,

$$f(x,y)=\frac{1}{2}(x+y-1)\cdot(x+y)+(1-y).$$

Verify whether this function is one-to-one and onto N.

*§6. Composition of Relations and Mappings[1]

A relation R can be treated as a mechanism that transforms any given set A into its image $R[A]$. If S is another relation, we can apply it to the set $R[A]$ to obtain its image under S, i.e., $S[R[A]]$. Given a third relation T, we can apply it to the set $S[R[A]]$ to obtain its image, $T\,[S[R[A]]]$, and so on. This process of successively applying several relations leads to the important notion of *composition* of relations. Before defining this notion, it is useful to prove the following lemma.

Lemma. *Two relations R and S are equal iff $R[x]=S[x]$ for every element x.*

Proof. Recall that R and S are equal iff they consist of exactly the same ordered pairs, that is, iff $(x,y)\in R \iff (x,y)\in S$, for all x, y. But, as was shown in §4, this can also be written as

$$y\in R[x] \iff y\in S[x] \text{ for all } x \text{ and } y.$$

Fixing x, we see from this that, whenever some element y belongs to the set $R[x]$, it also belongs to $S[x]$, and vice versa. In other words, the two sets $R[x]$ and $S[x]$ consist of the same elements. Thus we have

$$R[x]=S[x] \text{ for every } x,$$

as required. The converse is obtained by reversing the steps of the proof. Thus the lemma is proved. □

[1] This and other "starred" sections may be omitted in the first reading of Chapter 1. Indeed, the beginner is advised to postpone them, pending further directives.

This lemma shows that a relation R is uniquely determined if the sets $R[x]$ are given for all x. (Indeed, if any relation has the same image sets, it must coincide with R, by the lemma.) Therefore, *a relation can be defined by indicating the sets $R[x]$ for all x.*[2] We shall now apply this method to define the notion of the composite relation.

Definition.

By the *composite* of two relations R and S, denoted $R\circ S$ or RS, we mean the relation with images defined by

$$(R\circ S)[x]=R[S[x]] \text{ for every } x. \tag{1}$$

In other words, the image of any element x under the composite relation $R\circ S$ is obtained by first taking its image under S, i.e., $S[x]$, and then taking the image of the set $S[x]$ under R. Thus all images under $R\circ S$ are well defined; hence so is $R\circ S$. Note that formula (1) defines implicitly also the domain of $R\circ S$; it consists of those x whose images under $R\circ S$ are nonvoid.

Example.

Let

$$R=\begin{pmatrix}1&2&3\\2&3&4\end{pmatrix},\quad S=\begin{pmatrix}2&3\\1&5\end{pmatrix}.$$

Then RS consists of the pair $(2,2)$ alone, while SR consists of $(1,1)$ and $(2,5)$. This example shows that $RS\neq SR$; that is, *the composition of relations is, in general, not commutative* (even when they are mappings, as in this example). It is, however, associative, as is shown next.

Theorem 1. *For any relations R, S, T, we have $(RS)T=R(ST)$.*

Proof. By the lemma, it suffices to show that $((RS)T)[x]=(R(ST))[x]$, for every x. But, by definition (see formula (1) above), we obtain

$$((RS)T)\,[x]=(RS)\,[T[x]]=R\,[S\,[T[x]]]\,.$$

Similarly, $(R(ST))\,[x]=R\,[S\,[T[x]]]$. Thus the images coincide, as required, and all is proved. □

Theorem 2. *For any relations R and S, we have $(RS)^{-1}=S^{-1}R^{-1}$.*

The proof is left as an exercise (see Problem 4 below).

Theorem 3. *If R and S are functions, so also is RS. In particular, if R and S are* one-to-one *mappings, so is RS.*

[2] This is analogous to defining a function R by indicating $R(x)$ for all $x\in D_R$. In the present case, however, it is unnecessary to specify D_R because $R[x]$, unlike $R(x)$, is defined for *all* x.

Proof. Formula (1) above shows that $(RS)[x]$ contains *at most one* element if $R[S[x]]$ does, and this is clearly the case when S and R are *mappings*. The second clause likewise follows easily from Theorem 2. $\square$

Problems on the Composition of Relations

1. Find $(RS)T$, $R(ST)$, $(RT)S$, and $R(TS)$ by actual computation, if

$$R = \begin{pmatrix} 1 & 1 & 2 & 3 \\ 3 & 2 & 4 & 4 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 2 & 2 & 5 \\ 2 & 2 & 1 & 3 \end{pmatrix}, \quad T = \begin{pmatrix} 4 & 3 & 5 & 6 & 7 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}.$$

Comment on associativity and commutativity in these examples.

2. For any relation R and any positive integer n, define $R^n = R \circ R \circ \cdots \circ R$ (n times). Using the relations R, S, T of Problem 1, find the following:

(i) $R^3 \circ (R^{-1})^3$; (ii) $R^2 \circ (S^{-1})^2 \circ T$; (iii) $T^2 S^2 R^{-1}$.

Also, setting $R^{-n} = (R^{-1})^n$, find the following:

(iv) $R^{-2} S^2 T^{-1}$; (v) $S^{-3} T R^{-2}$.

3. Prove that $R \circ S = \{(x, y) \mid (\exists z)\ xSz,\ zRy\} = \{(x, y) \mid y \in R[S[x]]\}$.

4. Using the result of Problem 3, show that $(RS)^{-1} = S^{-1} \circ R^{-1}$. State and prove a similar formula for three relations and for n relations. Verify it also, by actual computation, for the three relations of Problem 1.

5. Which of the properties "reflexive", "symmetric", "transitive", and "trichotomic" does the relation R possess if $R \circ R \subseteq R$? Give a proof and compare with Problem 10 of §4.

6. Show that, for any relations R and S, $D_{RS} \subseteq D_S$ and $D'_{RS} \subseteq D'_R$. If, further, $D'_S \subseteq D_R$, then $D_{RS} = D_S$. (Use Problem 3.)

7. Show that, for every mapping $f\colon A \underset{\text{onto}}{\to} B$, we have $f \circ f^{-1} = I_B$, where $I_B = \{(y, y) \mid y \in B\}$ (= identity map on B); if, instead, $f\colon A \to B$ is one-to-one, we have $f^{-1} \circ f = I_A = \{(x, x) \mid x \in A\}$ (= identity map on A). Show by counterexamples that the second formula may fail if f is not one-to-one, and the first may fail if f is not onto B.

8. Let $\mathcal{T}$ be the family of all one-to-one maps of a set A onto itself. Prove the following:

(i) If f, $g \in \mathcal{T}$, then $f \circ g \in \mathcal{T}$.

(ii) If $f \in \mathcal{T}$, then $f^{-1} \in \mathcal{T}$, and $f \circ f^{-1} = f^{-1} \circ f = I_A$ (= identity map on A).

(iii) If $f \in \mathcal{T}$, then $f \circ I_A = I_A \circ f = f$. Note: By Theorem 1, we also have $(f \circ g) \circ h = f \circ (g \circ h)$ for all f, g, $h \in \mathcal{T}$. (A reader familiar with group theory will infer from all this that $\mathcal{T}$ is a *group*.)

9. Define a map of the xy-plane into itself by

$$f(x, y) = (x \cdot \cos\theta - y \cdot \sin\theta, x \cdot \sin\theta + y \cdot \cos\theta) \quad \text{(rotation)}.$$

Show that f is one-to-one and onto, and give a similar formula for the mapping $f^{-1} \circ g \circ f$, where (i) $g(x, y) = (x + 1, y)$, (ii) $g(x, y) = (x + 1, y + 1)$. Interpret geometrically.

10. Prove that a mapping $f\colon A \to B$ is one-to-one iff there is a map $g\colon B \to A$ with $g \circ f = I_A$.
[Hint: If f is one-to-one, fix some $a \in A$. Then define $g(y) = f^{-1}(y)$ if $y \in D'_f$, and $g(y) = a$ if $y \notin D'_f$.]

11. Prove that a mapping $f\colon A \to B$ is onto B if there is a map $h\colon B \to A$ such that $f \circ h = I_B$ (= identity map on B). Combining this with Problem 10, infer that f is one-to-one and onto if there are two maps g, $h\colon B \to A$ such that $g \circ f = I_A$ and $f \circ h = I_B$.
[Hint: If $f \circ h = I_B$, choose any $b \in B$ and find some $x \in A$ such that $f(x) = b$. (It suffices, e.g., to take $x = h(b)$. Why?)]

12. Prove the following:

(i) $f\colon A \to B$ is one-to-one iff $f \circ g = f \circ h$ implies $g = h$ for all maps g, $h\colon B \to A$.

(ii) If A has at least two elements, then $f\colon A \to B$ is onto B iff $g \circ f = h \circ f$ implies $g = h$ for all maps g, $h\colon B \to A$.

[Hint for part (ii): If f is not onto B, fix some $x_0, x_1 \in A$ ($x_0 \neq x_1$) and define two maps g, $h\colon B \to A$, setting: $(\forall y \in B)\ g(y) = x_0$; and $h(y) = x_0 = g(y)$ if $y \in D'_f$, while $h(y) = x_1$ if $y \notin D'_f$. Verify that $g \circ f = h \circ f$, though $g \neq h$. Thus $g \circ f = h \circ f$ *does not* imply $g = h$ if f is not onto.]

13. An equilateral triangle ABC (see Figure 7) is carried into itself by these rigid motions: clockwise rotations about its center through 0°, 120°, and 240° (call them r_0, r_1, r_2) and reflections in its altitudes AA', BB', CC' (call these reflections h_a, h_b, h_c, respectively). Treat these motions as mappings of the triangle onto itself, and set up for them a composition table (i.e., compute their mutual composites). Thus verify that the composite of any two of them is such a map itself; e.g., $r_1 r_2 = r_0$ (= the identity map); $r_1 h_a = h_c$; $h_a r_1 = h_b$, etc. (Note that $h_a r_1$ is the result of carrying out *first* the rotation r_1 and *then* the reflection h_a.) The maps r_0, r_1, r_2, h_a, h_b, h_c are called the *symmetries* of the triangle.

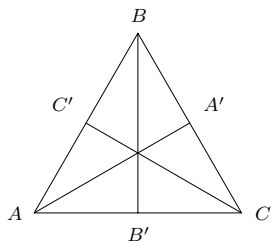


FIGURE 7

14. Set up and solve problems similar to 13 for

(i) the symmetries of the square (4 rotations and 4 reflections);

(ii) the symmetries of the rectangle (2 rotations and 2 reflections);

(iii) the symmetries of the regular pentagon (5 rotations and 5 reflections).

*§7. Equivalence Relations

In mathematics, as in everyday life, it is often convenient not to distinguish between certain objects that, however different, serve the same purpose and thus may be "identified" (i.e., regarded as the same) as far as this purpose is concerned. For example, different coins and bills of the same value may be regarded as equivalent in all money transactions. Parallel lines may be treated as the same in all angle measurements. Congruent figures may be "identified" in geometry. In all such cases some relation (like parallelism or congruence) plays the same role as equality. Such relations, called *equivalence relations*, resemble equality in that they are reflexive, symmetric and transitive. Usually they also have, to a certain degree, the so-called *substitution property*; that is, within certain limits, equivalent objects may be substituted for each other. We now give precise definitions.

Definition 1.

A binary relation E is called an *equivalence relation* on a set A if E is reflexive, symmetric, and transitive on A and moreover its domain D_E and its range D'_E coincide with A.[1]

Equivalence relations are usually denoted by special symbols resembling equality, such as $\equiv$, $\approx$, $\sim$, etc. The formula $(x, y) \in E$ or xEy, where E is such a symbol, is read "x is equivalent to y," "x is congruent with y," etc. Sometimes the phrase "modulo E" is added. Thus we write $x \equiv y$, or $x \equiv y \pmod{E}$, for xEy. If such a formula is true, we say that x and y are *E-equivalent*, or *equivalent modulo E*, or, briefly, *equivalent*.

Definition 2.

An equivalence relation E is said to have the *substitution property* with respect to another relation R if xRy implies $x'Ry'$ whenever $x \equiv x' \pmod{E}$ and $y \equiv y' \pmod{E}$. In this case we also say that E is *consistent* with R. In other words, consistency means that *the formula xRy does not alter its validity or nonvalidity if x and y are replaced by some equivalent elements*, $x' \equiv x$, and $y' \equiv y$.

Similarly, we say that E is *consistent* with an operation $\circ$ in a set A, or that E has the *substitution property* with respect to $\circ$, if $x \circ y \equiv x' \circ y'$ whenever $x, x', y, y' \in A$, $x' \equiv x$, and $y' \equiv y \pmod{E}$.

[1] Note that the domain D_E of E *must* coincide with its range D'_E due to symmetry. Explain!

The equality relation (i.e., the identity map on a set A) is itself an equivalence relation since it is reflexive, symmetric, and transitive. It has the (unlimited) substitution property since we have defined it as *logical identity*. Other examples (such as parallelism of lines, or congruence of figures) have been mentioned above; see also the problems below.

Definition 3.

If E is an equivalence relation on A, and if $p \in D_E$, we define the *E-class, or equivalence class modulo E, generated by p in A* to be the set of all those elements of A that are E-equivalent to p. Thus it is the set

$$\{y \in A \mid pEy\} = E[p] \quad (= \text{image of } p \text{ under } E).$$

If confusion is ruled out, we denote it simply by $[p]$ and call it the *E-class of p* (in A); p is called a *generator* or *representative* of $[p]$.[2] The family of all E-classes, generated in A by different elements, is called the *quotient set of A by E*, denoted A/E. **Note:** By definition, $x \in [p]$ *iff* $x \equiv p$.

Examples.

(a) If $E = I_A$ (the identity map on A), then $E[x] = [p] = \{p\}$ for each $p \in A$. Thus here each E-class consists of a *single* element (its generator).

(b) Under the parallelism relation between straight lines, an equivalence class consists of all lines parallel to a given line in space.

(c) Under congruence, an equivalence class consists of all figures congruent to a given figure.

Theorem 1. *If E is an equivalence relation on a set A, then we have the following*:

(i) *Every element $p \in A$ is in some E-class*; *specifically*, $p \in [p] \subseteq A$.

(ii) *Two elements $p, q \in A$ are E-equivalent iff they are in one and the same equivalence class*, *i.e.*, *iff* $[p] = [q]$.

(iii) *Any two E-classes in Q are either identical or disjoint*.

(iv) *The set A is the (disjoint) union of all E-classes*.

[2] As we shall see (Theorem 1(ii) below), any other element $q \equiv p$ is likewise a generator of $[p]$ because the E-classes generated by p and q *coincide* if $q \equiv p$ (i.e., $q \in [p]$).

Proof. (i) By definition, $x \in [p]$ iff $x \equiv p$. Now, if $p \in A$, reflexivity of E yields $p \equiv p$, whence $p \in [p] \subseteq A$, as asserted.

(ii) If $p \equiv q$, then, by symmetry and transitivity, $(\forall x \in A)\ p \equiv x$ iff $q \equiv x$. This means that $x \in [p]$ iff $x \in [q]$, i.e., $[p] = [q]$. Conversely, if $[p] = [q]$, then part (i) yields $q \in [q] = [p]$, i.e., $q \in [p]$, when $p \equiv q$, by the definition of $[p]$.

(iii) Suppose $[p] \cap [q] \neq \emptyset$, i.e., $(\exists x)\ x \in [p] \cap [q]$. Then $x \in [p]$ and $x \in [q]$, i.e., $x \equiv p \equiv q$, whence, by (ii), $[p] = [q]$. Thus $[p]$ and $[q]$ cannot have a common element unless $[p] = [q]$.

(iv) is a direct consequence of (i) and (iii). Thus all is proved. □

Part (iv) of this theorem shows that *every equivalence relation E on A defines a partition of A into E-classes*. The converse is likewise true, as we show next.

Theorem 2. *Every partition of a set A into disjoint sets A_i $(i \in I)$ uniquely determines an equivalence relation E on A, such that the sets A_i are exactly the E-classes in A.*

Proof. Given $A = \bigcup A_i$ (disj.),[3] define E as the set of all pairs (x, y) such that x and y belong to *one and the same* A_i. The relation E is easily shown to be reflexive, symmetric and transitive on A, with $D_E = D'_E = A$, so that E is an equivalence relation in A (we leave the details to the reader). Moreover, the E-classes clearly coincide with the sets A_i. Thus E has all the required properties.

Next, let E' be another equivalence relation on A, with the same properties, and take any $p \in A$. Then, by assumption, $E[p] = A_i$, where A_i is the partition set that contains p; also, $E'[p] = A_i$ for the same i. It follows that $(\forall p)$ $E[p] = E'[p]$, and this implies that $E = E'$ (by the lemma of §6). Thus any two such E and E' must coincide, i.e., E is unique. □

We see that there is a close connection between all equivalence relations on A and all partitions of A: Every equivalence relation defines (or, as we shall say, *induces*) a partition, and vice versa. Note that the quotient set A/E is exactly the family of the disjoint sets A_i whose union equals A, i.e., the family of the disjoint equivalence classes, under the equivalence relation E that corresponds to a given partition.

Now we can give a more exact mathematical interpretation to the procedure of "identifying" equivalent elements (see introductory remarks to this section). This procedure applies whenever an equivalence relation E *is consistent with some operation or relation R*, so that the substitution property holds with respect to R. Then, as far as R is concerned, equivalent elements behave *as if* they were identical, so that they may be treated as "copies" of one element.

[3] We use this notation to indicate that A is the union of *disjoint* sets $A_i (i \in I)$.

We achieve *actual* identity if we replace each element p of the set A by the equivalence class $[p]$ generated by p. Indeed, then, all E-equivalent elements are replaced by *one and the same* equivalence class and thus become *one* thing. Thus, *from the mathematical point of view, the "identification" of equivalent elements amounts to replacing the set A by the quotient set A/E.* In what follows, we shall often speak of "identifying" certain objects. The reader should, however, be aware of the fact that what is meant is actually the procedure outlined here, i.e., the replacement of A by the quotient set A/E.

Problems on Equivalence Relations

1. Prove in detail that the relation E defined in the proof of Theorem 2 is reflexive, symmetric, and transitive on A and that $D_E = D'_E = A$.

2. Which of the following relations on the set J of all integers are equivalence relations? If so, describe the E-classes, i.e., J/E.

(i) $E = \{(x, y) \mid x, y \in J;$ and $x - y$ is divisible by a fixed $n \in J\}$;

(ii) $E = \{(x, y) \mid x, y \in J;\ x - y$ is odd$\}$;

iii) $E = \{(x, y) \mid x, y \in J;$ and $x - y$ is a prime$\}$.

3. Are the equivalence relations of Problem 2 consistent with the addition, multiplication, and inequality relation $(<)$ defined in J?

Problems 4–10 are of theoretical importance for the construction of the rational number system from natural numbers (including 0), i.e., nonnegative integers.

4. Let N be the set of all integers ≥ 0, so that $N \times N$ is the set of all ordered *pairs* of nonnegative integers. Assuming the arithmetic of such integers to be known, let $(x, y)E(p, q)$ mean that $x + q = y + p$, and let $(x, y) < (p, q)$ mean that $x + q < y + p$, where $x,\ y,\ p,\ q \in N$. Without ever using subtraction or minus signs, show that E is an equivalence relation on $N \times N$, consistent with $<$. (Write $\equiv$ for E.) Also show that the relation $<$ is transitive and "*quasi-trichotomic*"; i.e., we have either

$$(x, y) < (p, q) \text{ or } (p, q) < (x, y) \text{ or } (x, y) \equiv (p, q),$$

but never two of these together.

5. Continuing Problem 4, define addition and multiplication in $N \times N$, setting

$$(x, y) + (p, q) = (x + p, y + q)$$

and

$$(x, y) \cdot (p, q) = (xp + yq, yp + xq).$$

Show that E is consistent with these operations. Also verify the following laws:

(i) If (x, y) and (p, q) belong to $N \times N$, so do their sum and product.

(ii) $(x,y)+(p,q)\equiv(p,q)+(x,y)$; $(x,y)\cdot(p,q)\equiv(p,q)\cdot(x,y)$.

(iii) $\{(x,y)+(p,q)\}+(r,s)\equiv(x,y)+\{(p,q)+(r,s)\}$, and similarly for multiplication:

$$\{(x,y)\cdot(p,q)\}\cdot(r,s)\equiv(x,y)\cdot\{(p,q)\cdot(r,s)\}.$$

(iv) $(x,y)+(0,0)\equiv(x,y)$; $(x,y)\cdot(1,0)\equiv(x,y)$.

(v) $(x,y)+(y,x)\equiv(0,0)$. (Hence we may write $-(x,y)$ for (y,x).)

(vi) $(x,y)\cdot\{(p,q)+(r,s)\}\equiv(x,y)\cdot(p,q)+(x,y)\cdot(r,s)$.

(vii) If $(p,q)<(r,s)$ then $(p,q)+(x,y)<(r,s)+(x,y)$. Similarly for multiplication, provided, however, that $(0,0)<(x,y)$.

Observe that $(x,y)<(0,0)$ iff $x<y$ (verify!); we call the pair (x,y) "negative" in this case. Show that $(x,y)<(0,0)$ iff $-(x,y)>(0,0)$.

6. The laws proved in Problems 4 and 5 show that ordered pairs (x,y) in $N\times N$, with inequalities and operations defined as above, "behave" like *integers* (positive, negative and 0) *except that equality* "$=$" *is replaced by* "$\equiv$". To avoid the latter we pass to equivalence classes. Let $[x,y]$ denote the E-class of the pair (x,y). Define addition and multiplication of such E-classes by

$$[x,y]+[p,q]=[x+p,y+q],\quad [x,y]\cdot[p,q]=[xp+yq,yp+xq].$$

Using the consistency of E (proved in Problem 5), show that these definitions are *nonambiguous*; i.e., the sum and product remain the same also when x, y, p, q are replaced by some x', y', p', q' such that $(x,y)\equiv(x',y')$ and $(p,q)\equiv(p',q')$. Then show that the laws (ii)–(vi) are valid for E-classes of the pairs involved, *with all equivalence signs* "$\equiv$" *turning into* "$=$".

7. Continuing Problems 4–6, define $[x,y]<[p,q]$ to mean that $(x,y)<(p,q)$, as in Problem 4. Show that this is unambiguous, i.e., the inequality holds also if (x,y) or (p,q) is replaced by an equivalent pair. Verify that Problem 5(vii), as well as the transitivity and "trichotomy" laws of Problem 4, hold for E-classes, with "$\equiv$" replaced by "$=$". (We now *define* "integers" to be the equivalence classes $[x,y]$.)

8. Let J be the set of *all* integers (positive or not), and let Q be the set of all ordered pairs $(x,y)\in J\times J$, with $y>0$. Assuming the arithmetic of integers to be known, let $(x,y)E(p,q)$ mean that $xq=yp$, and let $(x,y)<(p,q)$ mean that $xq<yp$, for (x,y) and (p,q) in Q. Without using division or fraction signs, answer the questions of Problem 4, with $N\times N$ replaced by Q. (Subtraction and minus signs are now permitted.)

9. In Problem 8, show that E is consistent with addition and multiplication defined in Q as follows:

$$(x,y)+(p,q)=(xq+yp,qy)\text{ and }(x,y)\cdot(p,q)=(xp,yq).$$

For such sums and products, establish the laws of Problem 5, with (iv) and (v) replaced by

(iv′) $(x,y)+(0,1)\equiv(x,y)\equiv(x,y)\cdot(1,1)$;

(v′) $(x,y)+(-x,y)\equiv(0,1)$;

(v″) if $x>0$, then $(x,y)\cdot(y,x)\equiv(1,1)$; and

(v‴) if $x<0$, then $(x,y)\cdot(-y,-x)\equiv(1,1)$.

Observe that pairs $(x,y)\in Q$ behave like fractions x/y in ordinary arithmetic (with "$=$" replaced by "$\equiv$" here).

10. Continuing Problems 8 and 9, let $[x,y]$ denote the E-class of the pair $(x,y)\in Q$, with E as in Problem 8. For such E-classes, define inequalities, addition and multiplication as for pairs in Problems 8 and 9, replacing (x,y) by $[x,y]$. Verify that these definitions are unambiguous, i.e., independent of the particular choice of the "representative pairs" (x,y) and (p,q) from the E-classes $[x,y]$ and $[p,q]$ (use the consistency properties of E). Verify that all laws proved in Problems 8 and 9 hold also for E-classes (with "$\equiv$" now turning into "$=$").

Note. Problems 4–10 show how, starting with *nonnegative* integers, one can construct first a system $N\times N$ and then a system Q that (on passage to suitable equivalence classes) behave exactly like integers and rational numbers, respectively. This is how integers and rationals are *constructed* from nonnegative integers, in precise mathematics.

11. A reader acquainted with group theory will verify that, if A is a group, and B its subgroup, then each of the following relations is an equivalence relation on A (we use multiplicative notation):

(i) $E=\{(x,y)\mid x,\ y\in A,\ x^{-1}y\in B\}$;

(ii) $E=\{(x,y)\mid x,\ y\in A,\ yx^{-1}\in B\}$.

Also show that if the group operation is commutative (i.e., $xy=yx$) then in both cases E is consistent with that operation.

§8. Sequences

One of the basic notions of analysis is that of a *sequence* (infinite or finite). It is closely connected with the theory of mappings and sets. Therefore we

consider it here, even though it involves the notion of *integers*, to be formally introduced in Chapter 2, along with real numbers.

Definition 1.

By an *infinite sequence* we mean a mapping (call it u) whose domain D_u consists of *all* positive integers 1, 2, 3, ... (it may also contain 0). A *finite sequence* is a mapping u in which D_u consists of positive (or nonnegative) integers less than some fixed integer p. The *range* D'_u may consist of *arbitrary* objects (numbers, points, lines, sets, books, etc.).

Note 1. In a wider sense, one may speak of "sequences" in which D_u also contains some negative integers, or excludes some positive integers. We shall not need this more general notion.

Note that a sequence, being a mapping, is a set of *ordered pairs*. For example,

$$u = \begin{pmatrix} 1 & 2 & 3 & \dots & n & \dots \\ 2 & 4 & 6 & \dots & 2n & \dots \end{pmatrix} \tag{1}$$

is an infinite sequence, with $D_u = \{1, 2, 3, \dots\}$; its range D'_u consists of the function's values

$$u(1) = 2,\ u(2) = 4,\ u(3) = 6,\ \dots,\ u(n) = 2n,\ n = 1, 2, \dots .$$

Instead of $u(n)$, we usually write u_n ("index notation"), and call u_n the *n-th term* of the sequence. If n is treated as a *variable*, u_n is called the *general term* of u, and $\{u_n\}$ is used to denote the entire sequence, as well as its range D'_u. The formula $\{u_n\} \subseteq B$ means that D'_u is contained in a set B; we then call u a *sequence of elements of* B, or a sequence *from* B, or *in* B. To uniquely determine a sequence u, it suffices to define its general term (by some formula or rule) for every $n \in D_u$. In Example (1) above, $u_n = 2n$.

Since the domain of a sequence is *known* to consist of integers, we often omit it and give only the range D'_u, specifying the terms u_n in the order of their indices n. Thus, instead of (1), we briefly write 2, 4, 6, ..., $2n$, ... or, more generally, $u_1, u_2, \dots, u_n, \dots$, along with the still shorter notation $\{u_n\}$. Nevertheless, whatever the notation, the sequence u (a set of ordered *pairs*) should not be confused with D'_u (the set of single terms u_n).

A sequence need not be a *one-to-one* mapping; it may have equal ("repeating") terms: $u_m = u_n$ $(m \neq n)$. For instance, in the infinite sequence 1, 1, ..., 1, ..., with general term $u_n = 1$, all terms are equal to 1, so that its range D'_u has only one element, $D'_u = \{1\}$. Nevertheless, by Definition 1, the sequence itself is infinite. This becomes apparent if we write it out in full notation:

$$u = \begin{pmatrix} 1 & 2 & 3 & \dots & n & \dots \\ 1 & 1 & 1 & \dots & 1 & \dots \end{pmatrix}. \tag{2}$$

Indeed, it is now clear that D_u contains all positive integers 1, 2, 3, ..., and u itself contains *infinitely many* pairs $(n, 1)$, $n = 1, 2, \dots$, even though $D'_u = \{1\}$. Sequences in which all terms are equal are referred to as *constant* or *stationary*.

In sequences (1) and (2) we were able to define the general term by means of a formula: $u_n = 2n$ or $u_n = 1$. This is not always possible. For example, nobody has yet succeeded in finding a formula expressing the general term of the sequence

$$1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, \dots \tag{3}$$

of so-called *prime* numbers (i.e., integers with no positive divisors except 1 and themselves).[1] Nevertheless, this sequence is well defined since its terms can be obtained step by step: start with *all* positive integers, 1, 2, 3, ...; then remove from them all multiples of 2 except 2 itself; from the remaining set remove all multiples of 3 except 3 itself, etc., ad infinitum. After the first step, we are left with

$$1, 2, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, \dots;$$

after the second step, we obtain

$$1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 25, \dots,$$

and so on, gradually obtaining (3).

Other cases of such "step-by-step" definitions (also called *algorithmic* or *inductive* definitions) will occur in the later work. In general, a sequence is regarded as well defined if some formula (or formulas) or rule has been given that makes it possible to find all terms of the sequence, either directly or by some "step-by-step" or other procedure. One should carefully avoid the misconception that, if several terms in a sequence conform to some law or formula, then the same law applies to all the other terms. For instance, if only the first three terms of (3) were given, it would be wrong to conclude that the sequence is necessarily 1, 2, 3, 4, 5, ..., n, ..., with general term $u_n = n$. Thus an infinite set can never be defined by giving a finite number of terms only; in this case one can only make a "plausible" guess as to the intended general term. Moreover, one may well think of sequences in which the terms have been chosen "at random," without any particular law. Such a "law" may, but need not, exist.

As noted above, the terms of a sequence need not be *numbers*; they may be arbitrary objects. In particular, we shall often consider sequences of *sets*: $A_1, A_2, \dots, A_n, \dots$, where each term A_n is a set (treated as *one* thing). The following definitions will be useful in the later work.

[1] For our purposes it is convenient to include 1 in this sequence, though usually 1 is not regarded as a prime number.

Definition 2.

A sequence of sets $\{A_n\}$, $n = 1, 2, \ldots$, is said to be *expanding* iff each term A_n is a subset of the next term A_{n+1}, i.e.,

$$A_n \subseteq A_{n+1}, \quad n = 1, 2, \ldots$$

(except if A_n is the *last* term in a finite sequence).

The sequence $\{A_n\}$ is *contracting* iff

$$A_n \supseteq A_{n+1}, \quad n = 1, 2, \ldots$$

(with the same remark). In both cases, $\{A_n\}$ is called a *monotonic*, or *monotone*, set sequence.

This definition imitates a similar definition for number sequences:

Definition 3.

A sequence of real numbers $\{u_n\}$, $n = 1, 2, \ldots$, is said to be *monotonic* or *monotone* iff it is either *nondecreasing* (i.e., $u_n \leq u_{n+1}$) or *nonincreasing* (i.e., $u_n \geq u_{n+1}$) for all terms. Notation: $\{u_n\}\uparrow$ and $\{u_n\}\downarrow$.

If the *strict* inequalities, $u_n < n_{n+1}$ ($u_n > u_{n+1}$, respectively) hold, the sequence is said to be *strictly monotonic* (*increasing* or *decreasing*).

Note 2. Sometimes we say "*strictly increasing*" (or "*strictly decreasing*") in the latter case.

For example, the sequences (1) and (3) above are strictly increasing. Sequence (2) (and *any* constant number sequence) is monotonic, but not strictly so; it is both nondecreasing and nonincreasing. Any sequence of concentric discs in a plane, with increasing radii, is an expanding sequence of sets (we treat each disc as the set of all points inside its circumference). If the radii decrease, we obtain a contracting sequence.

By a *subsequence* of a sequence $\{u_n\}$ is meant (roughly speaking) any sequence obtained by dropping some terms from $\{u_n\}$, *without changing the order of the remaining terms*, which then form the subsequence. More precisely, to obtain a subsequence, we must prescribe the terms that are supposed to *remain* in it. This is best done by indicating the *subscripts* of these terms. Note that all such subscripts necessarily form an *increasing* sequence of integers:

$$n_1 < n_2 < n_3 < \cdots < n_k < \cdots .$$

If these subscripts are given, they uniquely determine the corresponding terms of the subsequence:

$$u_{n_1}, u_{n_2}, u_{n_3}, \ldots$$

with general term (or k-th term)

$$u_{n_k}, \quad k = 1, 2, \ldots .$$

The subsequence is briefly denoted by $\{u_{n_k}\}$; in special cases, also other notations are used. Thus we have the following.

Definition 4.

Let $\{u_n\}$ be any sequence, and let $\{n_k\}$ be a strictly increasing sequence of integers from D_u. Then the sequence $\{u_{n_k}\}$, with k-th term equal to u_{n_k}, is called the *subsequence* of $\{u_n\}$, determined by the sequence of subscripts $\{n_k\} \subseteq D_u$, $k = 1, 2, 3, \ldots$.

For example, let us select from (3) the subsequence of terms with subscripts

$$2, 4, 6, \ldots, 2k, \ldots$$

(i.e., consisting of the 2nd, 4th, 6th, …, $2k$-th, … terms of (3)). We obtain

$$2, 5, 11, 17, 23, 31, 41, \ldots .$$

If, instead, the terms

$$u_1, u_3, u_5, \ldots, u_{2k-1}, \ldots$$

were selected, we would obtain the subsequence

$$1, 3, 7, 19, 29, 37, \ldots .$$

The first subsequence could briefly be denoted by $\{u_{2k}\}$ (here $n_k = 2k$); the second subsequence is $\{u_{2k-1}\}$, $n_k = 2k - 1$, $k = 1, 2, \ldots$.

Observe that, in every sequence u, the integers belonging to its domain D_u are used to "number" the terms of u, i.e., the elements of the range D'_u; e.g., u_1 is the *first* term, u_2 the *second*, and so on. This procedure is actually well known from everyday life: by numbering the buildings in a street or the books in a library, we put them in a certain order or sequence. The question now arises: given a set A, is it always possible to "number" its elements by integers? More precisely, is there a sequence $\{u_n\}$, finite or infinite, such that A is contained in its range:

$$A \subseteq D'_u = \{u_1, u_2, \ldots, u_n, \ldots\}?$$

As we shall see later, this question must, in general, be answered in the negative; the set A may be so large that even *all* integers are too few to number its elements. At this stage we only formulate the following definition.

Definition 5.

A set A is said to be *countable* iff A is contained in the range of some sequence (briefly: "*A can be put in a sequence*").

If, in particular, this sequence can be chosen finite, we call A a *finite* set ($\emptyset$ is finite, since $\emptyset \subseteq D'_u$ *always*). Thus all finite sets are countable.

Sets that are not finite are called *infinite*.

Sets that are not countable are called *uncountable*.

A finite set A is said to have exactly n *elements* iff it is the range of a sequence of n *distinct* terms; i.e., the range of a *one-to-one* map u with domain $D_u = \{1, 2, \dots, n\}$. The simplest example of an infinite countable set is $N = \{1, 2, \dots\}$.

Problems on Sequences

1. Find the first six terms of the sequence of numbers with general term:

(a) $u_n = 2$; (b) $u_m = (-1)^m$;
(c) $u_n = n^2 - 1$; (d) $u_m = -m/(m+1)$.

2. Find a suitable formula, or formulas, for the general term of a sequence that starts with

(a) 2, 5, 10, 17, 26, ...; (b) 2, −2, 2, −2, 2, ...;
(c) 2, −2, −6, −10, −14, ...; (d) 1, 1, −1, −1, 1, 1, −1, −1, ...;
(e) $\frac{3\cdot 2}{1}, \frac{4\cdot 6}{4}, \frac{5\cdot 10}{9}, \frac{6\cdot 14}{16}, \dots$; (f) $\frac{1}{2\cdot 3}, \frac{-8}{3\cdot 4}, \frac{27}{4\cdot 5}, \frac{-64}{5\cdot 6}, \frac{125}{6\cdot 7}, \dots$.

3. Which of the sequences in Problems 1 and 2 are monotonic or constant? Which have finite ranges (even though the sequences are infinite)?

4. Find the general term of the sequence obtained from $\{u_n\}$ by dropping

(a) the first term;

(b) the first two terms;

(c) the first p terms.

5. (Lagrange interpolation formula.) Given the first p terms $a_1, \dots, a_p$ of a number sequence, let $f(n,k)$ be the product of the $p-1$ numbers

$$n-1,\ n-2,\ \dots,\ n-(k-1),\ n-(k+1),\ \dots,\ n-p$$

(*excluding* $n-k$), for $n = 1, 2, \dots$, and $k = 1, 2, \dots, p$.

Setting $b_k = f(k,k)$, verify that $b_k \neq 0$ and that, for $n = 1, 2, \dots, p$, we have

$$a_n = a_1 f(n,1)/b_1 + a_2 f(n,2)/b_2 + \cdots + a_p f(n,p)/b_p. \qquad (*)$$

Thus (*) is a suitable formula for the general term of the sequence. Using it, find new answers to Problem 2(a)–(d), thus showing that there are many "plausible" answers to the questions posed.

6. Find the general term u_n of the number sequence defined *inductively*[2] by

(i) $u_1 = a$, $u_{n+1} = u_n + d$, $n = 1, 2, \dots$ (arithmetic sequence; a, d fixed);

(ii) $u_1 = a$, $u_{n+1} = u_n q$, $n = 1, 2, \dots$ (geometric sequence; a, q fixed);

(iii) $s_1 = u_1$, $s_{n+1} = s_n + u_{n+1}$, with u_n as in case (i); same for (ii);

(iv) $u_1 = a$, $u_2 = b$, $u_{n+2} = \frac{1}{2}(u_{n+1} + u_n)$, $n = 1, 2, \dots$ (a, b fixed).
[Hint: $u_{n+2} = u_3 + (u_4 - u_3) + (u_5 - u_4) + \cdots + (u_{n+2} - u_{n+1})$, where $u_3 = \frac{1}{2}(a+b)$. Show that the bracketed terms $(u_{k+1} - u_k)$ form a geometric series with ratio $\frac{1}{2}$, and compute its sum.]

[2] Problems 6–8 may be postponed until induction and other properties of natural numbers have been studied in more detail (Chapter 2, §§5–6).

7. Show that if a number sequence $\{u_n\}$ has no *largest* term, then it has a strictly increasing infinite subsequence $\{u_{n_k}\}$.
[Hint: Define u_{n_k} step by step. Let $u_{n_1} = u_1$. Then let n_2 be the *least* subscript such that $u_{n_2} > u_{n_1}$ (why does such u_{n_2} exist?). Next take the least n_3 such that $u_{n_3} > u_{n_2}$, and so on.]

***8.** Let $\{u_n\}$ be an infinite sequence of real numbers. By dropping from it the first k terms, we get a subsequence $u_{k+1}, u_{k+2}, \dots, u_{k+n}, \dots$ (call it the "k-subsequence"). Show that if every k-subsequence ($k = 1, 2, 3, \dots$) has a largest term (call it q_k, for a given k), then the original sequence $\{u_n\}$ has a nonincreasing subsequence formed from all such q_k-terms.
[Hint: Show that $q_k \geq q_{k+1}$, $k = 1, 2, \dots$, i.e., the maximum term q_k cannot increase as the number k of the dropped terms increases. Note that $\{u_n\}$ may have *several* terms equal to q_k for a given k; choose the one with the least subscript inside the given k-subsequence.]

***9.** From Problems 7 and 8 infer that *every infinite sequence of real numbers* $\{u_n\}$ *has an infinite monotonic subsequence*.
[Hint: There are two possible cases:

(i) either every k-subsequence (as described in Problem 8) has a largest term, or

(ii) some k-subsequence has no largest term (then apply to it the result of Problem 7 to obtain an increasing subsequence of it and hence of $\{u_n\}$).]

10. How many finite sequences of p terms, i.e., with domain $\{1, 2, \dots, p\}$, can one form, given that the range of the sequences is a fixed set of m elements?

11. Let $\{A_n\}$ be an infinite sequence of sets. For each n, let

$$B_n = \bigcup_{k=1}^{n} A_k, \quad C_n = \bigcap_{k=1}^{n} A_k, \quad D_n = \bigcap_{k=n}^{\infty} A_k, \quad E_n = \bigcup_{k=n}^{\infty} A_k.$$

Show that the sequences $\{B_n\}$ and $\{D_n\}$ are expanding, while $\{C_n\}$ and $\{E_n\}$ are contracting.

***12.** Given a sequence of sets $\{A_n\}$, $n = 1, 2, \dots$, we define

$$\overline{\lim}\, A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \text{ and } \underline{\lim}\, A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

and call these sets the *upper limit* and the *lower limit* of the sequence $\{A_n\}$, respectively. If they coincide, the sequence is said to be *convergent*, and we put

$$\underline{\lim}\, A_n = \overline{\lim}\, A_n = \lim A_n \quad (= \text{limit of } A_n).$$

Prove the following:

(i) $\bigcap_{n=1}^{\infty} A_n \subseteq \underline{\lim}\, A_n \subseteq \overline{\lim}\, A_n \subseteq \bigcup_{n=1}^{\infty} A_n$.

(ii) If $A_n \subseteq B_n$, $n = 1, 2, \dots$, then

$$\underline{\lim}\, A_n \subseteq \underline{\lim}\, B_n \text{ and } \overline{\lim}\, A_n \subseteq \overline{\lim}\, B_n.$$

(iii) Every monotonic sequence of sets is convergent, with $\lim A_n = \bigcap_{n=1}^{\infty} A_n$ if $\{A_n\}$ is contracting, and $\lim A_n = \bigcup_{n=1}^{\infty} A_n$ if $\{A_n\}$ is expanding.

***13.** Continuing Problem 12, prove the following:

(i) $E - \underline{\lim}\, A_n = \overline{\lim}(E - A_n)$ and $E - \overline{\lim}\, A_n = \underline{\lim}(E - A_n)$ for any set E.

(ii) $\underline{\lim}(A_n \cap B_n) = \underline{\lim}\, A_n \cap \underline{\lim}\, B_n$ and $\overline{\lim}(A_n \cup B_n) = \overline{\lim}\, A_n \cup \overline{\lim}\, B_n$.

(iii) $\underline{\lim}(A_n \cup B_n) \supseteq \underline{\lim}\, A_n \cup \underline{\lim}\, B_n$ and $\overline{\lim}(A_n \cap B_n) \subseteq \overline{\lim}\, A_n \cap \overline{\lim}\, B_n$. Investigate whether inclusion signs can be replaced by equality if one or both sequences are convergent.

***14.** Continuing Problem 12, prove the following:

(i) If the sets A_n are mutually disjoint, then $\underline{\lim}\, A_n = \overline{\lim}\, A_n = \emptyset$.

(ii) If $A_n = A$ for all n, then $\underline{\lim}\, A_n = \overline{\lim}\, A_n = A$.

(iii) $\{A_n\}$ converges iff for no x are there infinitely many n with $x \in A_n$ and infinitely many n with $x \notin A_n$.

*§9. Some Theorems on Countable Sets

We now derive some consequences of Definition 5 of §8.

Theorem 1. *If a set A is countable or finite, so also is any subset $B \subseteq A$, and so is the image $f[A]$ of A under any mapping f.*

Proof. If $A \subseteq D'_u$ for a sequence u (finite or not), then certainly $B \subseteq A \subseteq D'_u$. Thus B can be put in *the same* sequence, proving our first assertion.

Next, let f be any map, and suppose first that $D_f \supseteq A$. We may assume that A *fills* a sequence (if not, drop some terms); say, $A = \{u_1, u_2, \dots, u_n, \dots\}$. Then $f[A]$ consists *exactly* of the function values $f(u_1), f(u_2), \dots, f(u_n), \dots$. But this very fact shows that $f[A]$ can be put in a sequence $\{v_n\}$, with general term $v_n = f(u_n)$. Thus $f[A]$ is countable (finite if A is), as claimed. The case $A \not\subseteq D_f$ is treated in Problem 1 below. Thus all is proved. □

Theorem 2. *If a set A is uncountable, so also is any set $B \supseteq A$, and so is $f[A]$ under any* one-to-one *map f, with $D_f \supseteq A$. (Similarly for infinite sets.)*

Proof. The set $B \supseteq A$ cannot be countable or finite. Otherwise, its subset A would have the same property, by Theorem 1, contrary to assumption.

Next, if f is one-to-one, so is its inverse, f^{-1}. If further $A \subseteq D_f$, then $A = f^{-1}[f[A]]$ by Problem 9 of §5. Now, if $f[A]$ were countable or finite then, by Theorem 1, so would be its image under *any* map, such as f^{-1}. Thus the set $f^{-1}[f[A]] = A$ would be countable or finite, contrary to assumption. □

Corollary 1. *If all terms of an infinite sequence u are* distinct (*different from each other*), *then its range is an infinite, though countable, set.*

Proof. By assumption, u is a *one-to-one* map (its terms being *distinct*), with $D_u = N = \{1, 2, \dots\}$. The range of u is the u-image of its domain N, i.e., $u[N]$. Now, as N is infinite,[1] so also is $u[N]$ by Theorem 2. □

Theorem 3. *If the sets A and B are both countable, so is $A \times B$.*

Proof. If A or B is empty, then $A \times B = \emptyset$, and all is proved.

Thus let A and B be nonempty. As before, we may assume that they *fill* two sequences, $A = \{a_n\}$ and $B = \{b_m\}$. For convenience, we also assume that these sequences are infinite (if not, *repeat* some terms). Then, by definition, $A \times B$ is the set of all ordered pairs of the form (a_n, b_m), where n and m take on *independently* the values 1, 2, ... Call $n+m$ the *rank* of the pair (a_n, b_m). The only pair of rank 2 is (a_1, b_1). Of rank 3 are (a_1, b_2) and (a_2, b_1). More generally,

$$(a_1, b_{r-1}), (a_2, b_{r-2}), \dots, (a_{r-1}, b_1) \tag{1}$$

are the $r - 1$ pairs of rank r.

We now put all pairs (a_n, b_m) in *one* sequence as follows. We start with (a_1, b_1); then take the two pairs of rank 3; then the three pairs of rank 4, and so on. At the $(r-1)$-th step, we take all pairs of rank r in the order shown in (1). Continuing this process for all ranks ad infinitum, we obtain the sequence of pairs

$$(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_1, b_3), (a_2, b_2), \dots .$$

[1] A proof of this fact will be suggested in Chapter 2, §6, Problem 15.

By construction, this sequence contains *all pairs of any rank*, hence all pairs that form the set $A \times B$ (for every such pair has some rank r; so it *must* occur in the sequence). Thus $A \times B$ is put in a sequence. $\square$

As an application, consider the set Q of all *positive rationals*, i.e., fractions n/m where n and m are naturals. Let $n+m$ be called the *rank* of n/m, where n/m is written in *lowest terms*. By the same process (writing the fractions in the order of their ranks), we put Q in an infinite sequence of distinct terms:

$$1/1,\ 1/2,\ 2/1,\ 1/3,\ 3/1,\ 1/4,\ 2/3,\ 3/2,\ \dots\ .$$

Hence we obtain the following.

Corollary 2. *The set R of all rational numbers is countable.*

Indeed, we only have to insert the negative rationals and 0 in the above sequence, as follows:

$$0,\ 1,\ -1,\ \frac{1}{2},\ -\frac{1}{2},\ 2,\ -2,\ \frac{1}{3},\ -\frac{1}{3},\ 3,\ -3,\ \dots\ .$$

A similar "ranking" method also yields the following result.

Theorem 4. *The union of any sequence of countable sets $\{A_n\}$ is countable.*

Proof. We must show that $A = \bigcup_n A_n$ can be put in one sequence. Now, as each A_n is countable, we may set

$$A_n = \{a_{n1},\ a_{n2},\ \dots,\ a_{nm},\ \dots\},$$

where the *double* subscripts are to distinguish the sequences representing different sets A_n. As before, we may assume that all sequences are infinite.

Clearly $\bigcup A_n$ consists of the elements of *all* A_n combined, i.e., of *all* a_{nm} ($n,\ m \in N$). Call $n+m$ the *rank* of the term a_{nm}. Proceed as in Theorem 3 to obtain

$$A = \bigcup A_n = \{a_{11},\ a_{12},\ a_{21},\ a_{13},\ a_{22},\ a_{31},\ \dots\}.$$

Thus A can be put in a sequence. $\square$

Note 1. Theorem 4 is briefly stated as "*Any countable union of countable sets is countable*" ("*countable union*" means "union of a *countable* family of sets," i.e., one that can be put in a finite or infinite sequence $\{A_n\}$).

Note 2. In particular, Theorem 4 applies to *finite* unions. Thus, *if A and B are countable sets, so is $A \cup B$.* (So also are $A \cap B$ and $A - B$ since they are subsets of the countable set A; see Theorem 1.)

In the proof of Theorem 4, we see a set A whose elements a_{nm} carried *two* subscripts. To any pair (n, m) of such subscripts there corresponds a unique element a_{nm} of A. Thus we can define a function u (of two variables, n and m) by setting

$$u(n, m) = a_{nm}, \quad n,\ m \in N.$$

Its domain is the set $N \times N$ of all *pairs* (m, n) of positive (or nonnegative) integers. Such a function is called an infinite *double sequence*, briefly denoted by $\{u_{nm}\}$. Its range D'_u may consist of *arbitrary* objects, namely the function values $u(n, m)$, briefly u_{nm}.

Exactly as in Theorem 4, we obtain the following result.

Corollary 3. *The range of any double sequence $\{u_{nm}\}$ is a countable set.*

To show that *uncountable* sets exist also, we shall now prove the uncountability of the interval $[0, 1)$, i.e., the set of all reals x such that $0 \le x < 1$. We assume as known that each real $x \in [0, 1)$ has a unique infinite decimal expansion $0.x_1x_2 \dots x_n \dots$, where the x_n are the decimal digits, possibly zeros, and the sequence $\{x_n\}$ does not terminate *in nines* (e.g., instead of $0.4999\dots$, we write $0.50000\dots$). This fact is proved in Chapter 2, §13.

Theorem 5. *The interval $[0, 1)$ of the real axis is uncountable.*

Proof. We must show that *no* sequence can comprise *all* of $[0, 1)$.

Indeed, take *any* sequence $\{u_n\}$ from $[0, 1)$. Write each term u_n as an infinite decimal fraction; say, $u_n = 0.a_{n1}a_{n2} \dots a_{nm} \dots$. Then construct a *new* decimal fraction $z = 0.x_1x_2 \dots x_n \dots$, choosing the digits x_n as follows.

If a_{nn} (i.e., the nth digit of u_n) is 0, take $x_n = 1$; otherwise, take $x_n = 0$. Thus, in all cases, $x_n \ne a_{nn}$, i.e., *z differs from each u_n in at least one decimal digit* (namely the nth digit). It follows that z differs from all u_n and hence is not in the sequence $\{u_n\}$, even though $z \in [0, 1)$. Thus, no matter what the choice of $\{u_n\}$ was, we found some $z \in [0, 1)$, *not* in the range of that sequence. Hence *no* $\{u_n\}$ contains all of $[0, 1)$. $\square$

Note 3. Observe that the members a_{nn} used in that proof form the "diagonal" of the indefinitely extending square consisting of all a_{nn}:

$$\begin{array}{cccccc}
a_{11} & a_{12} & a_{13} & \dots\dots & a_{1n} & \dots \\
a_{21} & a_{22} & a_{23} & \dots\dots & a_{2n} & \dots \\
a_{31} & a_{32} & a_{33} & \dots\dots & a_{3n} & \dots \\
\dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\
a_{n1} & a_{n2} & a_{n3} & \dots\dots & a_{nn} & \dots \\
\dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots
\end{array}$$

Therefore the method used above is called the *diagonal process* (due to Cantor).

Now, by Corollary 2, all rationals *can* be put in a sequence. But, as shown above, no such sequence can cover all of $[0, 1)$. Thus $[0, 1)$ must contain numbers *that are not rational*, i.e., cannot be written as ratios of integers, n/m. Moreover, such numbers, called *irrational*, must form an uncountable set, for otherwise its union with the countable set of all rationals in $[0, 1)$ would be countable (by Note 2), whereas actually this union is the uncountable set $[0, 1)$.

The same argument applies to any other line interval with endpoints a and b $(a < b)$, since *any* such interval is uncountable (see Problem 2). Thus we have the following.

Corollary 4. *Between any two real numbers a and b $(a < b)$ there are uncountably many irrational numbers.*

Note 4. By Theorem 2, any superset of $[0, 1)$ is uncountable. In particular, so is the entire set of real numbers (the real axis).

We thus see that *the irrationals form an uncountable set*. In this sense, there are many more irrationals than rationals. Both sets are infinite. Thus there are *different* kinds of "infinities".

Problems on Countable and Uncountable Sets

1. Show that Theorem 1 holds also if $A \nsubseteq D_f$.
[Hint: Define a new map g on $A \cup D_f$ by $g(x) = f(x)$ if $x \in D_f$ and $g(x) = x$ if $x \notin D_f$. Noting that $D_g \supseteq A$, infer from what was already proved that $g[A]$ is countable, and hence so is $f[A]$ (why?).]

2. Let a and b be real numbers, $a < b$. Define a mapping f on $[0, 1)$ by setting $f(x) = a + x(b - a)$. Show that f is one-to-one and that it is *onto* $[a, b)$. Then, from Theorems 2 and 5, infer that $[a, b)$ is uncountable.

3. Show that if B is countable but A is not, then $A - B$ is uncountable.
[Hint: If $A - B$ were countable, so would be $(A - B) \cup B \supseteq A$.]

4. Show that every infinite set A contains a *countable* infinite set.
[Hint: Fix any element $x_1 \in A$; A cannot consist of x_1 alone (why?), so there is another element $x_2 \in A - \{x_1\}$. Again, $A \neq \{x_1, x_2\}$ (why?), so there is an element $x_3 \in A - \{x_1, x_2\}$, and so on. Proceeding step by step, we select from A an infinite sequence $\{x_n\}$ of *distinct* terms. Then $C = \{x_1, x_2, \dots, x_n, \dots\}$ is the required subset of A. (A reader acquainted with axiomatic set theory will observe that this proof uses the so-called *axiom of choice*.)]

5. Infer from Problem 4 that if A is infinite, then there is a mapping $f\colon A \to A$ that is one-to-one but *not onto* A.
[Hint: Choose $C = \{x_1, x_2, \dots, x_n \dots\}$ as in Problem 4. Then define f as follows: If $x \in A - C$, then $f(x) = x$; if, however, $x = x_n$ for some n, then $f(x) = f(x_n) = x_{x+1}$. Observe that *never* $f(x) = x_1$, and so f is not *onto* A. Verify however that f is one-to-one.]

6. Let $f\colon A \to B$ be a one-to-one map such that $B \subset A$, and let $x_1 \in A - B$. Inductively (step by step) define an infinite sequence:

$$x_2 = f(x_1),\ x_3 = f(x_2),\ \dots,\ x_{n+1} = f(x_n),\ \dots,\ n = 1,\ 2,\ \dots\ .$$

Observe that all x_n except x_1 are in B (why?), and so $x_n \neq x_1$, $n = 2, 3, \dots$. Show that *all* x_n are distinct (i.e., different from each other) and hence B *is an infinite set* by Corollary 1.

[Hint: Seeking a contradiction, suppose there is an n such that $x_n = x_m$ for some $m > n$, and take the *least* such n. Then $n - 1$ *does not* have this property, and so $x_{n-1} \neq x_m$ for *all* $m > n - 1$. As f is one-to-one, we get $f(x_{n-1}) \neq f(x_m)$, i.e., $x_n \neq x_{m+1}$, for *all* $m > n - 1$ (contradiction!).]

Combining this with Problem 5, infer that *a set A is infinite iff there is a map $f\colon A \to A$ that is one-to-one but not onto A.*

7. Using the result of Problem 6, show that the number n of elements in a finite set A is *uniquely* determined. More precisely, if $A =$ the range of a sequence u of distinct terms with $D_u = \{1, 2, \dots, n\}$, it is not the range of any sequence v with $D_v = \{1, 2, \dots, m\}$, $m \neq n$.
[Hint: Suppose this *is* the case, with $m < n$, say. Then show that the composite map $u \cdot v^{-1}$ is one-to-one (by Theorem 2 of §6) but not *onto* A, though its domain is A. Infer that A is *infinite* (contradiction!).]

Chapter 2

The Real Number System

§1. Introduction

Historically, the real number system is the result of a long gradual development that started with *positive integers* ("*natural numbers*") 1, 2, 3, ..., later followed by the invention of the *rational numbers* (i.e., fractions p/q where p and q are integers); it was completed by the discovery of irrational numbers.

It is possible to reproduce this gradual development also in exact theory, that is, to build up the real number system step by step from natural numbers. At this stage, however, we shall assume the set of all real numbers as already given, without attempting to reduce the notion of real number to simpler concepts. Also without definition (i.e., as so-called *primitive concepts*) shall we introduce the notions of the sum $(a + b)$ and the product, $(a \cdot b)$ or (ab), of two real numbers a and b, as well as the inequality relation $<$ (read: "*less than*"). The set of all real numbers taken together will be denoted by E^1 (read: "E one"). The formula "$x \in E^1$" means that x is in E^1, i.e., x is a real number.

Thus our primitive concepts are E^1 (set of all reals), $+$ (plus sign), $\cdot$ (multiplication sign), and $<$ (inequality sign).

Remark. Every mathematical theory must start with certain concepts accepted as primitive (i.e., without definition), since it is impossible to define *all* terms used. Indeed, any definition can only explain some terms by means of others. If the latter, too, were to be defined, *new* defining terms would be needed, and this process would never end. It is often only a matter of convention, *which* notions to accept as the first (i.e., the primitive) ones. Once, however, the choice has been made, all other notions should be defined in terms of the primitive ones. Similarly, it is impossible to *prove* all statements of a deductive theory. Certain propositions (called *axioms*) must be accepted as the *first* ones, without proof. Once, however, the axioms have been stated, all the following propositions (called *theorems*) must be proved, i.e., deduced in a strictly logical way from the axioms. This procedure characterizes every exact deductive science.

We now proceed to state a system of axioms for real numbers. The first nine axioms will be given in §2 (for a reason to be explained later, they will be called "*axioms of an ordered field*"). The last (10th) axiom will be formulated in a later section.

§2. Axioms of an Ordered Field

We shall assume as axioms (i.e., without proof) the following simple properties of real numbers. (The reader is certainly familiar with these properties from school algebra, where they are often regarded as "obvious", so that it might seem superfluous to mention them. We must, however, *state them as axioms* in accordance with our introductory remarks made in §1. Each axiom has a name given in parenthesis.)

A. Axioms of addition and multiplication.

I (Closure law) *The sum $x+y$ and the product xy of any two real numbers x and y are themselves real numbers.* In symbols:

$$(\forall x, y \in E^1) \quad (x+y) \in E^1,\ (xy) \in E^1.$$

II (Commutative laws) $(\forall x, y \in E^1)\ x+y = y+x,\ xy = yx.$

III (Associative laws) $(\forall x, y, z \in E^1)\ (x+y)+z = x+(y+z),\ (xy)z = x(yz).$

IV (Existence of neutral elements)

(a) *There exists a (unique) real number, called "zero" (0), such that, for all real x, $x+0=x$.*

(b) *There exists a (unique) real number, called "one" (1), such that $1 \neq 0$ and, for all real x, $x \cdot 1 = x$.* In symbols:

$$(\exists!\, 0 \in E^1)\ (\forall x \in E^1) \quad x+0=x,$$
$$(\exists!\, 1 \in E^1)\ (\forall x \in E^1) \quad x \cdot 1 = x, \quad 1 \neq 0.$$

The numbers 0 and 1 are called the *neutral elements* of addition and multiplication, respectively.

V (Existence of inverses)

(a) *For every real number x, there is a (unique) real number, denoted $-x$, such that $x+(-x)=0$.*

(b) *For every real number x, other than 0, there is a (unique) real number denoted x^{-1}, such that $x \cdot x^{-1} = 1$.* In symbols:

$$(\forall x \in E^1)\ (\exists!\, -x \in E^1) \quad x+(-x)=0,$$
$$(\forall x \in E^1 \mid x \neq 0)\ (\exists!\, x^{-1} \in E^1) \quad x \cdot x^{-1} = 1.$$

The numbers $-x$ and x^{-1} are called, respectively, the *additive inverse* (or the *symmetric*) and the *multiplicative inverse* (or the *reciprocal*) of x.

VI (Distributive law) $(\forall x, y, z \in E^1)\ (x+y)z = xz+yz.$

Note. The uniqueness assertions in Axioms IV and V could actually be omitted since they can be proved from other axioms.

B. Axioms of order.

VII (Trichotomy) *For any real numbers x and y, we have either $x<y$ or $y<x$ or $x=y$, but never two of these relations together.*

VIII (Transitivity) *If x, y, z are real numbers, with $x<y$ and $y<z$, then $x<z$.* In symbols:

$$(\forall x, y, z \in E^1) \quad x<y<z \text{ implies } x<z.$$

IX (Monotonicity of addition and multiplication)

(a) $(\forall x, y, z \in E^1)\ x<y$ *implies* $x+z<y+z.$

(b) $(\forall x, y, z \in E^1)\ x<y$ *and* $0<z$ *implies* $xz<yz.$

Note 1. As has already been mentioned, one additional (10th) axiom will be stated later.

Note 2. While every real number has an *additive* inverse (Axiom V(a)), only nonzero numbers have *reciprocals*. The number 0 *has no reciprocal*. (Axiom V(b).)

Note 3. Note the restriction $0<z$ in Axiom IX(b). It is easy to see that without this restriction the axiom would be false. For example, we have $2<3$, but $2(-1)$ is not less than $3(-1)$. No such restriction occurs in Axiom IX(a).

Due to the introduction of inequalities "$<$" and the Axioms VII–IX, the real numbers may be regarded as given in some *definite order*, under which smaller numbers *precede* the larger ones. (This is why Axioms VII–IX are called "axioms of *order*".) We express this fact briefly by saying that E^1 is an *ordered set*. More precisely, an *ordered set* is a set in which a certain relation "$<$" has been defined in such a manner that the trichotomy and transitivity laws are satisfied.

The ordering of real numbers can be visualized by "plotting" them as points on a directed line ("the real axis"), as shown below in Figure 8:

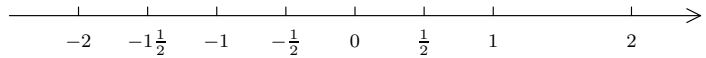


FIGURE 8

Therefore, real numbers are also often referred to as "*points*" of the real axis. We say, e.g., "the point x" instead of "the number x." We assume that the reader is familiar with this process of geometric representation of real numbers. We shall not dwell on its justification since it will only be used as illustration, not as proof.

It should be noted that the axioms only specify certain properties of real numbers *without indicating what these numbers actually are*. This question is left entirely open, so that we may regard real numbers as just any mathematical objects that are only supposed to satisfy our axioms but otherwise are *entirely arbitrary*. This makes our theory more general. Indeed, our theory also applies to any other set of objects (numbers or not numbers), provided only that they satisfy our axioms with respect to a certain relation of order ($<$) and certain operations ($+$) and ($\cdot$), which may, but need not, coincide with ordinary number addition and multiplication. *Whatever follows logically from the axioms must be true not only for real numbers but also for any other set that conforms with these axioms*. In this connection, we introduce the following definitions.

Definition 1.

A *field* F is any set of objects with two operations ($+$) and ($\cdot$) (usually called "addition" and "multiplication") defined in it, provided that these objects and operations satisfy the first six axioms (I–VI) listed above.

If this set is also equipped with an order relation ($<$) satisfying the additional three axioms VII–IX, it is called an *ordered field*.

In particular, *the real number system E^1 is an ordered field*. Of course, when speaking of ordered fields in general, the term "real number" in the axioms should be replaced by "element of F." Similarly, 0 and 1 should be interpreted as elements of the field satisfying Axiom IV(a) and (b), but not necessarily as ordinary numbers.

E^1 is not the only ordered field known in mathematics. Indeed, many examples of ordered and unordered fields are studied in higher algebra. We shall encounter some of them later.

As has been mentioned, everything that can be deduced from Axioms I–IX applies not only to E^1 but also to any other ordered field F (since F is supposed to satisfy these axioms). Therefore, we shall henceforth formulate our definitions and theorems in a more general way, speaking of "ordered fields" in general instead of E^1 alone. Of course, whatever we say about ordered fields applies in particular to E^1, and this particular example should be always kept in mind.

Definition 2.

An element x of an ordered field F is said to be *positive* or *negative* according as $x > 0$ or $x < 0$. The element 0 itself is neither positive nor negative.

Here and henceforth "$x > y$" means the same as "$y < x$". We also write "$x \le y$" for "$x < y$ or $x = y$"; similarly for $x \ge y$.

The numbers 0 and 1 have been introduced in Axiom IV, but we do not yet "officially" know what such symbols as 2, 3, 4, ..., etc. mean, since they have not yet been defined. Indeed, we have only introduced the notion of *real number*, but not that of *natural number* (or *integer*). Therefore, in our system, the latter must be defined in terms of our primitive concepts. Since, however, addition is already known, we can use it to define positive integers step by step, as follows:

$$2 = 1 + 1,\ 3 = 2 + 1,\ 4 = 3 + 1,\ 5 = 4 + 1,\ \text{etc.}$$

If this process is continued indefinitely, we obtain what is called the set of all "*positive integers*" (or "*natural numbers*"). We may say that a natural number is one that can be obtained from 0 by adding to it 1 a finite number of times. A similar process is, of course, possible not only in E^1 but in any field. Thus we may speak of "natural elements" in any field.

This may serve as a preliminary definition of natural numbers. A more exact definition will be given in §5.

Definition 3.

Given several elements a, b, c, d of a field F, we define

$$a + b + c = (a + b) + c,\ \ a + b + c + d = (a + b + c) + d,\ \text{etc.}$$

Similarly for multiplication.

§3. Arithmetic Operations in a Field

All arithmetic properties of real numbers can be deduced from the axioms stated in §2. We shall dwell on only some of these properties to illustrate the method of proving them. In this section we shall investigate inferences of the first six axioms, which hold in *every* (even unordered) field F.

Definition 1.

Given two elements x and y of a field F, we define their *difference*,

$$x - y = x + (-y).$$

In other words, *to subtract an element y means to add its additive inverse*, $-y$.

If $y \ne 0$, we also define the *quotient of x by y*,

$$\frac{x}{y} = x \cdot (y^{-1}),$$

also denoted by x/y. In other words, *to divide x by y means to multiply x by the reciprocal of y* (provided that y^{-1} exists, i.e., that $y \neq 0$).

In this way we have defined two new operations: *subtraction* (i.e., formation of differences) and *division* (i.e., formation of quotients). **Note:** *Division by 0 is undefined, hence inadmissible.* Since subtraction and division have been defined as special cases of addition and multiplication, respectively, we can apply to them our axioms to obtain the following corollaries.

Corollary 1. *The difference $x - y$ and the quotient x/y (where $y \neq 0$) of two real numbers x and y are themselves real numbers.* (*Similarly for differences and quotients of field elements in general.*)

In symbols:

$$(\forall x, y \in E^1) \quad (x - y) \in E^1,\ (x/y) \in E^1 \ \text{(the latter if } y \neq 0).$$

Corollary 2. *If a, b, c are elements of a field F, with $a = b$, then*

$$a + c = b + c \text{ and } ac = bc.$$

(*In other words, we may add one and the same element c to both sides of an equation $a = b$; similarly for multiplication.*)

In symbols:

$$(\forall a, b, c \in F) \quad a = b \text{ implies } a + c = b + c \text{ and } ac = bc.$$

Proof. By properties of equality, we obviously have $a + c = a + c$ (since the left side is the same as the right side). Now, as $a = b$, we may replace a by b on the right side. This yields $a + c = b + c$, as required. Similarly for $ac = bc$. □

The converse to this corollary is the following.

Corollary 3 (Cancellation law)**.** *If a, b, c are elements of a field F, then*

$$a + c = b + c \text{ implies } a = b.$$

If, further, $c \neq 0$, then

$$ac = bc \text{ implies } a = b.$$

(In other words, we may cancel a summand and a *nonzero* factor on both sides of an equation.)

Proof. Let $a + c = b + c$. By Corollary 2, we may add $(-c)$ on both sides of this equation to get

$$(a + c) + (-c) = (b + c) + (-c),$$

or, by associativity (Axiom III),

$$a + [c + (-c)] = b + [c + (-c)].$$

As $c + (-c) = 0$ (by Axiom V), we obtain $a + 0 = b + 0$, i.e., $a = b$ (by Axiom IV); similarly for multiplication. □

Theorem 1. *Given two elements, a and b, of a field F, there always exists a unique element x such that $a + x = b$; this element equals the difference $b - a$.* (*Thus $a + x = b$ means that $x = b - a$.*)

If, further, $a \neq 0$, there also is a unique element $y \in F$, with $ay = b$; this element equals the quotient b/a. (*Thus $ay = b$, $a \neq 0$, means that $y = b/a$.*)

In symbols:

$$(\forall a, b \in F)\ (\exists!\, x, y \in F) \quad a + x = b, \quad ay = b \ \text{(the latter if } a \neq 0).$$

We prove only the first part of the theorem, leaving the second (which is proved in the same way) to the reader.

It is easily checked that the equation $a + x = b$ is satisfied by $x = b - a$. In fact, we have (using commutativity, associativity, and Axioms IV and V)

$$a + x = a + (b - a) = (b - a) + a = [b + (-a)] + a = b + [(-a) + a] = b + 0 = b.$$

Thus the equation $a + x = b$ has *at least one* solution for x, namely $x = b - a$. To prove that this solution is unique, suppose that we have still another solution, x', say. Then we obtain $a + x = b$ and $a + x' = b$, so that $a + x = a + x'$, or $x + a = x' + a$. Cancelling a (by Corollary 3), we see that $x = x'$, so that the two solutions necessarily coincide.

Thus both the existence and uniqueness of the solution have been proved.

Theorem 1 shows that subtraction and division are *inverse* operations with respect to addition and multiplication. It can also be interpreted as a rule for transferring a summand or a factor from one side of an equation to the other.

Corollary 4. *For any element x of a field F, we have $0 - x = -x$. If, further, $x \neq 0$, then $1/x = x^{-1}$.*

In fact, we have, by definition,

$$0 - x = 0 + (-x) = -x \quad \text{(Axiom IV)}.$$

Similarly,

$$1/x = 1 \cdot x^{-1} = x^{-1}.$$

Corollary 5. *For any element x of a field F, we have*

$$x \cdot 0 = 0 \cdot x = 0.$$

(Hence we never have $0 \cdot x = 1$; this is why 0 cannot have a multiplicative inverse.)

In fact, by distributivity (Axiom VI) and by Axiom IV, we get

$$0x + 0x = (0 + 0)x = 0x = 0 + 0x.$$

Thus $0x+0x = 0+0x$. Cancelling $0x$ on both sides (Corollary 3), we obtain $0 \cdot x = 0$, and by commutativity, also $x \cdot 0 = 0$.

Corollary 6 (Rule of signs)**.** *For any elements a, b of a field F, we have*

(i) $a(-b) = (-a)b = -(a \cdot b)$;

(ii) $-(-a) = a$;

(iii) $(-a)(-b) = ab$.

Proof. Formula (i) means that $a(-b)$, and similarly $(-a)b$, equals the additive inverse of ab. Therefore, to prove its first part, we have to show that $a(-b) + ab = 0$ (for, this is the definition of the additive inverse). But, by distributivity, we have

$$a(-b) + ab = a[(-b) + b] = a \cdot 0 = 0$$

(by Corollary 5), as required.

Similarly we show that $(-a)b = -(ab)$ and that $-(-a) = a$.

Finally, (iii) is obtained from (i) when a is replaced by $(-a)$. Thus all is proved. □

§4. Inequalities in an Ordered Field. Absolute Values

As further examples of applications of our axioms, we now proceed to deduce some corollaries to Axioms VI–IX. They apply to any *ordered* field.

Corollary 1. *If x is a positive element of an ordered field F, then $-x$ is negative; and if x is negative, then $-x$ is positive.*

Proof. Given $x > 0$, we may add $(-x)$ on both sides, by Axiom IX. Then we obtain

$$x + (-x) > 0 + (-x), \text{ i.e., } 0 > -x,$$

as required. Similarly, it is shown that $x < 0$ implies $-x > 0$. □

Corollary 2 (Addition and multiplication of inequalities)**.** *If a, b, x, y are elements of an ordered field F, such that $a < b$ and $x < y$, then*

$$a + x < b + y$$

(i.e., we may always add two inequalities).

If, further, a, b, x, y are positive, then $a < b$ and $x < y$ implies $ax < by$ (i.e., the inequalities may be multiplied).

Proof. Both parts of the corollary are proved in a similar way, so we prove only the second part.

Suppose that $a < b$, and $x < y$, with a, b, x, y positive. Then, multiplying the first inequality by x and the second by b (Axiom IX(b)), we have

$$ax < bx \text{ and } bx < by.$$

Hence, by transitivity, $ax < bx < by$, i.e., $ax < by$, as required. □

Note 1. Multiplication of inequalities may fail if the numbers involved are not positive. For example, we have $-2 < 3$ and $-2 < 1$, but multiplication would lead to a false result: $4 < 3$. (However, it suffices that only b and x in Corollary 2 be positive.)

Corollary 3. *All nonzero elements of an ordered field have positive squares. That is, if $a \neq 0$, then $a^2 = a \cdot a > 0$. (Hence $1 = 1^2 > 0$.)*

Proof. As $a \neq 0$, we have, by trichotomy, either $a > 0$ or $a < 0$.

If $a > 0$, then we may multiply by a, obtaining $a \cdot a > 0 \cdot a = 0$, i.e., $a^2 > 0$.

If $a < 0$, then, by Corollary 1, $-a > 0$; so we may multiply the inequality $a < 0$ by $(-a)$, using again Axiom IX(b). We then obtain

$$a(-a) < 0 \cdot (-a) = 0, \text{ i.e., } -a^2 < 0, \text{ whence } a^2 > 0, \text{ as required.} \quad \square$$

Definition.

Given an element x of an ordered field F, we define its *absolute value*, denoted $|x|$, as follows:

If $x \geq 0$, then $|x| = x$; if, however, $x < 0$, then $|x| = -x$.

In particular, $|0| = 0$. It follows that $|x|$ *is always nonnegative*. In fact, if $x \geq 0$, then $|x| = x \geq 0$; and if $x < 0$, then by Corollary 1, $-x > 0$; and here $-x = |x| > 0$. Moreover, we always have

$$-|x| \leq x \leq |x|. \tag{1}$$

For, if $x \geq 0$, then $|x| = x$ by definition, and $-|x| = -x \leq 0 \leq x$. If, however, $x < 0$, then $|x| > x$ since $|x|$ is positive, while x is negative and $x = -|x|$. Thus (1) holds in both cases.

Corollary 4. *For any elements x and y of an ordered field F, we have $|x| < y$ iff $-y < x < y$.*

Proof. Suppose first that $|x| < y$. Then, by formula (1), we have $x \leq |x| < y$, whence $x < y$. It remains to prove that $-y < x$. This is certainly true if x is nonnegative (for $-y$ is negative here). If, however, x is negative, then by definition, $-x = |x|$, whence $-x < y$ (for $|x| < y$, by assumption); that is, $-y < x$. Thus, in all cases, $|x| < y$ implies $-y < x < y$.

The converse is proved in a similar way, by distinguishing the two cases: $x \geq 0$ and $x < 0$. The details are left to the reader. □

Note 2. This corollary has a simple geometric interpretation. Namely, if x is plotted on the real axis, then $|x|$ is its (undirected) distance from the origin 0. Thus the formulas $|x| < y$ and $-y < x < y$ express both the fact that this distance is $< y$.

Corollary 5. *For any elements a and b of an ordered field F, we have*

$$|ab| = |a| \cdot |b|.$$

If further $b \neq 0$, then

$$\frac{|a|}{|b|} = \left|\frac{a}{b}\right|.$$

For the proof, consider the four possible cases:

(1) $a \geq 0,\ b \geq 0$; (2) $a \geq 0,\ b < 0$; (3) $a < 0,\ b \geq 0$; and (4) $a < 0,\ b < 0$.

The result then easily follows by the definition of absolute value.

Corollary 6. *For any elements a and b of an ordered field F, we have*

(i) $|a + b| \leq |a| + |b|$;

(ii) $\big||a| - |b|\big| \leq |a - b|$.

(These are the so-called *triangle inequalities*.)

Inequality (i) can be proved by considering the four cases specified in the proof of Corollary 5, but it is much simpler to use Corollary 4. Indeed, by formula (1) on page 59, we have

$$-|a| \leq a \leq |a| \text{ and } -|b| \leq b \leq |b|.$$

Adding, we obtain

$$-(|a| + |b|) \leq a + b \leq |a| + |b|.$$

But by Corollary 4, with $x = a + b$ and $y = |a| + |b|$, this means that $|a + b| \leq |a| + |b|$, as required.

To prove (ii), let $x = a - b$. By part (i), $|x + b| \leq |x| + |b|$, i.e.,

$$|(a - b) + b| \leq |a - b| + |b|,$$

whence $|a| \leq |a - b| + |b|$, or

$$|a| - |b| \leq |a - b|.$$

Interchanging a and b, we also have $|b| - |a| \leq |a - b|$, and (ii) follows.

Corollary 7. *Given any two elements a and b $(a < b)$ of an ordered field F, there always is an element $x \in F$ such that $a < x < b$.* (This element is said to lie *between* a and b.)

This important proposition is often expressed by saying that *every ordered field* (*in particular*, E^1) *is densely ordered*. More generally, an ordered *set* F is said to be *densely ordered* if it has the property expressed in Corollary 7. In this connection, Corollary 7 will be referred to as the *density property* of real numbers, or the *density of an ordered field*.

Proof. It suffices to take

$$x = \frac{1}{2}(a + b).$$

Then Axiom II easily yields $a < x < b$. The details are left to the reader. □

Note 3. Corollary 7 shows that, given a real number a, there never exists a number "closest" or "next" to a. In fact, if b were such a number, then by Corollary 7, one could find a number x $(a < x < b)$ still closer to a.

Note 4. Having found *one* number, say x_1, between a and b, we can again apply Corollary 7 to find a number x_2 between x_1 and b, then again a number x_3 between x_2 and b, and so on. Since this process can be continued indefinitely, Corollary 7 may be strengthened to say that there are *infinitely many* real numbers between any two given numbers a and b; similarly for ordered fields in general.

As previously noted, the propositions proved in §§3 and 4 are only *examples* illustrating the deduction of arithmetic rules from axioms. Other such examples are given in problems below. We shall use them freely later.

These problems are to be treated as *logical* exercises, with the purpose of finding out *which particular axioms are needed in each case*. From the theoretical point of view, this is important in its own right. Practically, one might think of a computer programmed to deduce the rules of arithmetic purely mechanically from certain axioms. The computer does not "know" anything but the rules that have been fed into it. Even such "obvious" formulas as "$2 + 2 = 4$" the computer will have to *deduce* from axioms and definitions, as for example,

$$\begin{aligned} 2 + 2 &= 2 + (1 + 1) && \text{(definition of "2")} \\ &= (2 + 1) + 1 && \text{(associativity of addition)} \\ &= 3 + 1 && \text{(definition of "3")} \\ &= 4 && \text{(definition of "4").} \end{aligned}$$

Conclusion: To enable the computer to prove that $2+2 = 4$, one must "feed" into it at least the associative law of addition and the definitions of 2, 3, 4.

The main thing in such exercises is not to "jump" some axiom or definition (otherwise the computer will get "stuck"); use only one at a time! Do not omit parentheses in such expressions as $(a+b)+c$ without mentioning the definition of $a+b+c$. *Caution*: The commutative laws were stated for *two* elements only; such formulas as $abc = bac$, i.e., $(ab)c = (ba)c$, must be *proved*.

Problems on Arithmetic Operations and Inequalities in a Field

1. Supply the missing details (in particular, those "left to the reader") in the proofs of all corollaries stated in §§3 and 4.

2. Using the "preliminary definition" of natural numbers, deduce from our axioms that

(a) $2+3=5$;

(b) $3+4=7$;

(c) $2\cdot 2=4$;

(d) $3\cdot 2=6$.

Name the axioms used at each step (e.g., "associativity of addition," etc.).

3. Deduce from axioms, step by step, that in any field F we have the following:

(i) $abcd=cbad=dacb$; similarly for addition.

(ii) If $x\neq 0$ and $y\neq 0$, then $xy\neq 0$.
[Hint: If xy were zero, then multiplication by y^{-1} would yield $x=0$, contrary to our assumption.]

(iii) $(xy)^{-1}=x^{-1}y^{-1}$, provided that $x\neq 0$ and $y\neq 0$. Why must one assume that neither x nor y are zero?
[Hint: Proceed as in the proof of Corollary 6 of §3.]

(iv) If $x\neq 0$, $y\neq 0$ and $z\neq 0$, then $(xyz)^{-1}=x^{-1}y^{-1}z^{-1}$.

(v) If $x\neq 0$ and $y\neq 0$, then

$$\frac{a}{x}\cdot\frac{b}{y}=\frac{ab}{xy} \quad\text{and}\quad \frac{a}{x}+\frac{b}{y}=\frac{ay+bx}{xy}.$$

[Hint: By definition, $a/x=ax^{-1}$, $b/y=by^{-1}$, etc. Use axioms, previous corollaries, and the result of Problem 3(iii).]

(vi) $(a+b)(x+y)=ax+bx+ay+by$; (vi′) $(a+b)^2=a^2+2ab+b^2$.

(vii) $(a+b)(x-y)=ax+bx-ay-by$; (vii′) $(a+b)(a-b)=a^2-b^2$.

In all cases, arrange the proof in such a manner that, at each step, only *one* axiom, *one* definition or *one* previous corollary is used, and *name* it (except for the *closure law*, which is used at *each* step). Only Axioms I–VI may be used since F is not necessarily an *ordered* field.

4. Continuing Problem 3 (with the same directives), use Definition 3 of §2 to show that

$$(a+b+c)x=ax+bx+cx \text{ and } (a+b+c+d)x=ax+bx+cx+dx;$$

similarly for a sum of 5 terms (first define it!).

5. In the same manner as in Problem 3, prove the following for *ordered* fields:

(i) If $x>0$, then also $x^{-1}>0$.

(ii) If $x>y>z>u$, then $x>u$.

(iii) If $x>y\geq 0$, then

$$x^2>y^2 \text{ and } x^3>y^3\geq 0 \quad (\text{where } x^3=x^2x);$$

similarly,

$$x^4>y^4\geq 0 \quad (\text{where } x^4=x^3x).$$

Which (if any) of these propositions remain valid also if x or y is negative? Give proof.

(iv) If $x>y>0$, then $1/x<1/y$. What if $x>0>y$ or $0>x>y$?

(v) $|a+b+c|\leq |a|+|b|+|c|$ and $|a+b+c+d|\leq |a|+|b|+|c|+|d|$.

§5. Natural Numbers. Induction

At the end of §2, we showed how to select from E^1 the natural numbers 1, 2, 3, ..., starting with 1 and then adding 1 to each preceding number to get the following one. This process also applies to any other field F; the elements so selected are called the *natural elements* of F, and the set of all such elements (obtained by continuing the process indefinitely) is denoted by N. Note that, by this construction, *we always have* $n+1\in N$ *if* $n\in N$.

*A more precise approach to natural elements is as follows.[1] A subset S of a field F is called *inductive* iff

(i) $1\in S$ (S contains the unity element of F) and

(ii) $(\forall x\in S)\ x+1\in S$ (S contains $x+1$ whenever x is in S).[2]

Define N to be the intersection of *all* such subsets. We then obtain the following.

***Theorem 1.** *The set N so defined is inductive itself. In fact, it is the "smallest" inductive set in F (i.e., contained in any other such set).*

Proof. We have to show that, with our new definition,

(i) $1\in N$ and

[1] The beginner may omit all "starred" passages and simply assume Theorems 1′ and 2′ below as additional axioms without proof.

[2] Such subsets do exist; e.g., the entire field F is inductive since $1\in F$ and $(\forall x\in F)\ x+1\in F$, by the closure law.

(ii) $(\forall x \in N)\ x+1 \in N$.

Now, by definition, the unity 1 is in *each* inductive set; hence it also belongs to the intersection of such sets, i.e., to N. Thus $1 \in N$, as claimed.

Next, take any $x \in N$. Then, by our new definition of N, x is in *every* inductive set S. Hence, by property (ii) of such sets, also $x+1$ is in every such S; thus $x+1$ is in the intersection of *all* inductive sets, i.e., $x+1 \in N$, and so N is inductive, indeed.

Finally, by definition, N is the *common part* of all such sets, hence contained in each. $\square$

For applications, Theorem 1 is usually expressed as follows.

Theorem 1′ (First induction law)**.** *A proposition $P(n)$ involving a natural n holds for* all $n \in N$ *in a field F if*

(i) *it holds for $n = 1$ [$P(1)$ is true]; and*

(ii) *whenever $P(n)$ holds for $n = m$, it holds for $n = m+1$;* [$P(m) \implies P(m+1)$].

***Proof.** Let S be the set of all those $n \in N$ for which $P(n)$ is true; that is, $S = \{n \in N \mid P(n)\}$. We must show that actually *each* $n \in N$ is in S, i.e., $N \subseteq S$.

First, we show that S is inductive. By our assumption (i), $P(1)$ is true, so $1 \in S$.

Next, suppose $x \in S$. This means that $P(x)$ is true. But by assumption (ii), this implies $P(x+1)$, i.e., $x+1 \in S$. Thus $(\forall x \in S)\ x+1 \in S$ and $1 \in S$; so S *is* inductive. But then, by Theorem 1 (second clause), $N \subseteq S$. $\square$

This theorem is widely used to prove general propositions on natural elements, as follows. In order to show that some formula or proposition $P(n)$ is true for *every* natural n, we *first verify* $P(1)$, i.e., show that $P(n)$ holds for $n = 1$. We then show that

$$(\forall m \in N) \quad P(m) \implies P(m+1);$$

that is, *if* $P(n)$ holds for some value $n = m$, *then* it also holds for $n = m+1$. Once these two facts are established, Theorem 1′ ensures that $P(n)$ holds for *all* natural n.

Proofs of this kind are called *inductive*, or *proofs by induction*. Note that every such proof consists of *two* steps:

$$\text{(i) } P(1) \quad \text{and} \quad \text{(ii) } P(m) \implies P(m+1).$$

Special caution must be applied in step (ii). Here we *temporarily* assume that $P(n)$ has already been verified for some *particular* (but unspecified) value $n = m$.[3] From this assumption, we then try to *deduce* that $P(n)$ holds for $n = m+1$ as well. This fact must be *proved*; it would be a bad error to simply substitute $m+1$ for m in the assumed formula $P(m)$ since it was assumed for a *particular* value m, not for $m+1$. The following examples illustrate this procedure.[4]

Examples.

(A) *If m and n are natural elements, so are $m+n$ and mn.* To prove it, fix any $m \in N$. Let $P(n)$ mean that $m+n \in N$. We now verify the following:

(i) $P(1)$ *is true*; for $m \in N$ is *given*. Hence, by the very definition of N, $m+1 \in N$. But this means exactly that $P(n)$ holds for $n = 1$, i.e., $P(1)$ is true.

(ii) $P(k) \implies P(k+1)$ (here we use a different letter, k, since m is fixed already). Suppose that $P(n)$ holds for some *particular* $n = k$. This means that $m+k \in N$. Hence, by the definition of N, $(m+k)+1 \in N$; or, by associativity, $m+(k+1) \in N$. But this means exactly that $P(k+1)$ is true (if $P(k)$ is). Thus, indeed, $P(k) \implies P(k+1)$.

Since (i) and (ii) have been established, induction is complete; that is, Theorem 1′ shows that $P(n)$ holds for *each* $n \in N$, and this means that $m+n \in N$. As m and n are *arbitrary* naturals, our first assertion is proved.[5]

To show that $mn \in N$ also, we now let $P(n)$ mean that $mn \in N$ (for a *fixed* $m \in N$) and proceed similarly. We leave this to the reader.

(B) *If $n \in N$, then $n-1 = 0$ or $n-1 \in N$.* Indeed, let $P(n)$ mean that $n-1 = 0$ *or* $n-1 \in N$ (*one* of the two is required). We again verify the two steps:

(i) $P(1)$ *is true*; for if $n = 1$, then $n-1 = 1-1 = 0$. Thus one of the two desired alternatives, namely $n-1 = 0$, holds if $n = 1$. Hence $P(1)$ is true.

(ii) $P(m) \implies P(m+1)$. Suppose $P(n)$ holds for some *particular* value $n = m$ (inductive hypothesis). This means that either $m-1 = 0$ or $m-1 \in N$.

In the first case, we have $(m-1)+1 = 0+1 = 1 \in N$. But $(m-1)+1 = (m+1)-1$ by associativity and commutativity (verify!). Thus $(m+1)-1 \in N$.

[3] This temporary assumption is called the *inductive hypothesis*.

[4] Actually, these examples are basic theorems on naturals, to be well noted.

[5] Note the technique we applied here. Faced with *two* variables m and n, we *fixed* m and carried out the induction on n. This is a common procedure.

In the second case, $m-1 \in N$ implies $(m-1)+1 \in N$ by the very definition of N. Thus, in both cases, $(m+1)-1 \in N$, and this shows that $P(m+1)$ is true if $P(m)$ is.

As (i) and (ii) have been established, induction is complete.

(C) *In an* ordered *field, all naturals are* ≥ 1. Indeed, let $P(n)$ now mean that $n \geq 1$. As before, we again carry out the two inductive steps.

(i) $P(1)$ *holds*; for if $n = 1$, then certainly $n \geq 1$; so $P(n)$ holds for $n = 1$.

(ii) $P(m) \implies P(m+1)$. We make the inductive hypothesis that $P(m)$ holds for some *particular* m. This means that $m \geq 1$. Hence, by monotonicity of addition and transitivity (Axioms II and VIII), we have $m+1 \geq 1+1 > 1$ (the *latter* follows by adding 1 on both sides of $1 > 0$). Thus $m+1 > 1$ and certainly $m+1 \geq 1$, that is, $P(m+1)$ holds (if $P(m)$ does). Induction is complete.

(D) *In an* ordered *field,* $m, n \in N$ *and* $m > n$ *implies* $m - n \in N$. Indeed, fixing an arbitrary $m \in N$, let $P(n)$ mean "$m - n \leq 0$ or $m - n \in N$." Then we have the following:

(i) $P(1)$ *is true*; for if $n = 1$, then $m-n = m-1$. But, by Example (B), $m - 1 = 0$ or $m - 1 \in N$. This shows that $P(n)$ holds for $n = 1$; $P(1)$ is true.

(ii) $P(k) \implies P(k+1)$. Suppose $P(k)$ holds for some *particular* $k \in N$. This means that

$$m - k \leq 0 \text{ or } m - k \in N.$$

By Example (B), it easily follows that either

$$(m-k) - 1 \leq 0 \text{ or } (m-k) - 1 \in N;$$

that is,

$$\text{either } m - (k+1) \leq 0 \text{ or } m - (k+1) \in N.$$

But this shows that $P(k+1)$ holds (if $P(k)$ does). By induction, then, $P(n)$ holds for every $n \in N$; that is,

$$\text{either } m - n \leq 0 \text{ or } m - n \in N \text{ for every } n \in N.$$

Lemma. *For* no naturals m, n *in an ordered field is* $m < n < m+1$.

For, by Example (D), $n > m$ would imply $n - m \in N$, hence $n - m \geq 1$ (by Example (C)). But $n - m \geq 1$, or $n \geq m+1$, excludes $n < m+1$ (trichotomy). Thus $m < n < m+1$ is impossible for naturals.

Theorem 2. *In an ordered field, every nonempty subset of* N *(the naturals) has a* least *element, i.e., one not exceeding any other of its members.*[6]

***Proof.** Given $\emptyset \neq A \subseteq N$, we want to show that A has a *least* element. To do this, let

$$A_n = \{x \in A \mid x \leq n\} \quad n = 1, 2 \ldots .$$

That is, A_n consists of those elements of A that are $\leq n$ (A_n may be empty). Now let $P(n)$ mean

$$\text{"}\textit{either } A_n = \emptyset \textit{ or } A_n \textit{ has a least element.}\text{"} \tag{1}$$

We show by induction that $P(n)$ holds for each $n \in N$. Indeed, we have the following:

(i) $P(1)$ *is true*; for, by construction, A_1 consists of all naturals from A that are ≤ 1 (if any). But, by Example (C), the only such natural is 1. Thus A_1, if not empty, consists of 1 *alone*, and so 1 is also its *least* member. We see that either $A_1 = \emptyset$ or A_1 has a least element; i.e., $P(1)$ is true.

(ii) $P(m) \implies P(m+1)$. Suppose $P(m)$ holds for some *particular* m. This means that $A_m = \emptyset$ or A_m has a least element (call it m_0). In the *latter* case, m_0 is also the least member of A_{m+1}; for, by the lemma, A_{m+1} differs from A_m by the element $m+1$ *at most*, which is *greater* than all members of A_m.

If, however, $A_m = \emptyset$, then for the same reason, A_{m+1} (if $\neq \emptyset$) consists of $m+1$ *alone*; so $m+1$ is also its least element.

This shows that $P(m+1)$ is true (if $P(m)$ is). Thus the inductive proof is complete, and (1) holds for *every* A_n.

Now, by assumption, $A \neq \emptyset$; so we fix some $n \in A$. Then the set

$$A_n = \{x \in A \mid x \leq n\}$$

contains n, and hence $A_n \neq \emptyset$. Thus by (1), A_n must have a least element m_0, $m_0 \leq n$. But A differs from A_n only by elements $> n$ (if any), which are all $> m_0$. Thus m_0 is the desired *least* element of A as well. □

Theorem 2 yields a new form of the induction law for *ordered* fields.

Theorem 2′ (Second induction law)**.** *A proposition* $P(n)$ *holds for* each *natural* n *in an ordered field if*

(i′) $P(1)$ *holds, and*

(ii′) *whenever* $P(n)$ *holds for all naturals* n *less than some* $m \in N$, *it also holds for* $n = m$.

[6] This is the so-called *well-ordering property* of N. A simpler proof for E^1 will be given in §10. Thus the present proof may be omitted.

***Proof.** We use a so-called *indirect proof* or *proof by contradiction*. That is, instead of proving our assertion directly, we shall show that the *opposite* is false, and so our theorem must be true.

Thus assume (i′) and (ii′) and, seeking a contradiction, suppose $P(n)$ *fails* for some $n \in N$ (call such n "bad"). Then these "bad" naturals form a nonempty subset of N, call it A. By Theorem 2, A has a *least* member m. Thus m is the *least* natural for which $P(n)$ fails. It follows that all n *less* than m do satisfy $P(n)$ (among them is 1 by (i′)). But then, by our assumption (ii′), $P(n)$ also holds for $n = m$, which is impossible since m is "bad" by construction.

This contradiction shows that there cannot be any "bad" naturals, and the theorem is proved. □

Note. In inductive proofs, Theorem 2′ is used in much the same manner as Theorem 1′, but it leaves us more freedom in step (ii): instead of assuming that just $P(m)$ is true, we may assume that $P(1)$, $P(2)$, ..., $P(m-1)$ are true.

Problem. Verify Example (A) for mn. (See other problems in §6.)

§6. Induction (continued)

A similar induction law applies to *definitions*. It reads as follows.

A notion $C(n)$ is regarded as defined for every natural element of an ordered field F if

(i) *it has been defined for $n = 1$, and*

(ii) *some rule or formula is given that expresses $C(n)$ in terms of $C(1)$, $C(2)$, ..., $C(n-1)$, i.e., in terms of all $C(k)$ with $k < n$, or some of them.*

Such definitions are referred to as *inductive* or *recursive*. Step (ii), i.e., the rule that defines $C(n)$ in terms of all $C(k)$, $k < n$, or some of them, is called the *recursive* part of the definition. We have already encountered such definitions in Chapter 1, §8. The underlying intuitive idea is again a step-by-step procedure: first, we define $C(1)$; then, once $C(1)$ is known, we may use it to define $C(2)$; next, once both $C(1)$ and $C(2)$ are known, we may use them to define $C(3)$, and so on. The admissibility of inductive definitions can be proved rigorously,[1] in much the same manner as it was done in §5 for inductive *proofs*; however, we shall not go deeper into that problem.

The variable n in a recursive definition may run over the natural elements of any ordered field under consideration. However, for simplicity, we shall use only those inductive definitions in which n ranges over the natural elements *of* E^1, i.e., natural *numbers*. (Actually, this is no restriction; for, as we shall

[1] Cf., e.g., P. Halmos, *Naive Set Theory*, D. Van Nostrand.

show in §14, the natural elements in all ordered fields have exactly the same mathematical properties and may be "identified" with the natural numbers in E^1.) The expression $C(n)$ itself need not denote a number; it may be of quite arbitrary nature.

We shall now illustrate this procedure by several important examples of inductive definitions to be used throughout our later work.

Definition 1.

Given an element x of a field F, we define the *n-th power of x*, denoted x^n, for every natural number $n \in E^1$ ($n = 1, 2, 3, \dots$) by setting

$$\text{(i) } x^1 = x \text{ and (ii) } x^n = x^{n-1}x, \quad n = 2, 3, \dots .$$

By the inductive law expressed above, x^n is defined for every natural n. Intuitively, we may think of it as a step-by-step definition:

$$x^1 = x,\ x^2 = x^1x = xx,\ x^3 = x^2x = (xx)x = xxx,$$

and so on, indefinitely. Thus, formulas (i) and (ii) actually replace an infinite sequence of definitions, obtained consecutively by setting $n = 2, 3, 4, \dots$ in (ii) and substituting the value of x^{n-1} known from the preceding step.

If $x \neq 0$, we also define $x^0 = 1$ and $x^{-n} = \dfrac{1}{x^n}$, $n = 1, 2, \dots$ (division makes sense if $x \neq 0$). The expression 0^0 remains undefined.

Definition 2.

For every natural number n, we define recursively the expression $n!$ (read "n factorial") as follows:

$$\text{(i) } 1! = 1;\ \text{(ii) } n! = (n-1)! \cdot n, \quad n = 2, 3, \dots .$$

Thus, e.g., $2! = (1!) \cdot 2 = 2$; $3! = (2!) \cdot 3 = 6$, etc. We also define $0! = 1$.

Definition 3.

The sum and product of n elements $x_1, \dots, x_n \in F$ of a field, denoted by

$$x_1 + x_2 + \cdots + x_n \text{ and } x_1 \cdot x_2 \cdots x_n$$

(or $\sum_{k=1}^{n} x_k$ and $\prod_{k=1}^{n} x_k$), respectively, are defined recursively as follows:

Sums: $$\text{(i) } \sum_{k=1}^{1} x_k = x_1; \quad \text{(ii) } \sum_{k=1}^{n} x_k = \left(\sum_{k=1}^{n-1} x_k\right) + x_n, \quad n = 2, 3, \dots ;$$

Products: $$\text{(i) } \prod_{k=1}^{1} x_k = x_1; \quad \text{(ii) } \prod_{k=1}^{n} x_k = \left(\prod_{k=1}^{n-1} x_k\right) \cdot x_n, \quad n = 2, 3, \dots .$$

Note. If $x_1 = x_2 = \cdots = x_n = x$, we write nx for $\sum_{k=1}^{n} x_k$. Observe that here $n \in E^1$, while $x \in F$; thus nx is not, in general, a *product*, as defined in F. However, if $F \subseteq E^1$, nx coincides with the ordinary product in E^1 (cf. Problem 13).

Induction can be used to define the notion of an *ordered n-tuple* if the concept of an *ordered pair* is assumed to be known. In fact, an ordered triple can be regarded as an *ordered pair* of the form

$$((x_1, x_2), x_3),$$

that is, as a pair in which the left coordinate is itself a pair. Similarly, an ordered quadruple is a pair

$$((x_1, x_2, x_3), x_4)$$

in which the left coordinate is an ordered *triple* (x_1, x_2, x_3), and so on. This leads to the following definition.

Definition 4.

For any objects $x_1, x_2, \ldots, x_n$, the *ordered n-tuple* $(x_1, \ldots, x_n)$ is defined by

(i) $(x_1) = x_1$ (i.e., an ordered "*one-tuple*" (x_1) is x_1 itself);

(ii) $(x_1, \ldots, x_n) = ((x_1, \ldots, x_{n-1}), x_n)$, $n = 2, 3, 4, \ldots$.

Accordingly, we may now also define the Cartesian product

$$A_1 \times A_2 \times \cdots \times A_n$$

of n sets (see the end of §4 in Chapter 1) either as the set of all n-tuples $(x_1, \ldots, x_n)$ such that $x_k \in A_k$, $k = 1, 2, \ldots, n$, or directly by induction: Assuming the definition is known for two factors and writing $\prod_{k=1}^{n} A_k$ for $A_1 \times A_2 \times \cdots \times A_n$, we define

$$\text{(i)} \quad \prod_{k=1}^{1} A_k = A_1 \quad \text{and} \quad \text{(ii)} \quad \prod_{k=1}^{n} A_k = \left(\prod_{k=1}^{n-1} A_k\right) \times A_n, \quad n = 1, 2, \ldots .$$

Sometimes we start an inductive proof or definition not with $n = 1$ but with $n = 0$ or with $n = 2$, say. For example, Definition 2 could be stated thusly:

$$\text{(i) } 0! = 1; \quad \text{(ii) } n! = (n-1)! \cdot n, \; n = 1, 2, \ldots .$$

Formula (ii) may also be written as follows:

$$\text{(ii) } (n+1)! = n! \cdot (n+1), \; n = 0, 1, 2, \ldots ;$$

similarly in other cases of this kind.

Note. The notion of an ordered n-tuple as defined above *differs* from that of a *finite sequence* (cf. Chapter 1, §8, Definition 1). However, for all practical purposes, both behave in the same way; namely, two sequences, or two n-tuples, are the same iff the *corresponding* terms coincide (cf. Problem 16 below). Therefore, in most cases, we may "forget" about the difference between the two concepts.

Problems on Natural Numbers and Induction

1. Using induction (Theorem 1′ in §5), prove the following:

(i) $1^n = 1$ in any field;

(ii) $(\forall n \in N)\; 2^n \geq 2$ in any ordered field; specify the proposition $P(n)$.

2. Prove that if $x_1, \ldots, x_n$ are natural elements of a field, so are

$$\sum_{k=1}^{n} x_k \text{ and } \prod_{k=1}^{n} x_k.$$

Assume this known for $n = 2$, and use induction on n.

3. Prove that the sum and product of n elements of an ordered field are positive if all these elements are. (Use induction on n.)

4. Prove by induction that if $x_1, x_2, \ldots, x_n$ are nonzero elements of a field, so is $\prod_{k=1}^{n} x_k$; and

$$\left(\prod_{k=1}^{n} x_k\right)^{-1} = \prod_{k=1}^{n} x_k^{-1}.$$

Assume this known for $n = 2$.

5. Use induction over n to prove that for any field elements c, x_k and y_k:

$$\text{(i)} \quad c\left(\sum_{k=1}^{n} x_k\right) = \sum_{k=1}^{n} cx_k; \qquad \text{(ii)} \quad \sum_{k=1}^{n} (x_k \pm y_k) = \sum_{k=1}^{n} x_k \pm \sum_{k=1}^{n} y_k.$$

6. Prove by induction that in any ordered field

$$\left|\sum_{k=1}^{n} x_k\right| \leq \sum_{k=1}^{n} |x_k|.$$

7. Prove that in any ordered field, $a < b$ iff $a^n < b^n$, provided $a, b \geq 0$. Infer that $a^n < 1$ if $0 \leq a < 1$; $a^n > 1$ if $a > 1$ $(n = 1, 2, \ldots)$.

8. Use induction over n to prove that for any element ϵ of an ordered field F,

$$\text{(i) } (1+\epsilon)^n \geq 1 + n\epsilon \text{ if } \epsilon > -1; \quad \text{(ii) } (1-\epsilon)^n \geq 1 - n\epsilon \text{ if } \epsilon < 1$$

(Bernoulli inequalities). Infer that $2^n > n$, $n = 1, 2, \ldots$, in E^1.

9. Prove that in any field,

$$a^{n+1} - b^{n+1} = (a-b) \cdot \sum_{k=0}^{n} a^k b^{n-k}, \quad n = 1, 2, \ldots.$$

10. Prove in E^1,

(i) $1 + 2 + \cdots + n = \frac{1}{2} n(n+1)$;

(ii) $\sum_{k=1}^{n} k^2 = \frac{1}{6} n(n+1)(2n+1)$;

(iii) $\sum_{k=1}^{n} k^3 = \frac{1}{4} n^2 (n+1)^2$;

(iv) $\sum_{k=1}^{n} k^4 = \frac{1}{30} n(n+1)(2n+1)(3n^2+3n-1)$.

11. For any field elements a, b and natural numbers $m, n \in E^1$, prove the following:

(i) $a^m a^n = a^{m+n}$; (ii) $(a^m)^n = a^{mn}$; (iii) $(ab)^n = a^n b^n$.

If $a \neq 0$, then also

(iv) $\frac{a^n}{a^m} = a^{n-m}$; (v) $\left(\frac{b}{a}\right)^n = \frac{b^n}{a^n}$.

If $a, b \neq 0$ show that these laws hold for negative exponents, too. Also, prove the following:

(vi) $ma + na = (m+n)a$; (vii) $ma \cdot nb = (mn)(ab)$;

(viii) $n(a \pm b) = na \pm nb$.

[Hints: Fix m and use induction on n. The "natural multiples" nx can be defined inductively by $1 \cdot x = x$, $nx = (n-1)x + x$, $n = 1, 2, \ldots$.]

11′. Show by induction that each natural element x of an ordered field F can be uniquely represented as $x = n \cdot 1'$, where n is a natural *number* in E^1 ($n \in N$) and $1'$ is the unity in F; that is, x is the sum of n unities.

Conversely, show that each such $n \cdot 1'$ is a natural element of F.

Finally, show that, for $m, n \in N$, we have

$$m < n \text{ iff } mx < nx,$$

provided $x > 0$.

12. Define the *binomial coefficient*

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

for nonnegative integers n, k ($k \leq n$) in E^1. Verify *Pascal's law*:

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

Using it, prove inductively that $\binom{n}{k}$ is always a natural number. Then establish inductively the *binomial theorem*: for elements a, b of any field F and any natural number n,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

13. Show by induction that if $x_1 = x_2 = \cdots = x_n = x$, then

$$\sum_{k=1}^{n} x_k = nx \text{ and } \prod_{k=1}^{n} x_k = x^n \text{ (where } x \text{ is in any field).}$$

14. Show by induction that in any field

$$\sum_{k=1}^{n} (x_k - x_{k-1}) = x_n - x_0.$$

Deduce from it the formulas of Problem 10 directly.

[Hints: For Problem 10(i), take $x_k = k^2$. For Problem 10(ii), take $x_k = k^3$, etc. Substitute and simplify.]

15. Show by induction that every *finite* sequence $x_1, x_2, \ldots, x_n$ of elements of an ordered field contains a largest and a smallest term (which need *not* be x_n and x_1 since the sequence is not necessarily monotonic). Show by examples that the theorem fails for infinite sequences. Infer that the set of *all* natural numbers 1, 2, 3, ... is infinite. (For the definition of "finite" and "infinite", see Chapter 1, §8).

16. Prove by induction that two ordered n-tuples

$$(x_1, \ldots, x_n) \text{ and } (y_1, \ldots, y_n)$$

are equal iff $x_1 = y_1$, $x_2 = y_2, \ldots, x_n = y_n$. Assume this known for $n = 2$.

17. Show that if the sets A and B are finite (cf. Chapter 1, §8, Definition 5), so are $A \cup B$ and $A \times B$. By induction, prove this for n sets.

18. Solve Problems 6 and 7 of Chapter 1, §9 *by induction*.

19. Show *by induction* that if the finite sets A and B have m and n elements, respectively, then

(i) $A \times B$ has mn elements;

(ii) A has 2^m subsets;

(iii) If further $A \cap B = \emptyset$, then $A \cup B$ has $m + n$ elements.

20. Prove the *division theorem*: Let $N' = N \cup \{0\}$ be the set consisting of 0 and all naturals (N) in an ordered field. Then for any $m, n \in N'$ ($n > 0$), there is a unique pair $(q, r) \in N' \times N'$ such that

$$m = nq + r \text{ and } 0 \leq r < n$$

(q and r are called, respectively, the *quotient* and *remainder* from the division of m by n). If $r = 0$, we say that n *divides* m and write $n \mid m$). [Hints: Let q be the least element of

$$A = \{x \in N' \mid (x + 1)n > m\}$$

(why does it exist?) and put $r = m - nq$; show that $r \in N'$, $r < n$, using the fact that $q \in A$.

To prove uniqueness, let (q', r') be another such pair and show that the assumption $r < r'$ or $r' > r$ leads to a contradiction; thus $r = r'$, and hence $q = q'$.]

§7. Integers and Rationals

Definition 1.

All naturals in a field F, their additive inverses, and the zero element 0 are called the *integral elements* or *integers* (in F).

Below we denote by J the set of all integers in F and by N the set of all naturals, as before. In order to investigate J, we need a lemma.

Lemma. *If $m, n \in N$ in a field F, then $m - n$ is an integer in F ($m - n \in J$).*

Proof. We proceed by induction.[1] Fix $m \in N$, and let $P(n)$ mean $m - n \in J$.

(i) $P(1)$ *is true*. Indeed, $m - 1 = 0$ or $m - 1 \in N$ by Example (B) in §5. Thus $m - 1 \in J$, by definition. But this means that $P(n)$ holds for $n = 1$.

(ii) $P(k) \implies P(k+1)$. Suppose $P(n)$ holds for some particular $n = k$. This means that $m - k \in J$; that is, $m - k \in N$ or $m - k = 0$ or $-(m - k) \in N$. We must show that this implies $[m - (k + 1)] \in J$, i.e., $[(m - k) - 1] \in J$.

[1] If F is an *ordered* field, one can simply apply Example (D) in §5. Indeed, we have $m - n \in N$, $m - n = 0$, or $-(m - n) \in N$ accordingly as $m > n$, $m = n$, or $m < n$. Thus $m - n \in J$ by definition. This may suffice at a first reading.

Now, if $m - k \in N$, then $(m - k) - 1 = 0$ or $(m - k) - 1 \in N$ by Example (B) in §5. Hence $(m - k) - 1 \in J$, as required. This settles the case $m - k \in N$.

If $m - k = 0$, then $(m - k) - 1 = -1 \in J$ by definition.

Finally, if $-(m - k) \in N$, then $-(m - k) + 1 \in N$; that is, $-[m - (k + 1)] \in N$, and so, by definition, $[m - (k + 1)] \in J$. But this means that $P(k + 1)$ is true.

Thus, in all three cases, $P(k + 1)$ results from $P(k)$. This completes the induction, and so $P(n)$ holds for *every* $n \in N$, i.e., $m - n \in J$ for any $m, n \in N$. □

Theorem 1. *If x and y are integers in a field F, so are $x + y$ and xy.*[2]

Proof. As $x, y \in J$, we must consider the following possible cases.

(i) If x and y are both naturals, so are $x + y$ and xy by Example (A) in §5. Thus they are integers, as claimed.

(ii) If x or y is 0, all is trivial (we leave this case to the reader).

(iii) If x and y are both additive inverses of naturals, then $-x$ and $-y$ are *naturals*; hence so is their sum, $(-x) + (-y) = -(x + y)$. This shows that $x + y$ is the additive inverse of a natural element; so $x + y \in J$ by definition. Similarly $xy = (-x)(-y) \in N$; hence certainly $xy \in J$.

(iv) Suppose that one of x and y (say x) is a natural element while the other (y) is not. Then either $y = 0$ or $-y \in N$. The case $y = 0$ was taken care of in (ii). If, however, $-y \in N$, the lemma yields $x - (-y) \in J$; that is, $x + y \in J$, as claimed. Also, $x(-y) \in N$. Hence xy is an integer, being the additive inverse of the natural element $x(-y) = -xy$.

Thus, in all cases, $x + y \in J$ and $xy \in J$. The theorem is proved. □

We also have an induction rule for integers similar to that applying to natural elements.

Induction Law for Integers. *A proposition $P(n)$ holds for all integers n greater than a fixed integer p in an ordered field if*

(i′) $P(n)$ *holds for $n = p + 1$, and*

(ii′) *whenever $P(n)$ holds for all integers n such that $p < n < m$, then $P(n)$ also holds for $n = m$ ($m \in J$).*

This is proved from Theorem 2′ in §5 by substituting $x - p$ for n and noting that $x - p$ runs over all *natural* values when x takes on integral values greater than p. (Here we say that "induction *starts with* $p + 1$.")

Definition 2.

An element x of a field F is said to be *rational* iff $x = p/q$ for some integral elements p and q, with $q \neq 0$.[3]

[2] So also is $x - y$ since it reduces to $x + (-y)$, where x and $-y$ are integers.

[3] In particular, the rationals in E^1 are called rational *numbers*.

Theorem 2. *The sum, the difference, and the product of two rationals x and y in a field F are rational. So also is x/y if $y \neq 0$.*

Proof. Let $x = p/q$ and $y = r/s$, where p, q, r, s are integers, with q and s different from 0. Then, as is easily seen (cf. Problem 3 in §4),

$$x \pm y = \frac{ps \pm qr}{qs}, \quad xy = \frac{pr}{qs}, \quad \text{and} \quad \frac{x}{y} = \frac{ps}{qr}$$

(the latter provided that y and r, too, are different from zero). Thus $x \pm y$, xy, and x/y can be written as fractions with integral numerators and denominators. (The fact that numerators and denominators are integers follows from Theorem 1. It is also easily seen that these denominators are not 0 since q, r, $s \neq 0$.) By Definition 2, they are rational elements of F, as required. □

It follows, in particular, that $-x$ is rational whenever x is; similarly for $x^{-1} = 1/x$ if $x \neq 0$. All integers (including 0 and 1) are rationals since an integer m can be written as $m/1$.

It is easy to verify that Axioms I to IX remain valid if E^1 is replaced by the set R of all rational elements of an ordered field F. This means that R *is an ordered field*. It is called the *rational subfield* of F.

Problems on Integers and Rationals

1. Prove in detail the induction law for integers, stated above.

2. Show that the result of Problem 20 in §6, i.e., the *division theorem*, holds also with N' replaced by J, the set of all integers.

3. Verify that the set J of all integers in an ordered field F satisfies Axioms I–IX of §2 *except* Axiom V(b). Thus J is not a field.

Structures satisfying Axioms I–IX, except possibly IV(b) and V(b), are called *ordered commutative rings*. In particular, J is such a ring.

4. Verify that the set R of all rationals in F is a field if F is and an ordered field if F is.

5. Show that every rational r in an ordered field F has a unique representation $r = m/n$ *in lowest terms*, i.e., such that $n > 0$ and $|m|$ has *the least* possible value (along with $|n|$). Also prove that, in this case, m and n are *relatively prime*, i.e., have no common divisors > 1.

[Hint: If $r > 0$, let A be the set of all naturals m occurring in various representations $r = m/n$. Then apply Theorem 2 of §5. The rest follows from the minimality of m.]

6. Let A be a nonempty set of integers ($A \subset J$) in an ordered field F. Show that if all elements of A are greater than some integer p, then A has a *least* element.

[Hint: The differences $x - p$ ($x \in A$) are *naturals*; so by Theorem 2 of §5, one of them is the *least*; the corresponding x is the least in A.]

7. Let A be as in Problem 6. Show that if all elements of A are *less* than some integer p, then A has a *largest* element.

[Hint: Apply the result of Problem 6 to the set F of all additive inverses $-x$ of elements $x \in A$, noting that $-x > -p$ for all $x \in A$.]

8. From Problems 6 and 7 infer that in any ordered field, two nonzero integers m and n always have a least common multiple and a greatest common divisor.

[Hint: Show first that all common multiples (such as mn) are $\geq |m|$, while all common divisors are $\leq |m|$.]

9. Prove: Every integer $n > 1$ in an ordered field is the product of some finite sequence of *primes*, i.e., integers ≥ 2, each divisible only by 1 and itself.

[Hint: Let $P(n)$ mean that n can be so factored, and use induction. $P(n)$ is trivial if n is itself a prime (e.g., $n = 2$).

Now suppose $P(n)$ is true for all n less than some m. If m is not a prime, then $m = n_1 n_2$ for some integers n_1 and n_2 greater than 1 but *less than* m (why?); so by our assumption, n_1 and n_2 factor into primes, and the same follows for m.]

Note: It can be shown that the factorization into primes is unique except for the order in which they occur.

10. Show that there are *infinitely* many primes.

[Hint: Seeking a contradiction, suppose all primes can be put in a finite sequence

$$p_1, \ldots, p_n.$$

Then show that

$$1 + \prod_{k=1}^{n} p_k$$

is not divisible by any of the p_k (use the division algorithm theorem; cf. Problem 2). Infer from Problem 9 that $1 + \prod_{k=1}^{n} p_k$ is a prime *different* from all p_k ($k = 1, 2, \ldots, n$).]

11. Show that every strictly decreasing sequence of positive integers is necessarily finite.

[Hint: Use Problem 6 or Theorem 2 in §5.]

§8. Bounded Sets in an Ordered Field

Definition 1.

A subset A of an ordered field F is said to be *bounded below*, or *left-bounded*, if there is an element $p \in F$ such that $(\forall x \in A)\ p \leq x$.

The set A is *bounded above*, or *right-bounded*, if there is an element $q \in F$ such that $(\forall x \in A)\ x \leq q$.

In this case, p and q are called, respectively, a *lower* (or *left*) and an *upper* (or *right*) *bound* of A.

If A is both left- and right-bounded, it is simply referred to as *bounded* (by p and q). The empty set $\emptyset$ is always regarded as bounded, and all elements of F are considered both its lower and upper bounds.

Note. *The bounds p and q may, but need not, belong to the set A.*

If a set A is bounded below, it has *many* lower bounds; for if p is one of them, so also is every element less than p. Similarly, a right-bounded set always has many upper bounds.

All this applies, in particular, to sets of real numbers, i.e., sets in E^1.

Examples.

(1) The set of four numbers $\{1, -2, 3, 7\}$ is *bounded*, both above (e.g., by 7, 8, 9, 100, etc.) and below (e.g., by -2, -5, -12, etc.).

(2) The set of all natural numbers $N = \{1, 2, 3, \dots\}$ is bounded below (e.g., by 1, 0, $-\frac{1}{2}$, -9, etc.) *but not above*. (An exact proof of this fact will be given later, after the introduction of the missing 10th axiom, on which it is based.) On the other hand, the set of all negative integers is bounded above but not below.

(3) The set J of *all* integers has no lower and no upper bounds in E^1. In fact, given any number $p \in E^1$, one can always find an integer $> p$ and an integer $< p$. Thus no such p can be a lower or an upper bound for J.

Geometrically, an upper bound of a set $A \subset E^1$ is a point q on the real axis that lies on the *right* side of A, while a lower bound p lies on the *left* side; see Figure 9.

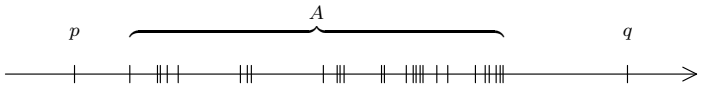


FIGURE 9

An especially important class of bounded sets form the so-called *intervals*.

Definition 2.

Given any real numbers a and b ($a \leq b$), we define

(i) the *open interval* (a, b) to be the set of all real numbers x such that $a < x < b$, i.e.,

$$(a, b) = \{x \in E^1 \mid a < x < b\};$$

(ii) the *closed interval* $[a, b]$ to be the set of all real numbers x such that $a \leq x \leq b$, i.e.,

$$[a, b] = \{x \in E^1 \mid a \leq x \leq b\}.$$

We also define, in a similar way, the *half-open interval* $(a, b]$ and the *half-closed interval* $[a, b)$ by the inequalities $a < x \leq b$ and $a \leq x < b$, respectively. The same definitions also apply to intervals in any ordered field F.

In all cases, a and b are called the *endpoints* of the interval. Note that a belongs to $[a, b]$ and $[a, b)$ but not to (a, b) and $(a, b]$, while b belongs to $[a, b]$ and $(a, b]$ but not to (a, b) and $[a, b)$ (square brackets are written beside those endpoints that are included in the interval). If $a = b$, i.e., if the endpoints coincide, the interval is said to be *degenerate*. In this case the closed interval $[a, a]$ consists of a single point, a, while $(a, a) = (a, a] = [a, a) = \emptyset$. (Why?) Every interval is a *bounded* set since its endpoints are its bounds by its very definition. Geometrically, intervals are segments of the real axis.

If an upper bound q of a set A is itself in A, then q is clearly the *greatest* element of A (i.e., one not exceeded by any other element of A). We then also call it the *maximum* of A, denoted $\max A$. Similarly, if A contains its lower bound p, then p is its *least* element, also called the *minimum* of A or, briefly, $\min A$. *A set A can have at most one maximum and one minimum*; for if, say, q and q' were both maxima, then by definition, $q \leq q'$ (since $q \in A$ and q' is an upper bound) and, similarly, $q' \leq q$, so that $q = q'$. However, *a set may have no maximum and no minimum even if it is bounded*; such a set is, for example, every open interval. (Why?) We denote by $\max(a, b)$ the larger of the two elements a and b; similarly for $\min(a, b)$ and for sets of several elements.

It is important to note that every nonempty *finite* set A in an ordered field must have a maximum and a minimum. This is easily proved by induction on the number n of elements in A; the details are left to the reader (cf. Problem 15 in §6). In particular, given n real numbers $x_1, x_2, \dots, x_n$, one of them must be the largest, i.e., $\max(x_1, \dots, x_n)$, and one of them must be the smallest, i.e., $\min(x_1, \dots, x_n)$.

§9. The Completeness Axiom. Suprema and Infima

In §8 it was shown that a right-bounded set of real numbers always has *many* upper bounds. The question arises as to whether or not there exists among them a *least* one. Similarly, one may ask whether or not a left-bounded set always has a *greatest lower bound*, i.e., one "closest" to the set.

Geometrically, this problem may be illustrated as follows. Figure 10 shows a bounded set M of real numbers plotted on the real axis.

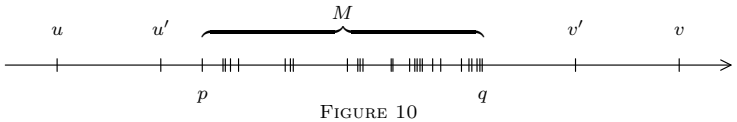


FIGURE 10

The points u and v on the axis represent a lower and an upper bound of M, respectively. It is, however, evident from Figure 10 that v is not the *least* upper bound since also the smaller number v' is an upper bound of M. Similarly, u is not the *greatest* lower bound since there is a greater lower bound, u'.

Now imagine that the point v moves along the axis in the direction of the set M but remaining to the right of all points of M. It is geometrically evident that v will eventually arrive at a certain position q where it can no longer continue its motion without passing some points of M, i.e., without ceasing to be an upper bound of M. This very position q (if it actually exists) is clearly that of the *least* upper bound. Similarly, by moving the point u in the positive direction, one arrives at a position p that corresponds to the greatest lower bound of M. Note that p and q need not be the minimum and maximum of M. For example, if M is the open interval (p, q), it has no minimum or maximum at all. Nevertheless, p and q are its greatest lower, and least upper, bounds. (To fix ideas, assume that M in Figure 10 has no maximum or minimum).

These geometric considerations, however plausible, cannot be considered a rigorous proof of the existence of the least upper and greatest lower bounds. This proof also cannot be derived from the nine axioms stated thus far. On the other hand, the existence of the least upper and greatest lower bounds is of very great importance for the entire mathematical analysis. Therefore, it has to be introduced as a special axiom, which, for reasons to be explained later, is called the *completeness axiom*. It is the last (tenth) axiom in our system.

Completeness Axiom.

X *Every nonvoid right-bounded set M of real numbers has a least upper bound* (also called the *supremum of M*, abbreviated $\sup M$ or l.u.b. M).

No special axiom is needed for lower bounds since the corresponding proposition can now be proved from the completeness axiom, as follows.

Theorem 1. *Every nonvoid left-bounded set M of real numbers has a greatest lower bound* (also called the *infimum of M,* abbreviated $\inf M$ or g.l.b. M).

Proof. Let B denote the (nonvoid) set of all lower bounds of M (such bounds exist since M is left-bounded). Clearly, each element of M is, in turn, an upper bound for B (because no element of B can exceed any element of M by the definition of a lower bound). Thus B is nonvoid and right-bounded. By the completeness axiom, B has a supremum, call it p. We shall now prove that p is also the required infimum of M. Indeed, we have the following:

(i) *p is a lower bound of M*; for p is, by definition, the *least* of all upper bounds of B. But, as we have seen, all elements of M are such upper bounds; so p cannot exceed anyone of them, as required.

(ii) *p is the greatest lower bound of M*. In fact, as p is an upper bound of B, it is not exceeded by any element of B. But, by definition, B contains *all* lower bounds of M; so p is not exceeded by any one of them.

This completes the proof. □

Note 1. Theorem 1 could, in turn, be assumed as an axiom. Then our completeness axiom could be *deduced* from it in a similar manner.

Note 2. The supremum and infimum of a set M (if they exist) are *unique*; for the infimum of M is, by definition, the greatest element of the set B of all lower bounds of M, i.e., $\max B$. But $\max B$ is unique, as shown at the end of §8; hence so is $\inf M$. Similarly for $\sup M$.

Note 3. To explain the "completeness axiom", consider again Figure 10 and imagine that the points p and q have been removed from the axis, leaving two "gaps" in it. Then the set M, though bounded, would have no supremum and no infimum since the required points would be missing. The completeness axiom asserts, in fact, that such "gaps" never occur, i.e., that the real axis is "complete".

As we mentioned, the completeness axiom is independent of the first nine axioms, i.e., cannot be deduced from them. In fact, there are ordered fields that do not satisfy it, though they certainly satisfy the first nine axioms. Such a field is, e.g., the field of all rational numbers (see §11). On the other hand, some ordered fields do have the completeness property, and E^1 is one of them. This justifies the following definition.

Definition.

An ordered field F is said to be *complete* iff every nonvoid right-bounded subset M of F has a supremum (i.e., a least upper bound) in F.

In particular, *E^1 is a complete ordered field* by the completeness axiom. We can now restate Theorem 1 in a more general form:

Theorem 1′. *In a complete ordered field F, every nonvoid left-bounded set $M \subset F$ has an infimum (i.e., a greatest lower bound).*

The proof is exactly the same as in Theorem 1.

Also the following corollaries will be stated for ordered fields in general. They apply, of course, to E^1 as well.

Corollary 1. *An element q of an ordered field F is the supremum of a set $M \subset F$ iff q satisfies these two conditions*:

(i) $(\forall x \in M)\ x \leq q$; *i.e., q is not exceeded by any element x in M.*

(ii) *Every element $p < q$ is exceeded by some x in M, i.e.,*

$$(\forall p < q)\ (\exists x \in M) \quad p < x.$$

A similar result holds for the infimum (with all inequalities reversed).

In fact, condition (i) states that q is an upper bound of M, while (ii) states that no *smaller* element $p \in F$ is such a bound (since it is exceeded by some $x \in M$). When combined, (i) and (ii) mean that q is the *least* upper bound.

Note 4. Every element $p < q$ can be written as $q - \epsilon$, where $\epsilon > 0$. Hence Condition (ii) in Corollary 1 can also be rephrased thusly:

(ii′) *For every field element $\epsilon > 0$, there is an $x \in M$ with $q - \epsilon < x$.*

In case $q = \inf M$, we have instead that

$$(\forall \epsilon > 0)\ (\exists x \in M) \quad q + \epsilon > x.^{1}$$

Corollary 2. *Let M be a nonempty set in an ordered field F, and let $b \in F$. If each element x of M satisfies the inequality $x \le b$ ($x \ge b$), so does $\sup M$ ($\inf M$, respectively), provided that $\sup M$ ($\inf M$) exists.*

In fact, the condition

$$(\forall x \in M) \quad x \le b$$

means that b is an upper bound of M. But $\sup M$ is the *least* upper bound of M, so $(\sup M) \le b$; similarly for $\inf M$.

Corollary 3. *If A and B are subsets of an ordered field, both nonvoid, and if $A \subseteq B$, then*

$$\sup A \le \sup B \text{ and } \inf A \ge \inf B,$$

provided that the suprema and infima involved exist. (Thus if new elements are added to a set A, its supremum cannot decrease and its infimum cannot increase.)

Proof. Let

$$p = \sup A \text{ and } q = \sup B.$$

As q is an upper bound of B, we have $x \le q$ for each $x \in B$. But, by assumption, B contains all elements of A. Hence, the inequality $x \le q$ holds also for each $x \in A$ (since $x \in B$ as well). As each $x \in A$ satisfies $x \le q$, Corollary 2 yields $\sup A \le q$, i.e.,

$$\sup A \le \sup B$$

(for $q = \sup B$); similarly for infima. □

[1] Here we may assume ϵ as small as we like (only $\epsilon > 0$); for if the required inequalities hold for a small ϵ, they certainly hold for any larger ϵ.

Note 5. If A is a *proper* subset of B ($A \subset B$), it does *not* follow that $\sup A < \sup B$, but only that $\sup A \le \sup B$ (and $\inf A \ge \inf B$). For example, the open interval (a, b) is a proper subset of the closed interval $[a, b]$, but their suprema and infima are the same, namely b and a. Similarly, if in Corollary 2 each $x \in M$ satisfies $x < b$ ($x > b$), it only follows that $\sup M \le b$ ($\inf M \ge b$), but not $\sup M < b$ ($\inf M > b$). For example, we have $x < b$ for all $x \in (a, b)$, but $\sup(a, b) = b$.

Corollary 4. *If a subset M of an ordered field F has a maximum q, then q is also its supremum. Similarly, the minimum of M (if it exists) is its infimum. The converse statements are, however, not true.*

The proof (which is obvious) is left to the reader.

Problems on Bounded Sets, Infima, and Suprema

1. Assume Theorem 1 as an *axiom* and deduce from it the completeness axiom.

2. Complete the proofs of Corollaries 1–3 (for infima) and Corollary 4.

3. Show that if $\inf A$ and $\sup A$ exist in an ordered field, then $\inf A \le \sup A$.

4. Prove that the endpoints of an open interval (a, b) ($a < b$) in an ordered field F are the infimum and supremum of (a, b).

5. In an ordered field F, let $A \subset F$ ($A \neq \emptyset$), and let cA denote the set of all products cx ($x \in A$) for some fixed element $c \in F$; so

$$cA = \{cx \mid x \in A\}.$$

Prove the following:

(i) If $c \ge 0$, then

$$\sup(cA) = c \cdot \sup A \text{ and } \inf(cA) = c \cdot \inf A,$$

provided that $\sup A$ (in the first formula) and $\inf A$ (in the second formula) exist.

(ii) If $c < 0$, then

$$\sup(cA) = c \cdot \inf A \text{ and } \inf(cA) = c \cdot \sup A,$$

provided again that $\inf A$ and $\sup A$ (as the case may be) exist. What if $c = -1$?

6. From Problem 5(ii), with $c = -1$, obtain a new proof of Theorem 1.
[Hint: If M is bounded below, show that $(-1)M$ is bounded above, then take its sup.]

7. Let A and B be subsets of an ordered field F. Assuming that the required l.u.b. and g.l.b. exist in F, prove the following:

(i) If $(\forall x \in A)\ (\forall y \in B)\ x \leq y$, then $\sup A \leq \inf B$.
[Hint: Each $y \in B$ is an upper bound of A and, hence, cannot be less than the *least* upper bound of A. Thus $(\forall y \in B)\ \sup A \leq y$, i.e., $\sup A$ is a lower bound of B, and so $\sup A \leq \inf B$ (cf. Corollary 2).]

(ii) If $(\forall x \in A)\ (\exists y \in B)\ x \leq y$, then $\sup A \leq \sup B$.

(iii) If $(\forall y \in B)\ (\exists x \in A)\ x \leq y$, then $\inf A \leq \inf B$.

(iv) If B consists of *all* upper bounds of A, then $\sup A = \inf B$.

8. In an ordered field F, let $A + B$ denote the set of all sums $x + y$, with $x \in A$ and $y \in B$ ($A \subseteq F$, $B \subseteq F$); so

$$A + B = \{x + y \mid x \in A,\ y \in B\}.$$

Prove that if $\sup A = p$ and $\sup B = q$ exist in F, then $p + q = \sup(A + B)$; similarly for infima.
[Hint: By Corollary 1 and Note 4, we must show (in the case of sup) that

(i) $(\forall x \in A)\ (\forall y \in B)\ x + y \leq p + q$ (which is easy), and

(ii′) $(\forall \epsilon > 0)\ (\exists x \in A$ and $y \in B)\ x + y > (p + q) - \epsilon$.

For (ii′), take any $\epsilon > 0$. By Note 4, there are $x \in A$ and $y \in B$, with

$$x > p - \frac{1}{2}\epsilon \text{ and } y > q - \frac{1}{2}\epsilon.$$

(Why?) Then

$$x + y > (p - \frac{1}{2}\epsilon) + (q - \frac{1}{2}\epsilon) = (p + q) - \epsilon,$$

as required.]

9. Continuing Problem 8, let A and B consist of *positive* elements only, and let

$$AB = \{xy \mid x \in A,\ y \in B\}.$$

Prove that if $\sup A = p$ and $\sup B = q$ exist in F, then $pq = \sup(AB)$; similarly for infima.
[Hint: Using Note 4, we may take $\epsilon > 0$ so small that

$$\frac{\epsilon}{p + q} < p, q;$$

take

$$x > p - \frac{\epsilon}{p + q} > 0 \text{ and } y > q - \frac{\epsilon}{p + q} > 0$$

and show that

$$xy > pq - \epsilon + \frac{\epsilon^2}{(p + q)^2} > pq - \epsilon.$$

For $\inf(AB)$, let $s = \inf B$, $r = \inf A$, $\epsilon > 0$. By density, there is $d < 1$, with

$$0 < d < \frac{\epsilon}{1 + r + s}.$$

Now take $x \in A$ and $y \in B$ with $x < r + d$, $y < s + d$, and show that $xy < rs + \epsilon$.]

10. Prove that if $a \geq b - \epsilon$ for *all* $\epsilon > 0$, then $a \geq b$. What if $(\forall \epsilon > 0)$ $a \leq b + \epsilon$?

***11.** Prove the *principle of nested intervals*: If $[a_n, b_n]$ are closed intervals in a *complete* field F, with

$$[a_n, b_n] \supseteq [a_{n+1}, b_{n+1}], \quad n = 1, 2, 3, \ldots,$$

then

$$\bigcap_{n=1}^{\infty} [a_n, b_n] \neq \emptyset.$$

[Hint: Let $A = \{a_1, a_2, \ldots, a_n, \ldots\}$. Show that A is right-bounded by each b_n. By completeness, let $\sup A = p$. Show that $a_n \leq p \leq b_n$, i.e.,

$$p \in [a_n, b_n], \quad n = 1, 2, \ldots,$$

and so

$$p \in \bigcap_{n=1}^{\infty} [a_n, b_n].]$$

12. Prove by induction that any union of finitely many bounded sets in an ordered field F is itself bounded in F (first prove it for *two* sets).

13. Prove that for any bounded subset $A \neq \emptyset$ of a complete ordered field F, there is a *smallest* closed interval C containing A ("smallest" means that C is a subset of any other such interval). Is this true with "closed" replaced by "open"?
[Hint: Let $C = [a, b]$, $a = \inf A$, $b = \sup A$.]

§10. Some Applications of the Completeness Axiom

From everyday experience, one knows that even a large distance y can be measured by a small yardstick x; one only has to mark x off sufficiently many times. This fact was noticed by ancient Greeks; it goes back to the Greek geometer and scientist Archimedes. Mathematically, it means that, given a positive number x (no matter how small) and another number y (no matter how large), there always is a natural number n such that $nx > y$. This fact, known as the *Archimedean property*, holds not only for real numbers (i.e., in E^1) but also in many other ordered fields. All such fields are called *Archimedean fields* to distinguish them from other fields in which this property fails. In particular, we shall now prove that every *complete* field (such as E^1) is Archimedean. That is, we have following.

Theorem 1 (Archimedean property)**.** *If x and y are elements of a complete ordered field F (e.g., E^1) and if $x > 0$, then there always is a natural $n \in F$ such that $nx > y$.*

We shall prove this theorem by showing that the opposite assertion is impossible since it leads to a contradiction; it will then follow that our theorem must be true.

Thus, given a fixed element $x > 0$, assume (seeking a contradiction) that there is *no* natural n with $nx > y$. Then, for *all* natural n, we have $nx \leq y$.

This means that y is an upper bound of the set of *all* products

$$nx \quad (n = 1, 2, 3, \dots);$$

call this set M. Clearly, M is nonvoid and bounded above (by y); so, by the assumed completeness of F, M has a supremum, say, $q = \sup M$. As q is an upper bound of M, we have (by the definition of M) that $nx \leq q$ for *each* natural element n. But if n is a natural element, so is $n + 1$. Thus, replacing n by $n + 1$, we get $(n + 1)x \leq q$, whence

$$nx \leq q - x, \quad n = 1, 2, 3, \dots .$$

In other words, $q - x$ (which is *less* than q since $x > 0$) is another upper bound of all nx, i.e., of the set M. But this is impossible because $q = \sup M$ is by definition the *least* upper bound of M; so no *smaller* element, such as $q - x$, can be its upper bound. This contradiction shows that the negation of our theorem must be false. The theorem is proved.

Note 1. The theorem also holds, with the same proof, for "natural multiples" $nx = x + x + \cdots + x$ as defined in §6 (see the note after Definition 3).

Note 2. Theorem 1 shows that no *complete* ordered field, such as E^1 can contain so-called "infinitely small" elements, supposedly $\neq 0$ but such that all their integral multiples are less than 1. (However, such elements do exist in non-Archimedean fields; and recent research, due to A. Robinson, made use of them in what is now generally called "Nonstandard Analysis".)

Corollary 1. *Given any element y in an Archimedean field F, there always are naturals $m, n \in N$ such that $-m < y < n$.*

Proof. Given any $y \in F$, use the Archimedean property (with $x = 1$) to find a natural $n \in F$ such that $n \cdot 1 > y$, i.e., $n > y$. Similarly there is another natural m such that $m > -y$, i.e., $-m < y < n$. □

Corollary 2. *In any Archimedean field, the set N of all naturals has no upper bound, and the set J of all integers has neither upper nor lower bounds. (The negative integers are not bounded below.)*

For, by Corollary 1, no element $y \in F$ can be an upper bound of N (being exceeded by $n \in N$), nor can it be a lower bound of the negative integers (since it exceeds some $-m$, $m \in N$).

Although our next theorem is valid in all Archimedean fields (see Problem 2 below), a simpler proof (avoiding the use of Theorem 2 of §5) can be given for *complete* fields, such as E^1. This is our purpose here.

Theorem 2. *In an Archimedean field F, every nonvoid right-bounded set of integers has a maximum, and every nonvoid left-bounded set of integers has a minimum.*

Proof for *complete* fields. Let M be a nonvoid right-bounded set of integers in a complete field F. By completeness, M has a supremum, call it q. The theorem will be proved if we show that $q \in M$ (for, an upper bound that belongs to the set is its maximum). To prove it, we assume the opposite, $q \notin M$, and seek a contradiction.

Consider the element $q-1$. As $q-1 < q$, Corollary 1 of §9 shows that $q-1$ is exceeded by some element $x \in M$. Since $q \notin M$, q cannot *equal* x. Therefore, as q is an upper bound of M, we have $x < q$, so that $q - 1 < x < q$. Now, as $x < q$, Corollary 1 of §9 yields another element $y \in M$ such that $x < y < q$, and so

$$q - 1 < x < y < q.$$

But this is impossible because x and y are *integers* (being elements of M), and no two *distinct* integers can lie between $q - 1$ and q (indeed, this would imply $0 < y - x < 1$, with $y - x$ a positive integer, contrary to what was shown in Example (C) of §5).

This contradiction shows that q *must* belong to M, and hence $q = \max M$, proving the first clause of the theorem. The second clause is proved quite similarly. We leave it to the reader. □

We now use Theorem 2 to obtain two further results.

Corollary 3. *Given any element x of an Archimedean field F, there always is a unique integer $n \in F$ such that*

$$n \leq x < n + 1.$$

(This integer is called the *integral part* of x, denoted $[x]$.)

Proof. By Corollary 1, there *are* integers $\leq x$. Clearly, the set of all such integers (call it M) is bounded above by x. Hence, by Theorem 2, M has a maximum; call it n. Thus, n is the *greatest* integer $\leq x$. It follows that $n + 1$ *cannot* be $\leq x$, and so $n + 1 > x \geq n$. Thus n has the desired property. This property, in turn, *implies* that $n = \max M$. Hence n is *unique*, as $\max M$ is. □

Examples. $[\frac{1}{2}] = 0$; $[-1\frac{1}{4}] = -2$; $[-4] = -4$; $[\sqrt{2}] = 1$.

As we saw in §4, any ordered field is *dense*:

$$\text{If } a < b \text{ in } F, \text{ there is } x \in F \text{ such that } a < x < b.$$

We shall now show that, in *Archimedean* fields, x can be chosen *rational*, even if a, b are not. We call this the *density of rationals*.

Theorem 3 (Density of rationals)**.** *Given any elements a and b $(a < b)$ in an Archimedean field F, there always is a rational $r \in F$ such that $a < r < b$.* (Briefly: *The rationals are dense in any Archimedean field.*)

Proof. Let $p = [a]$ (the integral part of a); so $p \in J$, $p \le a$. The idea of the proof is to start with p, and then to mark off a small "yardstick" $\frac{1}{n} < b - a$ several (say m) times until $p + \frac{m}{n}$ *lands inside the interval* (a, b) (see Figure 11).

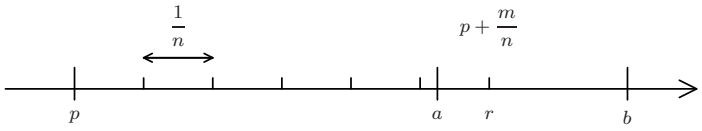


FIGURE 11

More precisely, as F is Archimedean, there are $n, m \in N$, with

$$n(b-a) > 1 \text{ and } m\Big(\frac{1}{n}\Big) > a - p.^{1}$$

Among all such m, fix the *least* one (it exists by Theorem 2). Then

$$a - p < \frac{m}{n} \text{ but } \frac{(m-1)}{n} \le a - p,^{2}$$

so that

$$p + \frac{m}{n} \le a + \frac{1}{n}.$$

Hence

$$a < p + \frac{m}{n} \le a + \frac{1}{n} < a + (b-a) \quad \Big(\text{for } \frac{1}{n} < b - a, \text{ by construction}\Big).$$

Setting

$$r = p + \frac{m}{n},$$

we find that

$$a < r < a + b - a = b.$$

[1] Here we apply the Archimedean property *twice*: first to find n, we take $x = (b-a)$ and $y = 1$; then (having fixed n) we find m, taking $x = \frac{1}{n}$, $y = a - p$.

[2] By the minimality of m.

Moreover, r is *rational*, being the sum of two rationals, p and $\frac{m}{n}$. (The number p is even an *integer*, namely the integral part of a.) Thus r is the desired rational, with $a < r < b$. □

Note 3. Having found one rational r_1, $a < r_1 < b$, we can apply Theorem 3 to find another rational r_2, with $r_1 < r_2 < b$, then a third rational r_3, with $r_2 < r_3 < b$, and so on, ad infinitum. Continuing, we obtain *infinitely many* rationals between a and b. Thus any interval (a, b), with $a < b$, in an Archimedean field (such as E^1) contains *infinitely many* rationals.

Problems on Complete and Archimedean Fields

1. Prove the second part of Theorem 2.

2. Prove Theorem 2 for *Archimedean* fields.
[Hint: If $M \neq \emptyset$ is left-bounded (right-bounded), its elements are greater (less) than some *integer* (why?); so one can use the results of Problems 6 and 7 of §7.]

3. From Theorem 2, prove the induction law of §7 for integers in E^1.
[Hint: Let A be the set of those integers $n \in E^1$ that satisfy $P(n)$ and are $> p$. Show (as in Theorem 2′ of §5) that A contains *all* integers $> p$.]

***4.** In Problem 11 of §9, show that if the intervals $[a_n, b_n]$ also satisfy (for a fixed $d > 0$)

$$b_n - a_n \le \frac{d}{n}, \quad n = 1, 2, \ldots,$$

then

$$\bigcap_{n=1}^{\infty} [a_n, b_n] \text{ contains } \textit{only one} \text{ point, } p,$$

and this p is both $\sup a_n$ and $\inf b_n$. Also show that, if F is only *Archimedean*, the same result follows, *provided that*

$$\bigcap_{n=1}^{\infty} [a_n, b_n] \neq \emptyset.$$

[Hint: Seeking a contradiction, suppose $\bigcap_{n=1}^{\infty}[a_n, b_n]$ contains *two* points p, q with $p - q = r > 0$, say. Then, using the Archimedean property, show that there is an $n \in N$ such that

$$r > \frac{d}{n} \ge b_n - a_n,$$

so that p and q cannot be *both* in $[a_n, b_n]$, let alone in $\bigcap_{n=1}^{\infty}[a_n, b_n]$.]

***5.** Prove that if the principle of nested intervals (cf. Problem 11 of §9) holds in some *Archimedean* field F, then F is complete.
[Outline: If M has an upper bound b, prove that $\sup M$ exists as follows.
Fix any $a \in M$ and let

$$d = b - a, \; c = \frac{1}{2}(a + b);$$

so c *bisects* $[a,b]$.

If there is an $a_1 \in M$ with $a_1 > c$, replace $[a,b]$ by the interval

$$[a_1, b] \subseteq [a,b],$$

noting that

$$b - a_1 < b - c = \frac{d}{2}.$$

If, however, all elements of M are $\leq c$, replace $[a,b]$ by

$$[a,c] \subseteq [a,b].$$

In both cases, the new smaller interval (call it $[a_1,b_1]$) is such that

$$[a_1,b_1] \subseteq [a,b],\ b_1 - a_1 \leq \frac{d}{2},\ a_1 \in M \text{ and } b_1 \text{ is an upper bound of } M.$$

Now let $c_1 = \frac{1}{2}(a_1 + b_1)$, and repeat this process for $[a_1,b_1]$ to obtain a new interval

$$[a_2,b_2] \subseteq [a_1,b_1];\ b_2 - a_2 \leq \frac{d}{4};\ b_2 \text{ an upper bound of } M,\ a_2 \in M.$$

Continuing this process indefinitely, obtain a contracting sequence of intervals $[a_n,b_n]$, with $b_n - a_n \leq d/2^n$ (cf. §6, Problem 8), such that $a_n \in M$ and b_n is an upper bound of M for each n. Then obtain p as in Problem 4 and show that $p = \sup M$, as required.]

6. Prove that an ordered field F is Archimedean iff, for any $x, y \in F$ with $x > 0$, there is a natural *number* $n \in E^1$, with $nx > y$.
[Hint: Use Problem 11′ of §6.]

§11. Roots. Irrational Numbers

An element of a ordered field is said to be *irrational* iff it is not rational, i.e., *cannot* be represented as a ratio m/n of two integers. As we shall see, irrationals exist in any complete ordered field. Irrational elements of E^1 are called *irrational numbers*. We shall also show that the completeness axiom implies the existence of the *n-th root* of any positive element. First, we must prove a lemma.

Lemma. *Let n be a natural number, and let $p \geq 0$ and $a \geq 0$ be elements of an ordered field F. If $p^n < a$ (respectively, $p^n > a$), then there is a positive element $x \in F$ such that $p < x$ and $x^n < a$ (respectively, $p > x$ and $x^n > a$).*

In other words, the given inequality $p^n < a$ ($p^n > a$) is still preserved if p increases (respectively, decreases) by a sufficiently small quantity ϵ.

Proof[1]**.** Let $p^n < a$ ($p \geq 0$), and consider the fraction

$$\frac{a - p^n}{(p+1)^n - p^n}.$$

It is *positive* because $p^n < a$, and so $a - p^n > 0$. Thus by density (Corollary 7 of §4), there is an element $\epsilon > 0$ in F, so small that $\epsilon < 1$ and also

$$\frac{a - p^n}{(p+1)^n - p^n} > \epsilon.$$

Expanding the binomial (cf. §6, Problem 12) and simplifying, we obtain

$$a - p^n > \epsilon[(p+1)^n - p^n] = \binom{n}{1}p^{n-1}\epsilon + \binom{n}{2}p^{n-2}\epsilon + \cdots + \binom{n}{n-1}p\epsilon + \epsilon. \quad (1)$$

Now, as $0 < \epsilon < 1$, we have $\epsilon \geq \epsilon^m$ for any natural m. Hence the inequality (1) can only be strengthened if we replace in it ϵ by various natural powers of ϵ. In this manner, we obtain

$$a - p^n > \binom{n}{1}p^{n-1}\epsilon + \binom{n}{2}p^{n-2}\epsilon^2 + \cdots + \binom{n}{n-1}p\epsilon^{n-1} + \epsilon^n.$$

Hence, transposing p^n to the right side and applying the binomial theorem, we have $a > (p+\epsilon)^n$. Thus, setting $x = p + \epsilon$, we obtain the required $x > p$, with $a > x^n$. This settles the case $p^n < a$ of the lemma.

The other case, $p^n > a$, is trivial if $a = 0$. Thus we assume $p^n > a > 0$. Then

$$\left(\frac{1}{p^n}\right) < \frac{1}{a}$$

and, by what was proved above (with p replaced by $\frac{1}{p}$ and a by $\frac{1}{a}$), there is some

$$y > \frac{1}{p},\ \text{with } y^n < \frac{1}{a},\ \text{i.e., } \left(\frac{1}{y}\right)^n > a.$$

Thus $\frac{1}{y}$ is the required element x, and the proof is complete. □

Theorem 1. *Given any element $a \geq 0$ in a complete ordered field F and a natural number $n \in E^1$, there always exists a unique element $p \geq 0$ ($p \in F$) such that $p^n = a$. This $p \geq 0$ is called the n-th root of a, denoted $p = \sqrt[n]{a}$.*

Proof. Let M be the set of all elements $x \geq 0$ such that $x^n \leq a$. M is nonempty since $0 \in M$. Also, M is right-bounded; e.g., one of its upper bounds is the element $a + 1$ (verify this!). Thus, by completeness, M has a supremum, call it p. Clearly, $p \geq 0$ since $p = \sup M$ and, by definition, all elements of M are ≥ 0. We shall now show that this p is the required element of F, i.e., that $p^n = a$.

[1] At a first reading, the beginner may omit this proof, noting only the lemma itself.

Indeed, if p^n were *less* than a, then by the previous lemma, there would be some $x > p$ such that $x^n < a$, i.e., $x \in M$. But this is impossible because no element x of M can exceed the supremum p of M.

On the other hand, if $p^n > a$, then again by the lemma, there is some $q < p$ ($q \geq 0$) with $q^n > a$. Then for every $x \in M$, we have $x^n \leq a < q^n$, whence (since everything is nonnegative) $x < q$. Thus q exceeds all elements $x \in M$, i.e., q is an upper bound of M. But this is impossible because $q < p$ and p is the *least* upper bound of M.

Thus we see that the inequalities $p^n < a$ and $p^n > a$ are impossible; and it follows by trichotomy that $p^n = a$, as asserted. It remains to prove the uniqueness of p. Suppose that there is yet another element $r \in F$ ($r > 0$) with $r^n = a = p^n$. Then

$$0 = r^n - p^n = (r - p)(r^{n_1} + r^{n-2}p + \cdots + p^{n-1}).$$

Dividing by the *positive* bracketed expression, we obtain $r - p = 0$, whence $r = p$ after all. This shows that p is indeed unique. □

Note 1. $\sqrt[n]{a}$ will always denote the *nonnegative* value of the root. As usual, we write $\sqrt{a}$ for $\sqrt[2]{a}$.

Theorem 2. *Every complete ordered field F (such as E^1) contains irrational elements. In particular, $\sqrt{2}$ is irrational.*

Proof. By Theorem 1, F contains the element $p = \sqrt{2}$, with $p^2 = 2$. Seeking a contradiction, we assume that $\sqrt{2}$ is rational, i.e.,

$$\sqrt{2} = \frac{m}{n}$$

for some natural elements m and n.

Now, by Theorem 2 of §10 (or Problem 5 of §7), we choose the *least possible* such m. Then m and n are *not both even* (otherwise reduction by 2 would yield a smaller m). From

$$\frac{m}{n} = \sqrt{2},$$

we obtain $m^2 = 2n^2$, whence m^2 *is even*. But, as is easily seen, only even elements have even squares. Thus m itself must be even; i.e., $m = 2r$ for some natural element r. It follows that $4r^2 = m^2 = 2n^2$, whence $2r^2 = n^2$; and the same argument shows that n *must be even*. But this is a contradiction since m and n are not both even.

This contradiction shows that, indeed, 2 is irrational, and thus the theorem is proved. □

Note 2. In a similar manner one can prove the irrationality of $\sqrt{n}$, where the natural n is not a full square. Moreover, one can show that *the irrationals are dense in E^1* (cf. Problem 4 below; also, Chapter 1, §9, Corollary 4).

Note 3. From Theorem 2 it follows that *the field of all rationals is not complete* (otherwise, it would contain irrational elements, contrary to its very definition), *even though it is Archimedean* (cf. Problem 6). Thus *there are incomplete Archimedean fields*.

Problems on Roots and Irrationals

1. Prove the irrationality of $\sqrt{3}$ and $\sqrt{5}$.

2. Prove that if a natural n is not a full square, then $\sqrt{n}$ is irrational.
[Hint: Consider first the case where n is not divisible by any square of a prime, i.e.,

$$n = p_1 p_2 \dots p_m,$$

where the p_k are *distinct* primes. The general case reduces to that case; for if $n = p^2 q$ then $\sqrt{n} = p\sqrt{q}$.]

3. Prove that if r is rational and q is not, then $r \pm q$ is irrational; so also are rq, r/q, and q/r if $r \neq 0$.
[Hint: Assume the opposite and find a contradiction.]

4. Prove that the irrationals are *dense* in any complete ordered field F; that is, between any two elements $a, b \in F$ ($a < b$) there is an irrational $x \in F$ ($a < x < b$), and hence there are infinitely many such x.
[Hint: By Theorem 3 of §10, there is a rational r that satisfies

$$a\sqrt{2} < r < b\sqrt{2}.$$

Put $x = r/\sqrt{2}$.]

5. Show by examples that the sum or product of two irrationals may be rational. Thus the irrationals do not form a field. Specify which field axioms fail for irrationals.

6. Show that the rationals in any ordered field form an *Archimedean* subfield.

7. Let $p \in E^1$,

$$A = \text{set of all rationals} < p,\ B = \text{set of all irrationals} < p.$$

Show that $p = \sup A = \sup B$. Solve a similar problem for infima.

8. Let A be the set of all positive rationals x in an ordered field F such that $x^2 < 2$. Without explicitly using $\sqrt{2}$ (which may not exist in F), show that A is bounded above but has no *rational* supremum. Thus give a direct proof that the rational subfield R of F is incomplete.
[Hint: Use the lemma and the fact (proved in Theorem 2) that for no $x \in R$, $x^2 = 2$.]

*§12. Powers with Arbitrary Real Exponents

In §11, we proved the existence and uniqueness of

$$\sqrt[n]{a} \quad (n = 1, 2, \dots)$$

for elements $a \geq 0$ in a complete ordered field. Using this, we shall now define the power a^r for any *rational* $r > 0$.

Definition 1.

Given any element $a \geq 0$ in a complete ordered field F and any rational number $r = m/n > 0$ (where m and n are natural numbers in E^1), we define

$$a^r = \sqrt[n]{a^m}.$$

Here we must clarify two facts:

(1) In case $n = 1$, we have

$$a^r = a^{m/1} = \sqrt[1]{a^m} = a^m.$$

Thus for *natural* values of r, our new definition agrees with the *original* meaning of a^m (as defined in §6), and so contradictions are excluded.

(2) Our definition does not depend on the particular representation of r in the form $\frac{m}{n}$, and thus is unambiguous. Indeed, if r is represented as a fraction in two different ways,

$$r = \frac{m}{n} = \frac{p}{q},$$

then $mq = np$, whence $a^{mq} = a^{pn}$, i.e., $(a^m)^q = (a^p)^n$.

Now, by definition, $\sqrt[n]{a^m}$ is exactly the element whose n-th power is a^m, i.e.,

$$(\sqrt[n]{a^m})^n = a^m.$$

Similarly, $(\sqrt[q]{a^p})^q = a^p$. Substituting this for a^m and a^p in the equation

$$(a^m)^q = (a^p)^n,$$

we get

$$(\sqrt[n]{a^m})^{nq} = (\sqrt[q]{a^p})^{nq}, \text{ whence } \sqrt[n]{a^m} = \sqrt[q]{a^p}$$

(by taking the nq-th root of both sides). Thus, indeed, all representations of r yield the same value of $\sqrt[n]{a^m} = a^r$, and so a^r is well defined.

By using our definition of $\sqrt[n]{a}$ (which can now also be written as $a^{1/n}$) and the formulas stated in Problems 11 and 7 in §6, the reader will easily verify that these formulas remain valid also for powers a^r $(a > 0)$ with rational exponents > 0 as defined above. That is, we have

$$\begin{aligned} &a^r a^s = a^{r+s};\ \ (a^r)^s = a^{rs};\ \ (ab)^r = a^r b^r;\ \ a < b \text{ iff } a^r < b^r\ \ (a, b, r > 0);\\ &a^r < a^s \text{ if } 0 < a < 1 \text{ and } r > s;\ \ a^r > a^s \text{ if } a > 1 \text{ and } r > s;\ \ 1^r = 1. \end{aligned} \tag{1}$$

Henceforth, we assume these formulas known for *rational* r, $s > 0$.

Next we define a^r for any *real* $r > 0$ and any element $a > 1$ in a complete field F. Let A_{ar} denote the set of all elements of F of the form a^x, where x is a rational number, $0 < x \leq r$; i.e.,

$$A_{ar} = \{a^x \mid 0 < x \leq r,\ x \text{ rational}\}.$$

By the density of rationals in E^1 (Theorem 3 of §10), such rationals x *exist*; so $A_{ar} \neq \emptyset$.

Moreover, A_{ar} is *right-bounded* in F. Indeed, fix any rational number $y > r$. By formulas (1), we have, for any positive *rational* $x \leq r$,

$$a^y = a^{x+(y-x)} = a^x a^{y-x} > a^x$$

(since $a > 1$, and $y - x > 0$ implies $a^{y-x} > 1$). Thus, a^y *is an upper bound of all* a^x *in* A_{ar}.

By the assumed completeness of F, $\sup A_{ar}$ *exists*; so we may (and do) define

$$a^r = \sup A_{ar}.^1$$

We also define

$$a^{-r} = \frac{1}{a^r}.$$

If $0 < a < 1$ (so that $\frac{1}{a} > 1$), we set

$$a^r = \left(\frac{1}{a}\right)^{-r} \left(\text{and } a^{-r} = \frac{1}{a^r}\right),$$

where

$$\left(\frac{1}{a}\right)^r = \sup A_{1/a,r},$$

as above. Summing up, we have the following.

Definition 2.

Given $a > 0$ in a complete field F and $r \in E^1$, we define the following:

(i) If $r > 0$ and $a > 1$, then

$$a^r = \sup A_{ar},$$

[1] Note that if r is itself a positive rational, then a^r is the *largest* a^x with $x \leq r$ (where a^r and a^x are as in Definition 1). Thus $a^r = \max A_{ar} = \sup A_{ar}$, and so our present definition *agrees* with Definition 1.

with A_{ar} as above.

(ii) If $r > 0$ and $0 < a < 1$, then

$$a^r = \frac{1}{(1/a)^r},$$

also written $(1/a)^{-r}$.

(iii) $a^{-r} = 1/a^r$ (this defines powers with negative exponents, too).

We also define $0^r = 0$ for any real $r > 0$, and $a^0 = 1$ for any $a \in F$, $a \neq 0$; 0^0 *remains undefined*.

The power a^r is also defined if $a < 0$, provided $r \in N$ (see §6), hence also if r is an *integer* < 0 (then $a^r = 1/a^{-r}$), and even if r is a rational m/n, with n *odd*, because $a^r = \sqrt[n]{a^m}$ has sense in this case, even if $a < 0$. (Why?) This does not work for other values of r. Therefore, in general, we assume $a > 0$.

Again, one can show that formulas (1) hold also for powers with *real* exponents, provided F is complete (see the problems below).

Problems on Powers

1. Verify formulas (1) for powers with positive *rational* exponents r, s.

2. Prove that if A consists of positive elements only, then $q = \sup A$ iff we have

(i) $(\forall x \in A)\ x \leq q$, and

(ii) $(\forall d > 1)\ (\exists x \in A)\ \dfrac{q}{d} < x$.

[Hint: Use Corollary 1 of §9.]

In Problems 3–9, the field F is assumed *complete*.

3. Prove that (i) $a^{r+s} = a^r a^s$ and (ii) $a^{r-s} = a^r/a^s$ for $r, s \in E^1$ and $a > 0$ in F.

[Hint: (i) If $r, s > 0$ and $a > 1$, use Problem 9 of §9, to get

$$a^r a^s = \sup A_{ar} \cdot \sup A_{as} = \sup(A_{ar} \cdot A_{as}).$$

Verify that

$$\begin{aligned} A_{ar} \cdot A_{as} &= \{a^x a^y \mid x, y \in R,\ 0 < x \leq r,\ 0 < y \leq s\} \\ &= \{a^z \mid z \in R,\ 0 < x \leq r + s\}, \end{aligned}$$

where $R =$ rationals. Hence, deduce

$$a^r a^s = \sup(A_{a,r+s}) = a^{r+s}$$

by Definition 2.

(ii) If $r > s > 0$ and $a > 1$ then, by (i), $a^{r-s}a^s = a^r$; so $a^{r-s} = a^r/a^s$. For the cases $r < 0$ or $s < 0$, or $0 < a < 1$, use above results and Definition 1(ii)–(iii).]

4. From Definition 2, prove that if $r > 0$ $(r \in E^1)$, then

$$a > 1 \iff a^r > 1$$

for $a \in F$ $(a > 0)$.

5. Prove for $r, s \in E^1$ that

(i) $r < s \iff a^r < a^s$ if $a > 1$;

(ii) $r < s \iff a^r > a^s$ if $0 < a < 1$.

[Hint: By Problems 3–4,

$$a^s = a^{r+(s-r)} = a^r a^{s-r} > a^r$$

since $a^{s-r} > 1$ if $a > 1$ and $s - r > 0$. If $0 < a < 1$, use Definition 2(ii).]

6. Prove that

$$(ab)^r = a^r b^r \text{ and } \left(\frac{a}{b}\right)^r = \frac{a^r}{b^r}$$

for $r \in E^1$ and positive a, $b \in F$.

[Hint: Proceed as in Problem 3.]

7. Given a, $b > 0$ in F and $r \in E^1$, prove the following:

(i) $a > b \iff a^r > b^r$ if $r > 0$; and

(ii) $a > b \iff a^r < b^r$ if $r < 0$.

[Hint:

$$a > b \iff \frac{a}{b} > 1 \iff \left(\frac{a}{b}\right)^r > 1$$

if $r > 0$, by Problems 4 and 6.]

8. Prove that $(a^r)^s = a^{rs}$ for $r, s \in E^1$ and $a \in F$ $(a > 0)$.

[Outline: First let $r, s > 0$ and $a > 1$. Use Problem 2 to show that

$$(a^r)^s = a^{rs} = \sup A_{a,rs} = \sup\{a^{xy} \mid x, y \in R,\ 0 < xy \leq rs\},$$

with $R = \{\text{rationals}\}$. Thus, prove the following:

(i) $(\forall x, y \in R \mid 0 < xy \leq rs)\ a^{xy} \leq (a^r)^s$, which is easy; and

(ii) $(\forall d > 1)\ (\exists x, y \in R \mid 0 < xy \leq rs)\ (a^r)^s < da^{xy}$. To do this, fix any $d > 1$ and set $b = a^r$. Then

$$(a^r)^s = b^s = \sup A_{bs} = \sup\{b^y \mid y \in R,\ 0 < y \leq s\}.$$

Hence there is some $y \in R$ $(0 < y \leq s)$ such that

$$(a^r)^s < d^{\frac{1}{2}}(a^r)^y. \text{ (Why?)}$$

Fix that y. Now,

$$a^r = \sup A_{ar} = \sup\{a^x \mid x \in R,\ 0 < x \leq r\};$$

so

$$(\exists x \in R \mid 0 < x \leq r) \quad a^r < d^{\frac{1}{2y}} a^x. \text{ (Why?)}$$

Combining all, and using formulas (1) for *rational* x, y, obtain

$$(a^r)^s < d^{\frac{1}{2}}(a^r)^y < d^{\frac{1}{2}}(d^{\frac{1}{2y}}a^x)^y = da^{xy},$$

proving (ii). Proceed.]

*§13. Decimal and Other Approximations

The reader is certainly familiar with decimal approximations of real numbers; e.g.,

$$\sqrt{2} = 1.414213\ldots .$$

A terminating decimal fraction is a sum of powers of 10 multiplied by certain coefficients (the "digits"); e.g.,

$$1.413 = 1\cdot 10^0 + 4\cdot 10^{-1} + 1\cdot 10^{-2} + 3\cdot 10^{-3}.$$

The idea behind decimal approximations is best explained geometrically. Given a real number $x > 0$, we first find a "coarse" decimal approximation of the form $10^{s-1} \le x < 10^s$, where s is an integer (possibly 0 or negative).

Note 1. Such an s exists and is unique. For, by the binomial theorem,

$$10^n = (1+9)^n = 1 + 9n + \cdots > 9n;$$

hence, by the Archimedean property, $10^n > 9n > x$ for large n. Similarly, $10^m > \dfrac{1}{x}$ for some natural m, and so

$$10^{-m} < x < 10^n.$$

Thus, the set of all integers n such that $10^n > x$ is nonvoid and bounded below (e.g., by $-m$). By Theorem 2 of §10, there is a *least* such n; call it s. Then

$$10^s > x \ge 10^{s-1},$$

as required.

Thus, x is in the interval $[10^{s-1}, 10^s)$. To find a better approximation, we subdivide this interval into 9 equal subintervals of length 10^{s-1}. Then x must be in one of these subintervals; let it be

$$[x_1,\ x_1 + 10^{s-1}),$$

where x_1 is some multiple of 10^{s-1}; say,

$$x_1 = m_1 \cdot 10^{s-1}.$$

Thus,

$$x_1 \le x < x_1 + 10^{s-1}.$$

Next we subdivide $[x_1,\ x_1 + 10^{s-1})$ into 10 still smaller subintervals of length 10^{s-2}. Again one of them must contain x; let it be

$$[x_2,\ x_2 + 10^{s-2}),$$

where x_2 is obtained from x_1 by marking off some multiple of 10^{s-2}; say,

$$x_2 = x_1 + m_2 \cdot 10^{s-2}.$$

Then we subdivide the interval $[x_2,\ x_2 + 10^{s-2})$ into 10 still smaller intervals, of length 10^{s-3}, and so on. At the n-th step, x is enclosed in an interval

$$[x_n,\ x_n + 10^{s-n}),$$

approximating x to within 10^{s-n}. Thus one obtains decimal approximations as accurate as is desired.

Instead of using powers of 10, one could use powers of any other number $q > 1$ to obtain, quite similarly, approximations to within q^{s-n}. Moreover, this is possible not only in E^1, but in any Archimedean field F.

Indeed, fixing $q > 1$ and any $x > 0$ in F, we find, exactly as before, a whole number s such that

$$q^{s-1} \le x < q^s.$$

Then, by the Archimedean property of F there is an integer m_1 *in* F such that

$$(m_1 + 1)q^{s-1} > x.$$

Taking the *least* such m_1, we also achieve that

$$m_1 q^{s-1} \le x.$$

(Why?) For brevity, let $x_1 = m_1 q^{s-1}$, so

$$x_1 \le x < x_1 + q^{s-1}.$$

We also put $x_0 = 0$. *Note that* $1 \le m_1 < q$. For if $m_1 \ge q$, then

$$m_1 q^{s-1} \ge q q^{s-1} = q^s > x,$$

contrary to $m_1 q^{s-1} \le x$.

Now, proceeding by induction, suppose that the x_n and the integers m_n in F have already been defined (up to some n) in such a manner that

$$x_n \le x < x_n + q^{s-n}, \quad x_n = x_{n-1} + m_n q^{s-n}, \quad \text{and} \quad 0 \le m_n < q. \tag{1}$$

Then let $m_{n+1} + 1$ be the *least* integer in F, with

$$x < x_n + (m_{n+1} + 1)q^{s-(n+1)};$$

equivalently, m_{n+1} is the largest integer such that

$$x_n + m_{n+1} \cdot q^{s-(n+1)} \leq x.$$

Setting

$$x_{n+1} = x_n + m_{n+1} \cdot q^{s-(n+1)},$$

we have

$$x_{n+1} \leq x < x_{n+1} + q^{s-(n+1)}.$$

Moreover, $0 \leq m_{n+1} < q$; for if $m_{n+1} \geq q$, then

$$x_n + x_{n+1} \cdot q^{s-(n+1)} \geq x_n + qq^{s-(n+1)} = x_n + q^{s-n} > x \quad \text{(by (1))},$$

contrary to our choice of m_{n+1}.

Thus, by induction, we obtain two infinite sequences $\{x_n\}$ and $\{m_n\}$ in F such that the m_n are integers in F ($0 \leq m_n < q$), and (1) holds for all n. We call x_n the *n-th q-ary approximation of x* (from below). In particular, if $q = 2$, $q = 3$, or $q = 10$, we speak of *binary*, *ternary*, or *decimal* approximations, respectively. If the integers m_n (called *q-ary digits*) and s are given, they determine all x_n uniquely. Indeed, setting $n = 1, 2, 3, \ldots$ in the second part of (1), we obtain (with $x_0 = 0$), step by step,

$$x_n = m_1 q^{s-1} + m_2 q^{s-2} + \cdots + m_n q^{s-n}, \quad n = 1, 2, 3, \ldots . \tag{2}$$

The infinite sequence $s, m_1, m_2, \ldots, m_n, \ldots$ is called the *q-ary* (e.g., *binary*, *ternary*, *decimal*) *expansion* of x. Customarily, one briefly writes $x = m_1 m_2 \ldots$, indicating the value of s by placing a dot (the "*q-ary point*") at an appropriate step (namely, after the coefficient m_s of q^0).

Note 2. If s is negative (say, $s = -p$), we insert $p + 1$ zeros before m_1 and place the "dot" after the first zero so inserted.

Note 3 If all m_n from some digit onward are equal to some m, we say that $\{m_n\}$ *terminates* in m (any such repeating digit or group of digits is called the *period* of $\{m_n\}$). This m *cannot be* $q - 1$ (cf. Problem 3). If $m = 0$, we simply say that $\{m_n\}$ *terminates*, and we may omit the zeros at its "end". Then, for sufficiently large n, $x_n = x$; that is, formula (2) expresses x *exactly*.

Examples.

(1) The decimal expansion of $40/33$ is $1.2121212\ldots$, also written $1.2(12)$ where (12) is the repeating "period" of the expansion. Here $s = 1$ since $10^1 > 40/33 > 10^0$; and $m_1 = 1$, $m_2 = 2$, $m_3 = 1$, and so on. In practice, the digits m_n are found by the familiar division algorithm.

(2) The *binary* expansion of 10 is $1010.000\ldots$ (briefly, 1010). Here $s = 4$ since $2^4 > 10 > 2^3$; we have $10 = 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$, i.e., $m_1 = 1$, $m_2 = 0$, $m_3 = 1$, $m_4 = m_5 = m_6 = \cdots = 0$. The expansion terminates, and we may omit the zeros at its "end", leaving, however, the zero preceding the "binary" point, so as to indicate the value of s. Observe that the digits m_n in a binary expansion can have only the value 0 or 1 since $0 \leq m_n < q = 2$. Similarly, in ternary expansions, m_n is either 0 or 1 or 2.

In practice, the q-ary expansion of x is obtained by "trying" to represent x as a sum of powers of q (i.e., $q^{s-1}, q^{s-2}, \ldots$) multiplied by suitable coefficients $m_n < q$ ($m_n \geq 0$); the latter are the digits. If the process does not terminate, one obtains an infinite sequence of q-ary approximations x_n, as in formula (2). In all cases, we have the following.

Theorem 1. *Every element $x > 0$ in an Archimedean field F is the supremum of the set $\{x_1, x_2, \ldots, x_n, \ldots\}$ of its q-ary approximations ($q > 1$, $q \in F$).*

Proof. By the definition of the x_n, we have

$$x_n \leq x < x_n + q^{s-n}, \quad n = 1, 2, \ldots .$$

Thus none of the x_n exceeds x, and so x is an upper bound of all x_n.

It remains to show that x is the *least* upper bound. Seeking a contradiction, suppose there is a smaller upper bound y, $y < x$. Then we have

$$x_n \leq y < x < x_n + q^{s-n},$$

and hence

$$0 < x - y < (x_n + q^{s-n}) - x_n = q^{s-n},$$

i.e., $0 < x - y < q^s/q^n$, or

$$q^n(x - y) < q^s, \quad n = 1, 2, 3, \ldots .$$

But this is incompatible with the Archimedean property. (Why?) Thus the theorem is proved. □

If the field F is *complete* and q is an *integer* > 1, the process described above can be reversed. More precisely, we have the following.

Theorem 2. *Let s be an integer in E^1, and let q and m_n ($n = 1, 2, 3, \ldots$) be integers in a complete field F, with $q > 1$, $0 \leq m_n \leq q - 1$, and $m_1 \geq 1$. If the sequence $\{m_n\}$ does not terminate in $q - 1$, there is a unique element $x > 0$ in F, whose q-ary expansion, as defined above, is exactly $s, m_1, m_2, \ldots, m_n, \ldots$.*

Proof. With q, s and m_n as above, define

$$x_n = \sum_{k=1}^{n} m_k q^{s-k}, \quad y_n = \sum_{k=1}^{n} (q-1) q^{s-k}, \quad n = 1, 2, 3, \ldots,$$

so that the x_n are as in (2). As $m_k \leq q - 1$, we have $x_n \leq y_n$. Moreover, as $\{m_k\}$ does not terminate in $q - 1$, we have $m_k < q - 1$ for infinitely many k, and hence $x_n < y_n$ for large n so that $d_n = y_n - x_n > 0$, and the differences

d_n *increase* with n. So also do x_n and y_n. Let d be one of the $d_n > 0$. Then, for sufficiently large n, $y_n - x_n = d_n > d > 0$, and we obtain

$$q^{s-1} \le \sum_{k=1}^{n} m_k q^{s-k} = x_n < y_n - d = \sum_{k=1}^{n} (q-1)q^{s-k} - d = q^s - q^{s-n} - d < q^s - d.$$

Thus, the set of all x_n is bounded above by $q^s - d$; so, by completeness, it has a supremum; call it x. By Corollary 2 of §9,

$$q^{s-1} \le x = \sup x_n \le q^s - d < q^s.$$

Also, for $p > n$, we obtain as above (for sufficiently large p and some $d_0 > 0$)

$$\begin{aligned} x_p - x_n &= \sum_{k=n+1}^{p} m_k q^{s-k} \\ &< \sum_{k=n+1}^{p} (q-1)q^{s-k} - d_0 \\ &= q^{s-n} - q^{s-p} - d_0 \\ &< q^{s-n} - d_0, \end{aligned} \tag{3}$$

whence $x_p < x_n + q^{s-n} - d_0$. Keeping n fixed and passing to $\sup_{p>n} x_p$, we get

$$x = \sup_{p>n} x_p \le x_n + q^{s-n} - d_0 < x_n + q^{s-n}.$$

Thus, $x_n \le x < x_n + q^{s-n}$ for each n.

Finally, from $x_n = \sum_{k=1}^{n} m_k q^{s-k}$, we obtain

$$x_{n+1} = \sum_{k=1}^{n+1} m_k q^{s-k} = x_n + m_{n+1} q^{s-(n+1)}.$$

This, combined with the previously obtained inequalities,

$$x_n \le x < x_n + q^{s-n} \text{ and } q^{s-1} \le x < q^s,$$

shows that the x_n coincide with the q-ary approximations of x as defined in (1) and (2), and that $s, m_1, m_2, \dots, m_n, \dots$ is the q-ary expansion of x, as required. The proof is complete. □

Thus, we see that, for any integer $q > 1$ in a complete field F, there is a *one-to-one correspondence* between positive elements $x \in F$ and their q-ary expansions, i.e., sequences $s, m_1, m_2, \dots, m_n, \dots$, *not terminating in* $q-1$ and such that $0 \le m_n < q$ and $m_1 \ge 1$ (with s an integer in E^1, and m_n integers in F). By Theorem 1, x is the supremum of all x_n, i.e., sums of the

form (2). This supremum is denoted by

$$\sum_{k=1}^{\infty} m_k q^{s-k}.$$

The representation of x as a supremum of finite sums is not unique. For example, in decimal notation, $2 = 2.0000\dots$; but 2 is also the supremum of approximations of the form $1.9999\dots$. However, as noted above, our definitions exclude $q-1$ as a period, and so uniqueness is achieved.

Problems on Decimal and q-ary Approximations[1]

1. Why is there a largest integer m_n such that $x_{n-1} + m_n q^{s-n} \le x$?

2. Given $c > 0$ and $0 < r < 1$, show that $c/(1-r)$ is the supremum of all sums

$$\sum_{k=0}^{n} cr^k, \quad n = 1, 2, \dots .$$

[Hint: Compute $\sum_{k=0}^{n} cr^k$ from Problem 9 in §6.]

3. Why can $q-1$ never occur as a period, by our definitions?
[Hint: If $(\forall n > p)$ $m_n = q-1$, formula (2) yields $x_n = x_p + q^{s-p} - q^{s-n}$. (Verify!) From Problem 2, show that $x = \sup x_n = x_p + q^{s-p}$, contrary to formula (1).]

4. Write in binary and ternary notation the following decimal expressions:
a) 2.311; b) 23.11; c) 231.11; d) 231110; e) 45/4; f) 1/3.

5. Write the following binary fractions in decimal and ternary notation:
a) 1.0101; b) 1001,001; c) 10100.1; d) 0.0001001; e) 0.0010001.

6. Explain how (and why) decimal expansions of *rationals* m/n can be obtained by repeated division (cf. Problem 20 of §6). Similarly for q-ary expansions, with q an integer > 1.

7. Let q be an integer > 1. Show that the q-ary expansion of x is *periodic*[2] iff x is a rational, m/n.
[Hint: If $x = m/n$, consecutive division by n yields remainders $< n$. As there are only finitely many such remainders, they must eventually repeat. For the converse, use Problem 2 and Theorem 1.]

8. Using the result of Problem 2, find x from its *periodic* q-ary expansion:
a) $x = 0.00(13)$, $q = 10$; b) same with $q = 4$; c) same with $q = 5$.

9. Answer the question ("why?") posed at the end of the proof of Theorem 1.

[1] In these problems, $q > 1$, x, x_n and m_n are elements of an *Archimedean* ordered field F, defined as above in §13. The m_n are *integers* in F.

[2] The sequence $\{m_n\}$ is called *periodic* iff it terminates in consecutive repetitions of a finite subsequence $(p_1, p_2, \dots, p_k)$, possibly (0).

*§14. Isomorphism of Complete Ordered Fields

We shall now show that, in a sense, there is *only one* complete ordered field. That is, all such fields have the same mathematical properties as E^1 and thus cannot be distinguished *mathematically* from E^1.

Definition 1.

Two fields, F and F', are said to be *isomorphic* iff there is a one-to-one mapping $f: F \underset{\text{onto}}{\leftrightarrow} F'$ such that (denoting addition and multiplication in both fields by the same symbols, $+$ and $\cdot$)

$$(\forall\, x, y \in F) \quad f(x+y) = f(x) + f(y) \text{ and } f(x \cdot y) = f(x) \cdot f(y). \tag{1}$$

If F and F' are *ordered* fields, we also require that

$$(\forall\, x, y \in F) \quad x < y \iff f(x) < f(y). \tag{2}$$

In other words, the mapping f (called an *isomorphism* between F and F') establishes a one-to-one correspondence between elements $x \in F$ and $f(x) \in F'$ that *carries the sum and product of any elements $x, y \in F$ into the sum and product, respectively, of $f(x)$ and $f(y)$ in F'*. We briefly say that f *preserves* the operations in F and F'. In the ordered case, the map f is also supposed to preserve *order* (formula (2)). Writing briefly x' for $f(x)$, we may say that, under the correspondence $x \leftrightarrow x'$, sums correspond to sums and products correspond to products:

$$(x+y) \leftrightarrow (x'+y'), \quad xy \leftrightarrow x'y'.$$

Thus, any formula valid in F can be "translated" into a formula valid in F'; one only has to replace

$$x,\ y,\ z,\ \ldots \in F \text{ by } x',\ y',\ z',\ \ldots \in F'.$$

Anything that can be proved in F can also be proved in F', and conversely. In ordered fields, this applies to inequalities as well, due to (2). Thus F' behaves exactly like F, as far as field operations and inequalities are concerned. Therefore, it is customary not to distinguish between two isomorphic fields F and F', even though their elements may be objects of different nature. (Compare this to playing one and the same game of chess or cards with two different sets of chessmen or decks of cards: it is not the color or shape of the chessmen but the game itself that really matters.)

Consequently, if F and F' are isomorphic, we treat them as just two "copies" of the same field; we call F' the *isomorphic image* of F (under the isomorphism f) and briefly write

$$F \cong F', \text{ or } F \overset{f}{\cong} F'.$$

The same definitions and conventions also apply if F and F' are any *sets* (not necessarily fields) with some "addition" and "multiplication" defined in them, satisfying the closure law but not necessarily the other field axioms. If only *one* operation in F and F' (say, addition) is considered, or defined, the isomorphism f is supposed to preserved this particular operation: $f(x+y) = f(x) + f(y)$. We then say that F and F' are isomorphic with *respect to addition* (though, possibly, not with respect to multiplication).[1] *Order isomorphism* (2) may apply to any ordered *sets*, regardless of operations.

Note. If the map f satisfies (1) but is not necessarily one-to-one or onto F', we call it a *homomorphism* (of F into F').

Examples.

(a) Let $F = E^1$ and let F' be the set of all ordered *pairs* of the form $(x, 0)$, $x \in E^1$. For such pairs, define

$$(x,0) + (y,0) = (x+y,0),\ (x,0) \cdot (y,0) = (xy,0);$$

and

$$(x,0) < (y,0) \iff x < y.$$

It is easy to verify that F' is an ordered field under these operations, and the mapping $x \leftrightarrow (x,0)$ is an isomorphism satisfying both (1) and (2). Thus $E^1 \cong F'$.

(b) Let N be the set of all natural numbers, and let N'' be the set of all *even* elements of N. Define the mapping $f: N \to N''$ by $f(x) = 2x$. This map is one-to-one and onto N''. (Verify!) Moreover,

$$(\forall\, x, y \in N) \quad f(x+y) = 2(x+y) = 2x + 2y = f(x) + f(y).$$

Thus f preserves *addition*; so it is an isomorphism *with respect to addition* (but not with respect to multiplication). It also preserves the order since we have

$$x < y \text{ iff } 2x < 2y, \text{ i.e., } f(x) < f(y).$$

Thus, $N \cong N''$, with respect to addition and order.

(c) The identity map $I: F \leftrightarrow F$, defined by $I(x) = x$, obviously preserves any operations or ordering defined in F; e.g., for multiplication, we have $I(xy) = xy = I(x) \cdot I(y)$. Thus, I is an isomorphism of F *onto itself*: $F \overset{I}{\cong} F$.

Below, N and R will denote the naturals and rationals *in* E^1, while N' and R' are the corresponding sets in some arbitrary ordered field F. The

[1] Of course, it does not matter whether the operation involved is denoted by $(+)$ or some other symbol and whether it is called "addition" or some other name. It may also occur that the operations in F and F' have *different* names and are *differently* denoted.

unity element of F is denoted by $1'$ to distinguish it from $1 \in E^1$. From §6 (Definition 3 and the subsequent note), we recall that, for any $n \in N$ and $a \in F$,

$$n \cdot a = na = a + a + \cdots + a \quad (n \text{ terms}).$$

We shall now define ra for any $r \in R$ and $a \in F$.

Definition 2.

Given any element a of a field F, and a rational number $r = \dfrac{m}{n} \in E^1$ $(m, n \in N)$, we define

$$r \cdot a = ra = \frac{ma}{n \cdot 1'}$$

($1'$ being the unity of F). We also put

$$(-r) \cdot a = -ra \text{ and } 0 \cdot a = 0' \in F.$$

Note that $ra \in F$ in all cases.

This definition is unambiguous, inasmuch as it does not depend on the particular representation of r as a fraction m/n. For, if $r = m/n = p/q$ for some m, n, p, $q \in N$, then $mq = np$, whence $(mq)(a \cdot 1') = (np)(a \cdot 1')$. It easily follows [cf. Problem 11(vii) in §6] that $(ma) \cdot (q1') = (pa)(n1')$, and hence

$$\frac{pa}{q1'} = \frac{ma}{n1'} = ra.$$

Thus, indeed, ra is uniquely determined. Moreover, if $r \in N$, i.e., $r = m/1$, then

$$ra = \frac{ma}{1 \cdot 1'} = \frac{ma}{1'} = ma.$$

Thus, for a natural r, Definition 2 agrees with our previous definition of the natural multiple ma, and so there is no danger of contradiction. We now obtain the following.

Theorem 1. *For any elements a and b of a field F and any rational numbers r and s (in E^1), we have the following*:

(i) $ra + sa = (r + s)a$;

(ii) $ra \cdot sb = (rs)(ab)$;

(iii) $r(a + b) = ra + rb$;

(iv) *if F is an ordered field, we also have $ra < sa$ iff $r < s$, provided $a > 0'$.*

Indeed, if r, $s \in N$, all this follows from Problems 11(vi)–(viii) and 11′ of §6. The general case (r, $s \in R$) easily follows by Definition 2. We leave the details to the reader.

Theorem 2. *The ordered subfield R of E^1 (i.e., the field of all rational numbers) is isomorphic with the rational subfield R' of any other ordered field F (with zero element $0'$ and unity $1'$).*

Proof. As was noted at the end of §7, R and R' are ordered fields (subfields of E^1 and F, respectively). To establish their isomorphism, we define a mapping $f: R \to R'$ by setting

$$f(x) = x \cdot 1' \quad \text{for } x \in R.$$

Then, by Theorem 1, $(\forall x, y \in R)$

$$f(x + y) = (x + y) \cdot 1' = x1' + y1' = f(x) + f(y)$$

and

$$f(xy) = (xy) \cdot 1' = (x1') \cdot (y1') = f(x) \cdot f(y).$$

Thus, f preserves the operations.

By part (iv) of the theorem, f also preserves order. This also implies that f is one-to-one; for, if

$$x \neq y \quad (x, y \in R),$$

then either

$$x < y \text{ or } x > y,$$

whence

$$f(x) < f(y) \text{ or } f(x) > f(y),$$

and in both cases $f(x) \neq f(y)$, as required.

It only remains to show that f is *onto* R', i.e., *that each element $r' \in R'$ has the form $r' = f(x)$ for some $x \in R$.* Let

$$r' = \frac{m'}{n'} \quad (m', n' \text{ naturals in } F).$$

Now, by Problem 11′ of §6, we have

$$m' = m \cdot 1' \text{ and } n' = n \cdot 1'$$

for some $m, n \in N$. Hence,

$$r' = \frac{m'}{n'} = \frac{m \cdot 1'}{n \cdot 1'}.$$

Setting $x = m/n$, we have, by definition,

$$f(x) = x \cdot 1' = \frac{m1'}{n1'} = r'.$$

Thus, our assertion is proved in case $r' > 0'$.

If, however, $r' < 0'$, then $-r' > 0'$; so, by what was proved above, $-r' = f(x) = x \cdot 1'$ for some $x \in R$, and it easily follows that $r' = (-x) \cdot 1' = f(-x)$.

Finally, $0' = 0 \cdot 1' = f(0)$, by definition. Thus, f is, indeed, *onto* R'. This completes the proof. $\square$

Observe that the map f carries naturals and integers of E^1 onto those of F. Thus, we have also proved that *the set N of all natural numbers (in E^1) is isomorphic, with respect to addition, multiplication, and order, to the set N' of all naturals in any ordered field F*, and similarly for the integers (J and J'). Because of the isomorphism established above, we may regard R and R' as "copies" of one and the same set and not distinguish between them. Similarly for N and N', or J and J'. Thus, henceforth, we adopt the convention that R, N, and J are *the same* in each ordered field F, so that each F contains the rational numbers, R, *themselves* ($R = R'$). In particular, $0 = 0'$, $1 = 1'$, and $r \cdot 1' = r \cdot 1 = r$ for any rational r.

Next, let F be *complete*. Then one can define ra ($a \in F$) for any *real* r, in much the same manner as we defined a^r in §12. Fixing first some $r > 0$ in E^1 and $a > 0'$ in F, let

$$A_{ra} = \{xa \mid x \in R,\ 0 < x \le r\};$$

i.e., A_{ra} is the set of all xa (defined as in Definition 2), with $x \in R$, $0 < x \le r$. Clearly, $A_{ra} \neq \emptyset$ and, by Theorem 1(iv), A_{ra} is right-bounded in F by any ya, with $y \in R$, $y > r$. Thus, by completeness, $\sup A_{ra}$ exists in F; so we define

$$ra = \sup A_{ra} \quad (r > 0,\ a > 0'). \tag{3}$$

If, in particular, r is rational then, by Theorem 1(iv), ra is the *largest* of all xa in A_{ra}; so ra (as in Definition 2) equals $\sup A_{ra} = \max A_{ra}$. Thus, in the rational case, our new definition of ra agrees with Definition 2. Finally, if $r < 0$, we put $ra = -(-r)a$, and if $a < 0'$, we define $ra = -[r(-a)]$. Thus, ra is defined for all $r \in E^1$ and all $a \in F$.

It is easy to verify that Theorem 1 remains valid for *arbitrary* real r and s (provided F is complete); cf. Problems 2–4 below. We now have the following.

Theorem 3. *Any complete ordered field F is isomorphic with E^1.*

Proof. As before, we define $f\colon E^1 \to F$ by setting

$$f(r) = r \cdot 1' \quad \text{for } r \in E^1.$$

Exactly as in Theorem 2, it follows that f preserves the operations and the order and is one-to-one. Only the fact that f is *onto* F requires a different proof.

Given any $q \in F$, we have to find an $r \in E^1$ such that $q = f(r) = r \cdot 1'$. First let $q > 0'$, and let

$$Q' = \{x \in R' \mid 0 < x \le q\};$$

i.e., Q' consists of all *rational* x such that $0 < x \le q$ in F. Clearly, q is an upper bound of Q'. Moreover, there is no *smaller* upper bound; for if $p < q$ then, by the density of rationals in the complete field F, there is $x \in R'$ with $p < x < q$ ($x > 0$), so that $x \in Q'$ and $x > p$, and hence p is not an upper bound of Q'. Thus $q = \sup Q'$. It also follows that Q' has *rational* upper bounds (take any rational $y > q$).

Since $R' = R$, we may also regard Q' as a set of rationals *in* E^1, with rational upper bounds *in* E^1. Thus Q' also has a supremum in E^1; call it r. Let us denote Q' by Q *when it is regarded as a subset of* E^1.[2] Thus

$$Q = \{x \in R \mid 0 < x \le r\} \text{ in } E^1,$$

while

$$Q' = \{x \in R' \mid 0' < x \le q\} \text{ in } F.$$

More precisely, the sets Q and Q' correspond to each other under the isomorphism

$$x \leftrightarrow x \cdot 1'.$$

Thus Q' is exactly the set of all elements in F of the form

$$x \cdot 1' \quad (x \in R,\ 0 < x \le r).$$

In other words, $Q' = A_{ra}$ with $a = 1'$, and

$$q = \sup Q' = \sup A_{ra} = r \cdot 1' = f(r);$$

i.e., q has the form $f(r)$ for some $r \in E^1$, as required. This proves our assertion in case $q > 0'$.

On the other hand, if $q < 0'$, then $-q > 0'$ and hence, by what was proved above, $-q = f(s)$ for some $s \in E^1$. Hence, by definition, $f(-s) = q$; so our assertion is true in the negative case as well.

Finally, by definition, $0' = f(0)$. Thus *every* element of F has the form $f(r)$, $r \in E^1$, and so f is indeed *onto* F. This completes the proof. $\square$

The theorems that we have proved show that, except for isomorphic "copies", there is only one complete ordered field (E^1), only one rational ordered field (R), and only one ordered system of naturals (N). We express this briefly by saying that E^1, R, and N are unique *to within isomorphism*. Due to this, we may henceforth treat natural multiples na ($n \in N$, $a \in F$) as *products* in F; similarly for rational multiples ra ($r \in R$, $a \in F$).

While the uniqueness of E^1 is thus established, there still remains the question of its *existence*. Indeed, right from the start, E^1 was introduced only *axiomatically*; that is, we have *assumed* that there is some set E^1 with two

[2] To make the same distinction, we also continue writing R', N', $1'$, and $0'$ for the rationals, naturals, unity, and zero *of* F, even though $R' = R$ by our convention.

operations $(+)$ and $(\cdot)$ and an order relation $<$ satisfying our Axioms I–X (including completeness). However, this fact was never *proved*. In the next section, we shall take up the problem of *constructing* E^1 from simpler structures, thus proving its existence.

Problems on Isomorphisms

1. Complete the proof of Theorem 1.

2. Prove parts (i)–(iii) of Theorem 1 for positive *real* r, s and positive a, b in a complete field F.
[Hint: Proceed as in Problems 8 and 9 in §9 to show that

$$\sup A_{ra} \cdot \sup A_{sb} = \sup A_{rs,ab} \text{ and } \sup A_{ra} + \sup A_{sb} = \sup(A_{r+s,a+b}).$$

Then apply formula (3) from p. 108, noting that Theorem 1 holds for *rational* r, s.]

3. Solve Problem 2 for *arbitrary* r, $s \in E^1$ and a, $b \in F$.
[Hint for part (i): Let first $r > s > 0$, $a > 0'$. As $r - s > 0$, Problem 2 yields

$$(r-s)a + sa = (r-s+s)a = ra,$$

whence $(r-s)a = ra - sa$.

This holds also if $s > r > 0$ since, by definition, $(r-s)a = -(s-r)a$, where $s - r > 0$; so, as shown above, $(s-r)a = sa - ra$, and hence

$$(r-s)a = -(sa - ra) = ra - sa.$$

Thus $(r \pm s)a = ra \pm sa$ for *positive* r, s, a. Now, if $r > 0 > s$ and $a > 0'$, then $-s > 0$ and hence $(r+s)a = [r-(-s)]a = ra + sa$. Similarly in the other cases.]

4. Prove part (iv) of Theorem 1 for any *real* r, s and any a, b in a complete ordered field F.
[Hint:

$$r < s \implies s - r > 0 \implies (s-r) \cdot 1' > 0',$$

by the very definition of multiples ra for *positive* r, a (here $a = 1'$). But, by Problem 3,

$$(s-r) \cdot 1' = s1' - r1' = f(s) - f(r);$$

thus $f(s) - f(r) > 0'$, as required. Conversely, if $f(r) < f(s)$, we cannot have $r \geq s$ (why?), and so $r < s$.]

Give also a *direct* proof based on properties of suprema (without referring to Problem 3).

5. Let F and F' be two fields, with zero-elements 0 and $0'$, and unities 1 and $1'$, respectively. Prove that if $f\colon F \underset{\text{onto}}{\leftrightarrow} F'$ is an isomorphism, then

(i) $f(0) = 0'$;

(ii) $f(1) = 1'$;

(iii) $(\forall x \in F)\ f(-x) = -f(x)$, and $f\left(\frac{1}{x}\right) = \frac{1'}{f(x)}$ (the latter if $x \neq 0$).

Also show (by induction) that

$$x \in N \text{ iff } f(x) \in N',$$

i.e., $f[N] = N'$ (with N and N' as in the text). Hence, infer that $f[J] = J'$ and $f[R] = R'$.
[Hint for part (i): To prove that $f(0)$ is the zero element of F', show that $(\forall y \in F')$ $y + f(0) = y$, noting that $y = f(x)$ for some x (why?), and using (1) from p. 104. Use similar arguments for parts (ii) and (iii).]

6. With the notation of Problem 5, let F and F' be *ordered* fields, $F \overset{f}{\cong} F'$. Prove by induction that

$$(\forall n \in N) \quad f(n) = n \cdot 1',$$

and infer that

$$(\forall r \in R) \quad f(r) = r \cdot 1',$$

with $r \cdot 1'$ as in the text. Also show that if $p = \sup A$ $(A \subset F)$, then $f(p) = \sup f[A]$ in F', and similarly for infima. (The last part also holds for order-isomorphisms of ordered *sets*, regardless of operations.)

7. Continuing Problem 6, show that if F and F' are *Archimedean* fields, with $F \overset{f}{\cong} F'$, then *necessarily*

$$(\forall x \in F) \quad f(x) = x \cdot 1'$$

(with $x \cdot 1'$ defined as in the text, for $x \in E^1$). Thus there is *at most one* isomorphism $f\colon F \underset{\text{onto}}{\leftrightarrow} F'$.

8. Show that the relation of isomorphism is reflexive, symmetric, and transitive, i.e., an equivalence relation.
[Hint: $F \overset{I}{\cong} F$ by Example (c). Show that $F \overset{f}{\cong} F'$ and $F' \overset{g}{\cong} F''$ implies $F' \overset{f^{-1}}{\cong} F$ and $F \overset{h}{\cong} F''$, where $h(x) = g(f(x))$.]

*§15. Dedekind Cuts. Construction of E^1

I. In the problems of §7 in Chapter 1, we sketched a method of constructing integers from naturals, and rationals from integers. Now we shall show how *reals* can be constructed from rationals. More generally, we shall show how an Archimedean field R can be extended to a complete one, and consider a

similar problem for ordered *sets* in general.[1] This can be done by using so-called *Dedekind cuts* (R. Dedekind, German mathematician, 1831–1916). We define them now for any ordered set R.

Definition 1.

A *Dedekind cut* (briefly, *cut*) in an ordered set R is a pair (A, B) of nonempty subsets of R such that A is exactly the set of all lower bounds of B, and B is the set of all upper bounds of A, in R.

A cut (A, B) is called a *gap* (in R) iff $A \cap B = \emptyset$.

If (A, B) is *not* a gap, i.e., $A \cap B \neq \emptyset$, then $A \cap B$ consists of a *single* element; for, by the definition of (A, B), any element $p \in A \cap B$ is an upper bound of A (since $p \in B$) and hence $p = \max A$ (for $p \in A$); similarly, $p = \min B$. Thus, by the uniqueness of $\max A$ and $\min B$,

$$p = \max A = \min B \text{ is } \textit{unique}.$$

From Definition 1, it also follows that

$$y \leq x \in A \implies y \in A; \text{ and } y \geq x \in B \implies y \in B.$$

(Why?) In the examples below, R is the set of all rationals.

Examples.

(1) Let $p \in R$, let

$$A = \{x \in R \mid x \leq p\}, \quad B = \{x \in R \mid x \geq p\}.$$

This yields a cut (A, B); it is not a gap, for $\max A = \min B = p \in A \cap B$.

(2) Let

$$A = \{x \in R \mid x \leq 0 \text{ or } x^2 < 2\}, \quad B = \{x \in R \mid x > 0 \text{ and } x^2 > 2\}.$$

Then (A, B) is a cut. (Verify!) It is a gap since $A \cap B = \emptyset$. Also, $\max A$ and $\min B$ do not exist in R (cf. §11, Problem 8).

Thus, we see that there are cuts of both kinds in R: gaps and nongaps.

Theorem 1. *For any cut (A, B) in an ordered set R, we have $R = A \cup B$.*

Indeed, by Definition 1,

$$A \subseteq R \text{ and } B \subseteq R,$$

whence $A \cup B \subseteq R$. Conversely, if $x \in R$ and, say, $x \notin A$, then x is *not* a lower bound of B; i.e.,

$$x > y \text{ for some } y \in B.$$

[1] We recall from §2 that an *ordered set* is a set in which a transitive and trichotomic relation "$<$" is defined. The notions of upper and lower bound, supremum, infimum, etc. are defined in such a set exactly as in ordered fields. Similarly for "completeness".

But, as noted above, $x > y \in B \implies x \in B$. Similarly, if $x \notin B$, then $x \in A$. Thus, x must be in one of A and B, i.e., $x \in A \cup B$. □

Theorem 2. *For any cuts (A, B) and (A', B') in an ordered set R, we have either*

$$A \subset A' \text{ or } A \supset A' \text{ or } A = A'.$$

Moreover,

$$A \subset A' \iff B \supset B'.$$

Proof. If $A \supseteq A'$, then either $A = A'$ or $A \supset A'$, so there is nothing to prove.

So suppose A' has an element r *not* in A. Then, by Theorem 1, $r \in B$. Hence r is an upper bound of A, i.e., $(\forall x \in A)\ x \leq r$. As

$$x \leq r \in A' \implies x \in A',$$

we get $(\forall x \in A)\ x \in A'$, i.e., $A \subset A'$.[2] Thus we have either $A \supseteq A'$, or else $A \subset A'$, as asserted. We leave to the reader the proof that $A \subset A'$ is equivalent to $B \supset B'$. □

We shall now show that any ordered set R can be made *complete* by adding to it new elements, so as to "fill" its gaps. The nature of these elements may be arbitrary; it is only required that they be different from the original ("old") elements of R. Thus, *for each gap (A, B) in R, we introduce a new element p in such a manner that different elements p correspond to different gaps (A, B)*; we shall say that this p is *determined* by the corresponding gap (A, B), and conversely.

If (A, B) is *not* a gap then, as was shown above, there is in R an element $p = \max A = \min B$; in this case, too, we shall say that p is *determined* by the cut (A, B). Thus *each cut (A, B) in R determines a certain element p that is "new" or "old" according as (A, B) is, or is not, a gap.*[3] The set consisting of the "old" and "new" elements together is called the *completion* of R, denoted $\overline{R}$. By what was said above, there is a one-to-one correspondence between all elements of $\overline{R}$ and all cuts in R; the "new" elements correspond to *gaps* in R.

For brevity, we write "$p \equiv (A, B)$" to mean that p is determined by (A, B).

Definition 2.

For any elements $p \equiv (A, B)$ and $q \equiv (A', B')$ in $\overline{R}$, we write

$$p < q \text{ iff } A \subset A', \text{ and } p \leq q \text{ iff } A \subseteq A'.$$

Similarly for $p > q$ and $p \geq q$.

[2] A is a *proper* subset of A' because $r \in A'$, while $r \notin A$, by assumption.

[3] p is said to be "old" if $p \in R$ and "new" if $p \notin R$.

The relation "$<$" so defined is trichotomic on $\overline{R}$ by Theorem 2. It is also transitive (for so is $\subset$). Thus, it makes $\overline{R}$ an *ordered set*. Moreover, it agrees with the *original* ordering of R if p, $q \in R$. Indeed, in this case (A, B) and (A', B') are *not* gaps, and so

$$A = \{x \in R \mid x \leq p\}, \quad A' = \{x \in R \mid x \leq q\}.$$

Hence it easily follows (by Corollary 3 of §9) that $A \subseteq A'$ iff $p \leq q$, and $A \subset A'$ iff $p < q$, under the *original* meaning of "$p < q$" in R. (Verify!)

Theorem 3. *For any $p \equiv (A, B)$ in R,*

$$p = \sup A = \inf B.$$

If, further, $p < q$ in $\overline{R}$, there always are $x, y \in R$ such that

$$p \leq x < y \leq q.$$

Proof. All this is trivial if $p \in R$, i.e., if (A, B) is not a gap. Thus, we assume $A \cap B = \emptyset$, i.e., p is a "new" element.

First we show that p is an upper bound of A. Take any element $r \in A$. As $A \subset R$, $r \in R$; so r is determined by a cut (A'', B'') (no gap!), with $r = \max A''$. Hence,

$$(\forall x \in A'') \quad x \leq r \in A,$$

implying

$$(\forall x \in A'') \quad x \in A.$$

Thus, $A'' \subseteq A$, i.e., $r \leq p$ (by Definition 2). As this holds for *any* $r \in A$, p is indeed an upper bound of A. Similarly it is shown that p is a lower bound of B. We shall briefly say that p "bounds" A and B.

As the next step, let $p < q \equiv (A', B')$. Then, by definition, $A \subset A'$; so we can find some $y \in A' - A$. As $y \notin A$, we have $y \in B$; so, by what was proved above, $p \leq y \leq q$ (for q bounds A', and $y \in A'$). Moreover, as (A, B) is a gap, B has no minimum in R; thus, B must also contain some $x < y$, so that

$$p \leq x < y \leq q.$$

This proves the second clause of the theorem. It also shows that *no $q > p$ can be a lower bound of B* (for it exceeds some $x \in B$). Thus, p is the *greatest* lower bound of B, i.e., $p = \inf B$. Similarly for $p = \sup A$. $\square$

Note 1 *It follows that if a set $M \subseteq \overline{R}$ has an upper (lower) bound q in $\overline{R}$, then M must also have such a bound in R.* For example, if $q \equiv (A', B')$ is an upper bound, then $(\forall b \in B')\ q \leq b$; so b is another bound of M, and $b \in R$.

Theorem 4. *The completion $\overline{R}$ of any ordered set R is a complete ordered set.* (This justifies the name "completion".)

Proof. The fact that $\overline{R}$ is an ordered set was established above. We only have to show that any nonempty right-bounded subset M of $\overline{R}$ has a supremum in $\overline{R}$.

Now, by Note 1, such an M has upper bounds *belonging to* R. Let $B \neq \emptyset$ be the set of all *such* upper bounds on M, so that $B \subseteq R$. In turn, let A be the set of all lower bounds of B *in* R. (They exist, by Note 1, for B has left bounds in M.) As is easily seen, (A, B) is a cut in R; so it determines an element $p \equiv (A, B)$. We shall show that $p = \sup M$.

Indeed, by Theorem 3, $p = \inf B$; so p is *not less* than any lower bound of B, e.g., any $m \in M$. Thus $(\forall m \in M)\ m \leq p$; i.e., *$p$ is an upper bound of M*. Now, seeking a contradiction, suppose there is a *smaller* upper bound r, $r < p$. Then, again by Theorem 3, $r \leq x < p$ for some $x \in R$. Hence, x too is an upper bound of M, and since $x \in R$, it must belong to B, by the definition of B. But this is impossible since $x < p = \inf B$. This contradiction shows that p is the *least* upper bound of M, $p = \sup M$. $\square$

II. Thus far we have only assumed that R is an ordered *set*. Now suppose that it is an *ordered field*. Then we not only can construct the complete ordered set $\overline{R}$ as above but also define operations in it, as follows.

Definition 3.

Let R be an ordered field and let $\overline{R}$ be as above. Assuming that $p, q \in \overline{R}$, $p \equiv (A, B)$, and $q \equiv (A', B')$, we have the following:

(i) We define

$$p + q = \inf(B + B'),$$

where $B + B'$ is the set of all sums $x + y$, with $x \in B$ and $y \in B'$. (These sums are *defined* in R, since $B \subset R$ and $B' \subset R$.)

(ii) We define

$$-p = \inf(\sim A),$$

where $\sim A$ is the set of all additive inverses $-x$ of elements $x \in A$ (similarly for $\sim B$).

Note that $\sim B$ is exactly the set of all lower bounds *of* $\sim A$ in R, and, conversely, $\sim A$ consists of all upper bounds of $\sim B$. Thus *$(\sim B, \sim A)$ is a cut*. By Theorem 3, the element determined by $(\sim B, \sim A)$ equals $\inf(\sim A)$, i.e., $-p$. Thus $-p \equiv (\sim B, \sim A)$.

(iii) If $p > q$ and $q > 0$, we define

$$pq = \inf(BB'),$$

where BB' is the set of all products xy, with $x \in B$, $y \in B'$.

We also put $p \cdot 0 = 0 \cdot p = 0$. In case $p < 0$, $q < 0$, we put

$$pq = (-p)(-q).$$

If $p < 0 < q$, we define $pq = -((-p)q)$, and if $q < 0 < p$, then

$$pq = -(p(-q)),$$

so as to preserve the rule of signs. This reduces everything to the positive case; for if $p < 0$, then $-p > 0$, as easily follows from part (ii) of the definition.

(iv) If $p > 0$, we define

$$p^{-1} = \inf(A^{-1}),$$

where A^{-1} is the set of all reciprocals of *positive* elements $x \in A$. (Such elements exist if $p > 0$; why?) Finally, if $p < 0$, we put $p^{-1} = -(-p)^{-1}$.

Observe that all the infima required above exist in $\overline{R}$ because $\overline{R}$ is complete (by Theorem 4) and *all sets involved are left-bounded*. For, by Definition 1, B and B' have lower bounds $r \in A$ and $r \in A'$, respectively. Thus

$$(\forall x \in B)\ (\forall y \in B') \quad r \le x,\ r' \le y.$$

As R is a field and x, y, r, $r' \in R$, we may add the inequalities and obtain

$$r + r' \le x + y \text{ for all } x + y \text{ in } B + B';$$

so $r + r'$ is a lower bound of $B + B'$.

Also, as A is right-bounded by some $s \in B$, $-A$ is left-bounded by $-s$.

All this is still simpler in parts (iii) and (iv); for the assumption $p > 0$, $q > 0$ implies that B and B' consist of positive elements only (why?); so 0 is a lower bound of BB' in (iii), and similarly in (iv). Thus, indeed, *all the required infima are well-defined elements of* $\overline{R}$; hence so are $p + q$ and pq. *This proves the closure laws in* $\overline{R}$.

Finally note that if $p, q \in R$, then our definition of $p + q$ and pq agrees with the *original* meaning of $p + q$ and pq in the field R. For, by Theorem 3, if

$$p = \inf B \text{ and } q = \inf B',$$

then

$$p + q = \inf(B + B') \text{ and } pq = \inf(BB')$$

in R (cf. Problems 8 and 9 of §9). We can now prove our main result.

Theorem 5. *With operations and inequalities* $(<)$ *defined as above, the completion of an Archimedean field* R *is a complete ordered field.*

Proof. Closure laws, trichotomy, transitivity, and completeness have already been verified above. The easy verification of Axioms II–IV is sketched in Problems 9–12 below. It remains to verify V, VI, and IX.

Axiom V(a). Given an element $p \equiv (A, B)$ in $\overline{R}$, we must show that $p + (-p) = 0$, where $-p \equiv (\sim B, \sim A)$ by Definition 3(ii). This amounts to proving that $0 = \inf[B + (\sim A)]$, where

$$B + (\sim A) = \{y - x \mid y \in B,\ x \in A\},$$

by Definition 3(i).

Now, as (A, B) is a cut, we have

$$(\forall x \in A)\ (\forall y \in B) \quad y \ge x,$$

i.e., $y - x \ge 0$. Thus, 0 is a lower bound of the set $B + (\sim A)$, and we must only show that 0 is the *greatest* lower bound.

Seeking a contradiction, suppose there is a *larger* lower bound, $r > 0$. Then, fixing any $x \in A$, we have

$$(\forall y \in B) \quad r \le y - x$$

since r is a lower bound of *all* such $y - x$. Thus

$$(\forall y \in B) \quad y \ge r + x,$$

i.e., $r + x$ is a lower bound of B, and hence $r + x \in A$, by the definition of a cut. We see that

$$(\forall x \in A) \quad r + x \in A.$$

As this applies to *any* $x \in A$, we may replace x by $r + x$, and thus obtain $r + (r + x) = 2r + x \in A$. Repeating this process, we obtain

$$nr + x \in A, \quad n = 1, 2, \ldots$$

for any $x \in A$; hence $nr + x \le y$ for $y \in B$ (for each $y \in B$ is an upper bound of A). Thus, fixing $x \in A$ and $y \in B$, we get

$$nr \le y - x, \quad n = 1, 2, \ldots$$

contrary to the assumed Archimedean property of R. This contradiction shows that, indeed, $B + (\sim A)$ has no lower bounds > 0, and completes the proof.

Axiom V(b) is proved quite analogously in case $p > 0$. One only has to replace everywhere addition and subtraction by multiplication and division. Accordingly,

$$0,\ -p,\ B + (\sim A),\ y - x,\ r + x, \text{ and } nr$$

are replaced, respectively, by

$$1,\ p^{-1},\ BA^{-1},\ \frac{y}{x},\ rx, \text{ and } r^n,$$

but essentially the argument is the same. Note that the binomial expansion yields

$$r^n = (1 + a)^n = 1 + na + \cdots > na$$

if we put $r = 1 + a$ (using the fact that $r > 1$ here). Thus, by the Archimedean property

$$r^n > na > \frac{y}{x}$$

for large n, and this yields the required contradiction in the last part of the proof. The details are left to the reader. Finally, the proof for $p < 0$ easily follows from the positive case, by the formula $p^{-1} = -(-p)^{-1}$. Thus Axiom V is verified in full.

Axiom VI. Let

$$p \equiv (A, B),\ q \equiv (A', B'),\ r \equiv (A'', B'').$$

We must show that $(p + q)r = pr + qr$.

Assume first that p, q, $r > 0$. Then it easily follows that

$$(p + q)r = \inf[(B + B')B''] \text{ and } pr + qr = \inf(BB'' + B'B'');$$

cf. Problem 10(c). Thus all reduces to proving that

$$(B + B')B'' = BB'' + B'B''.$$

But, by definition, the elements of $(B + B')B''$ have the form $(b + b')b''$, and those of $BB'' + B'B''$ have the form $bb'' + b'b''$ ($b \in B$, $b' \in B'$, $b'' \in B''$, all in R). Thus, by the distributive law for R (a *field*, by assumption),

$$(b + b')b'' = bb'' + b'b'',$$

and so the sets $(B + B')B''$ and $BB'' + B'B''$ coincide. This settles the case p, q, $r > 0$.

Moreover, if $p > q > 0$ and $r > 0$, we also have

$$(p - q)r + qr = [(p - q) + q]r,$$

by what was proved above (replacing p by $p - q$). Hence $(p - q)r = pr - qr$.

This holds also if $q > p > 0$, since

$$(p - q)r = -[(q - p)r] = -(qr - pr) = pr - qr.$$

Thus

$$(p \pm q)r = pr \pm qr$$

for p, q, $r > 0$. Now also the other cases can be handled. For example, if $p > 0 > q$ and $r > 0$, then $-q > 0$ and

$$(p + q)r = [p - (-q)]r = pr - (-q)r = pr + qr.$$

Axiom IX(a). Let again $p \equiv (A, B)$, $q \equiv (A', B')$, $r \equiv (A'', B'')$, $p > q$. We must show that

$$p + r > q + r.$$

Now, by Definition 2 and Theorem 2, $p > q$ implies $A \supset A'$ and $B' \supset B$; hence

$$B' + B'' \supseteq B + B''.$$

(Verify!) Thus, by Corollary 3 in §9,

$$\inf(B + B'') \geq \inf(B' + B''),$$

i.e., $p + r \geq q + r$. Equality is excluded here, for $p + r = q + r$ would imply $p = q$ (by Axioms III–V, which we assume as proved already for $\overline{R}$), and this is contrary to $p > q$. Thus $p + r > q + r$, as claimed.

Axiom IX(b) is proved similarly for p, $q > 0$ and is obvious if $p = 0$ or $q = 0$.

In the general case (with $r > 0$, always), $p > q$ implies $p - q > 0$, whence

$$(p - q)r > 0 \cdot r = 0,$$

i.e., $pr > qr$, by distributivity (Axiom VI). □

Thus the theorem is proved. In particular, *we can apply it to the field R of all rational numbers* (for R is Archimedean).[4] By Theorem 5, the completion $\overline{R}$ of R satisfies all axioms valid for real numbers, and so we may simply *define* E^1 to be $\overline{R}$. In this case, the "old" elements of $\overline{R}$ are the rationals, and hence the "new" ones are the *irrationals*.

Problems on Dedekind Cuts

1. Prove that in any cut (A, B),

$$y \leq x \in A \text{ implies } y \in A, \text{ and } y \geq x \in B \implies y \in B.$$

2. Verify that (A, B) in Example 2 is a cut.

3. Prove that if (A, B) and (A', B') are cuts, then $A \subset A'$ iff $B \supset B'$.

4. Prove in detail the assertions immediately preceding Theorem 3.

5. Complete the proof of Theorem 3 by showing that p is a lower bound of B. Also, carry out the proof for the case $p \in R$, $q \notin R$.

6. Prove for *any* $p \in \overline{R}$ that

$$p \equiv (A, B) \text{ iff } A = \{x \in R \mid x \leq p\} \text{ and } B = \{x \in R \mid x \geq p\}.$$

[4] For if $x, y \in R$ and $x, y > 0$, then $y/x \in R$ and so $y/x = m/n$ for some $m, n \in N$. Hence $y/x \leq m$, i.e., $y \leq mx < (m + 1)x$, and the Archimedean property follows.

7. Complete the proof of Theorem 4 by showing that (A, B) is indeed a cut.

8. Prove that $(\sim B, \sim A)$ is a cut if (A, B) is, and that $p < 0$ iff $-p > 0$.

9. From Definitions 3(i) and (iii) prove the following:

(a) $B + B' = \{x \in R \mid x \geq p + q\}$ if $p, q \in R$; and $B + B' = \{x \in R \mid x > p + q\}$ if $p \notin R$ or $q \notin R$.

(b) If $p, q > 0$, then

$$BB' = \{x \in R \mid x \geq pq\}$$

if $p, q \in R$, and

$$BB' = \{x \in R \mid x > pq\}$$

otherwise. Hence we infer the following.

(c) *$p + q$ determines a cut (A^*, B^*) in which $B^* = (B + B') \cup \{p + q\}$, or $B^* = B + B'$* (cf. Problem 6); *similarly for pq if $p, q > 0$.*
[Hint for (a): First show that

$$\{x \in R \mid x > p + q\} \subseteq B + B':$$

Let $x \in R$, $x > p + q$; or $x > \inf(B + B')$ (by the definition of $p + q$). Then x is *not* a lower bound of $B + B'$ (why?); so $x > b + b'$ for some $b \in B$ and $b' \in B'$. Let $t = x - b'$; so $t + b' = x > b + b'$. Hence $t \in R$ and $t > b \in B$, *implying* $t \in B$ (cf. Problem 1). Thus $x = t + b'$, with $t \in B$ and $b' \in B$, i.e., $x \in B + B'$, as required.

Next, prove the converse inclusion in case $p \notin R$ or $q \notin R$. Finally, consider the case $p, q \in R$.]

10. Using the results of Problem 9, prove that if $p \equiv (A, B)$, $q \equiv (A', B')$, and $r \equiv (A'', B'')$, then the following are true.

(a) $(p + q) + r = \inf[(B + B') + B''] = \inf[B + (B' + B'')] = p + (q + r)$. (First show that $(B + B') + B'' = B + (B' + B'')$.)

(b) $(pq)r = p(qr)$. (First assume p, q, and r are greater than 0, then extend this to *all* $p, q, r \in \overline{R}$ by the rule of signs.)

(c) $(p + q)r = \inf[(B + B')B'']$ and $pr + qr = \inf(BB'' + B'B'')$.

[Hint: Observe that

$$\inf(B + B') = \inf[(B + B') \cup \{p + q\}].$$

(Why?) Thus, it does not matter whether $p + q \in B + B'$. Hence, using Problem 9(c), we may safely assume that

$$p + q \equiv (A^*, B^*),$$

with $B^* = B + B'$, disregarding the case $B^* = (B + B') \cup \{p + q\}$. Then, by Definition 3(i),

$$(p + q) + r = \inf(B^* + B'') = \inf[(B + B') + B''],$$

etc.]

11. Show that $(\forall p, q \in \overline{R})$

$$pp^{-1} = 1, \ p + q = q + p, \text{ and } pq = qp.$$

12. Verify Axiom IV for $\overline{R}$: $p + 0 = p$ and $p \cdot 1 = p$.
[Hint: 0 corresponds to a cut (A_0, B_0) with $B_0 = \{x \in R \mid x \geq 0\}$. If $p \equiv (A, B)$, then $p = \inf B$, by Theorem 3. Show that $\inf B = \inf(B + B_0)$, since $(\forall x \in B_0)$ $x \geq 0$ and so $b + x \geq b$.]

13. Prove *Dedekind's theorem*: An ordered set is complete iff it has no gaps.

§16. The Infinities. *The $\underline{\lim}$ and $\overline{\lim}$ of a Sequence

I. As we know, a set $A \neq \emptyset$ in E^1 has a l.u.b. (g.l.b.) if A is bounded above (below, respectively), *but not otherwise*. In order to avoid this inconvenient restriction, we now add to E^1 two new objects of arbitrary nature ("two pebbles") and call them "*minus infinity*" $(-\infty)$ and "*plus infinity*" $(+\infty)$, with the convention that $-\infty < +\infty$ and $-\infty < x < +\infty$ for all $x \in E^1$.

It is readily seen that, with this convention, the laws of trichotomy and transitivity (Axioms VII and VIII) remain valid. The set consisting of all reals and the two infinities is called the *extended real number system*. We denote it by E^* and call its elements *extended real numbers*. The ordinary reals are also called *finite* numbers, while $\pm\infty$ are the only two *infinite* elements of E^*. (**Caution**: They are *not* real numbers. *E^* is not a field.*)

At this stage we do not define any operations involving $\pm\infty$ (though this can be done). However, the notions of upper and lower bound, maximum, minimum, supremum, and infimum *are* defined in E^* exactly as in E^1. In particular, $-\infty = \min E^*$ and $+\infty = \max E^*$. Thus, in E^*, *all* sets are bounded by $-\infty$ and $+\infty$.[1]

It follows that *in E^* every set $A \neq \emptyset$ has a l.u.b. and a g.l.b.* For if A has no upper bound in E^1, it still has the upper bound $+\infty$ *in* E^*, which in this case is the *unique* (hence also the *least*) upper bound; thus $\sup A = +\infty$.[2] It is also customary to define $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$ (this is the *only* case where $\sup A < \inf A$). *All properties of l.u.b. and g.l.b. stated in §9 remain valid in E^**, with the same proof. The only exception is Note 4, since $+\infty - \epsilon$ and $-\infty + \epsilon$ make no sense.

[1] Therefore, when speaking of "bounded" sets in E^*, one usually has in mind those bounded *in* E^1, i.e., having *finite* bounds.

[2] Unless A consists of $-\infty$ *alone*, in which case $\sup A = -\infty$. Similarly, $\infty = \inf A$ if there is no other lower bound.

We can now define *intervals* in E^* exactly as in E^1 (see §8), allowing also infinite values of a, b, x. Thus

$$(-\infty, a) = \{x \in E^* \mid -\infty < x < a\} = \{x \in E^1 \mid x < a\},$$
$$[a, +\infty) = \{x \in E^* \mid a \le x < +\infty\},$$
$$(-\infty, \infty) = \{x \in E^* \mid -\infty < x < \infty\} = E^1,$$
$$[-\infty, +\infty] = \{x \in E^* \mid -\infty \le x \le +\infty\} = E^*,$$

etc. Intervals with *finite* endpoints are said to be finite; all other intervals are called infinite. If $a \in E^1$, the intervals $(-\infty, a)$, $(-\infty, a]$, $(a, +\infty)$, $[a, \infty)$ are actually subsets of E^1, as is $(-\infty, +\infty)$. Thus we may speak of infinite intervals *in* E^1 as well.

***II. Upper and Lower Limits.**[3] We have already mentioned that a real number p is called the *limit* of a sequence $\{x_n\} \subseteq E^1$ ($p = \lim\limits_{n\to\infty} x_n$) iff

$$(\forall \epsilon > 0)\ (\exists k)\ (\forall n > k) \quad |x_n - p| < \epsilon, \text{ i.e., } p - \epsilon < x_n < p + \epsilon; \tag{1}$$

in this definition, ϵ is in E^1 and n and k are in N.

This may be stated thusly: "For sufficiently large n ($n > k$), x_n becomes and *stays* as close to p as we like ('ϵ-close')." We also define the following:

$$\lim_{n\to\infty} x_n = +\infty \iff (\forall a \in E^1)\ (\exists k)\ (\forall n > k) \quad x_n > a, \tag{2}$$

and

$$\lim_{n\to\infty} x_n = -\infty \iff (\forall b \in E^1)\ (\exists k)\ (\forall n > k) \quad x_n < b. \tag{3}$$

Note that (2) and (3) make sense *in* E^1, *too*, since the symbols $\pm\infty$ do not occur on the right side of the formulas. Formula (2) means that x_n becomes *arbitrarily large* (larger than any $a \in E^1$ given in advance) for sufficiently large n ($n > k$). The interpretation of (3) is analogous. We shall now develop a more general and unified approach for E^*, allowing *infinite* terms x_n, too.

Let $\{x_n\}$ be any sequence in E^*. For each n, let A_n consist of all terms *from* x_n *onward*:

$$A_n = \{x_n, x_{n+1}, \dots\}.$$

Thus,

$$A_1 = \{x_1, x_2, \dots\},\ A_2 = \{x_2, x_3, \dots\},\ \text{etc.}$$

The A_n form a *contracting sequence* (Chapter 1, §8), as $A_1 \supseteq A_2 \supseteq \cdots$.

Now, for each n let

$$p_n = \inf A_n \text{ and } q_n = \sup A_n,$$

[3] Before taking up this topic, the reader should review §§8 and 3 (quantifiers) of Chapter 1.

also denoted

$$p_n = \inf_{k\ge n} x_k,\ q_n = \sup_{k\ge n} x_k.$$

(These infima and suprema *always* exist in E^*, as noted above.) Since $A_n \supseteq A_{n+1}$, Corollary 3 of §9 yields

$$\inf A_n \le \inf A_{n+1} \le \sup A_{n+1} \le \sup A_n.$$

Thus,

$$p_1 \le p_2 \le \cdots \le p_n \le p_{n+1} \le \cdots \le q_{n+1} \le q_n \le \cdots \le q_2 \le q_1, \tag{4}$$

and so $\{p_n\}\uparrow$, while $\{q_n\}\downarrow$ in E^*. Also, *each q_m is an upper bound of all p_n* and hence $q_m \ge \sup_n p_n$ (= l.u.b. of all p_n). It follows that this l.u.b. (call it $\underline{L}$) is a *lower bound of all q_m*, and so

$$\underline{L} \le \inf_m q_m.$$

We set $\overline{L} = \inf_m q_m$.

Definition 1.

For each sequence $\{x_n\} \subseteq E^*$, we define its *upper limit* $\overline{L}$ and its *lower limit* $\underline{L}$, denoted

$$\overline{L} = \overline{\lim}\, x_n \ (\text{or } \limsup_{n\to\infty} x_n) \text{ and } \underline{L} = \underline{\lim}\, x_n = \liminf_{n\to\infty} x_n,$$

as follows. We put

$$(\forall n) \quad q_n = \sup_{k\ge n} x_k \text{ and } p_n = \inf_{k\ge n} x_k,$$

as before. Then we set

$$\overline{L} = \overline{\lim}\, x_n = \inf_n q_n \text{ and } \underline{L} = \underline{\lim}\, x_n = \sup_n p_n, \text{ all in } E^*. \tag{5}$$

Here and below, $\inf_n q_n$ is the inf of *all* q_n, and $\sup_n p_n$ is the sup of *all* p_n.

Corollary 1. *For any sequence in E^*,*

$$\inf_n x_n \le \underline{\lim}\, x_n \le \overline{\lim}\, x_n \le \sup_n x_n.$$

For, as we noted before,

$$\underline{L} = \sup_n p_n \le \inf_m q_m = \overline{L}.$$

Also,

$$\underline{L} \ge p_n = \inf A_n \ge \inf A_1 = \inf_n x_n \text{ and}$$
$$\overline{L} \le q_n = \sup A_n \le \sup A_1 = \sup_n x_n,$$

with A_n as above.

Examples.

(a) $x_n = 1/n$, $n = 1, 2, \dots$. Here

$$q_1 = \sup\Big\{1, \frac{1}{2}, \dots, \frac{1}{n}, \dots\Big\} = 1,\ q_2 = \frac{1}{2},\ q_n = \frac{1}{n}.$$

Hence

$$\overline{L} = \inf_n q_n = \inf\Big\{1, \frac{1}{2}, \dots, \frac{1}{n}, \dots\Big\} = 0,$$

as easily follows by Theorem 2, §§8–9, and the Archimedean property. (Verify!) Also,

$$p_1 = \inf_{k\geq 1} \frac{1}{k} = 0,\ p_2 = \inf_{k\geq 2} \frac{1}{k} = 0, \dots,\ p_n = \inf_{k\geq n} \frac{1}{k} = 0.$$

Since all p_n are 0 so is $\underline{L} = \sup_n p_n$. Thus, here $\underline{L} = \overline{L} = 0$.

(b) Consider the sequence

$$1, -1, 2, -\frac{1}{2}, \dots, n, -\frac{1}{n}, \dots.$$

Here

$$p_1 = -1 = p_2,\ p_3 = -\frac{1}{2} = p_4, \dots;\ p_{2n-1} = -\frac{1}{n} = p_{2n}.$$

Thus

$$\underline{\lim}\, x_n = \sup_n p_n = \sup\Big\{-1, -\frac{1}{2}, \dots, -\frac{1}{n}, \dots\Big\} = 0.$$

On the other hand, $q_n = +\infty$ for all n. (Why?) Thus,

$$\overline{\lim}\, x_n = \inf_n q_n = +\infty. \quad \text{(Why?)}$$

Theorem 1.

(i) *If $x_n \geq b$ for infinitely many n, then $\overline{\lim}\, x_n \geq b$ as well.*

(ii) *If $x_n \leq a$ for all but finitely many n,*[4] *then $\overline{\lim}\, x_n \leq a$ as well.*

Similarly for lower limits (with all inequalities reversed).

Proof. (i) If $x_n \geq b$ for infinitely many n, then such x_n must occur in *each* set $A_m = \{x_m, x_{m+1}, \dots\}$. Hence $(\forall m)\ q_m = \sup A_m \geq b$; so $\overline{L} = \inf_m q_m \geq b$, by Corollary 2 of §9.

[4] In other words, for all except (at most) a finite number of terms x_n. This is stronger than just "*infinitely many n*" (allowing infinitely many *exceptions* as well). **Caution**: Avoid confusing "*all but finitely many*" with just "*infinitely many*".

(ii) If $x_n \leq a$ except for *finitely many* n, let n_0 be the *last* of these "exceptional" n. Then, for $n > n_0$, $x_n \leq a$, i.e., the set $A_n = \{x_n, x_{n+1}, \dots\}$ is bounded above by a; so $q_n = \sup A_n \leq a$. Hence, certainly $\overline{L} = \inf_n q_n \leq a$. □

Corollary 2.

(i) *If $\overline{\lim}\, x_n > a$, then also $x_n > a$ for infinitely many n.*

(ii) *If $\overline{\lim}\, x_n < b$, then $x_n < b$ for all but finitely many n.*

Similarly for lower limits (with all inequalities reversed).

Proof. Assume the opposite and find a contradiction to Theorem 1. □

To unify our definitions, we now introduce some useful notions. By a *neighborhood of p* ($p \in E^1$), briefly G_p,[5] we mean any interval of the form $(p-\epsilon, p+\epsilon)$, $\epsilon > 0$. If $p = +\infty$ (resp., $p = -\infty$), G_p is an infinite interval of the form $(a, +\infty]$ (resp., $[-\infty, b)$), with $a, b \in E^1$. We can now combine formulas (1)–(3) in one equivalent definition.

Definition 2.

An element $p \in E^*$ (finite or not) is called the *limit* of a sequence $\{x_n\} \subset E^*$ if each G_p (no matter how small it is) contains all but finitely many x_n, i.e., all x_n from some x_k onward.

In symbols,

$$(\forall G_p)\ (\exists k)\ (\forall n > k) \quad x_n \in G_p. \quad \Big(\text{Notation: } p = \lim x_n \text{ or } \lim_{n\to\infty} x_n.\Big) \tag{6}$$

Indeed, if $p \in E^1$, then $x_n \in G_p$ means that $p - \epsilon < x_n < p + \epsilon$, as in (1). If, however, $p = +\infty$ (resp., $p = -\infty$), it means that $x_n > a$ (resp., $x_n < b$), as in (2) and (3).

Theorem 2. *We have $q = \overline{\lim}\, x_n$ in E^* iff these two conditions hold:*

(i′) *Each neighborhood G_q contains x_n for infinitely many n.*

(ii′) *If $q < b$, then $x_n \geq b$ for at most finitely many n.*[6]

Proof. If $q = \overline{\lim}\, x_n$, Corollary 2 yields (ii′). It also shows that any interval (a, b), with $a < q < b$, contains *infinitely many* x_n (for there are infinitely many $x_n > a$, and only finitely many $x_n \geq b$, by (ii′)).

Now, if $q \in E^1$, $G_q = (q-\epsilon, q+\epsilon)$ is such an interval; so we obtain (i′). The cases $q = \pm\infty$ are analogous; we leave them to the reader.

Conversely, assume (i′) and (ii′). Seeking a contradiction, let $q < \overline{L}$; say, $q < b < \overline{\lim}\, x_n$. Then Corollary 2(i) yields $x_n > b$ for infinitely many n, contrary to our assumption (ii′). Similarly, $q > \overline{\lim}\, x_n$ would contradict (i′). Thus necessarily $q = \overline{\lim}\, x_n$. □

[5] This terminology and notation anticipates some more general ideas.

[6] A similar theorem (with all inequalities reversed) holds for $\underline{\lim}\, x_n$.

Theorem 3. *We have*

$$q = \lim x_n \text{ in } E^* \text{ iff } \underline{\lim}\, x_n = \overline{\lim}\, x_n = q.$$

Proof. Suppose $\underline{\lim}\, x_n = \overline{\lim}\, x_n = q$. If $q \in E^1$, then every G_q is an interval (a,b), $a < q < b$; so Corollary 2(ii) and its analogue for $\underline{\lim}\, x_n$ imply (with q treated as *both* $\overline{\lim}\, x_n$ and $\underline{\lim}\, x_n$) that $a < x_n < b$ for *all but finitely many* n. Thus, by Definition 2, $q = \lim x_n$, as claimed.

Conversely, if $q = \lim x_n$, then any G_q (no matter how small) contains all but finitely many x_n. Hence, *so does any interval* (a,b) *with* $a < q < b$; for it contains some small G_q. Now, exactly as in the proof of Theorem 2, one *excludes* $q \neq \overline{\lim}\, x_n$ and $q \neq \underline{\lim}\, x_n$. This settles the case $q \in E^1$. The cases $q = \pm\infty$ are quite analogous. □

Problems on Upper and Lower Limits of Sequences in E^*[7]

1. Complete the missing details in the proofs of Theorems 2 and 3, Corollary 1, and Examples (a) and (b).

2. State and prove the analogues of Theorems 1 and 2, and Corollary 2, for $\underline{\lim}\, x_n$.

3. Find $\overline{\lim}\, x_n$ and $\underline{\lim}\, x_n$ if

(a) $x_n = c$ (constant);

(b) $x_n = -n$;

(c) $x_n = n$;

(d) $x_n = (-1)^n n - n$.

Does $\lim x_n$ exist in each case?

⇒4. A sequence $\{x_n\}$ is said to *cluster* at $q \in E^*$, and q is called its *cluster point*, iff each G_q contains x_n for infinitely many values of n. Show that both $\underline{L}$ and $\overline{L}$ are cluster points ($\underline{L}$ the *least* and $\overline{L}$ the *largest*).

[Hint: Use Theorem 2, and its analogue for $\underline{L}$. To show that no $p < \underline{L}$ (or $q > \overline{L}$) is a cluster point, assume the opposite and find a contradiction to Corollary 2.]

⇒5. Prove that

(i) $\overline{\lim}(-x_n) = -\underline{\lim}\, x_n$;

(ii) $\overline{\lim}(ax_n) = a \cdot \overline{\lim}\, x_n$ if $0 \leq a < +\infty$.

6. Prove that $\overline{\lim}\, x_n < +\infty$ ($\underline{\lim}\, x_n > -\infty$) iff $\{x_n\}$ is bounded above (below) in E^1.

[7] The problems marked by ⇒ are theoretically important. Study them!

7. If $\{x_n\}$ and $\{y_n\}$ are *bounded in* E^1, then

$$\begin{aligned}\overline{\lim}\, x_n + \overline{\lim}\, y_n &\geq \overline{\lim}(x_n + y_n) \geq \overline{\lim}\, x_n + \underline{\lim}\, y_n \\ &\geq \underline{\lim}(x_n + y_n) \geq \underline{\lim}\, x_n + \underline{\lim}\, y_n.\end{aligned}$$

Give a proof.

⇒8. Prove that if $p = \lim x_n$ *in* E^1, then

$$\underline{\lim}(x_n + y_n) = p + \underline{\lim}\, y_n.$$

Similarly for $\overline{L}$.

⇒9. Prove that if $\{x_n\}$ is monotone, then $\lim x_n$ exists *in* E^*. Specifically, if $\{x_n\}\uparrow$ then $\lim x_n = \sup_n x_n$, and if $\{x_n\}\downarrow$ then $\lim x_n = \inf_n x_n$.

⇒10. Prove that

(i) if $\lim x_n = +\infty$ and $(\forall n)\ x_n \leq y_n$, then also $\lim y_n = +\infty$;

(ii) if $\lim x_n = -\infty$ and $(\forall n)\ y_n \leq x_n$, then also $\lim y_n = -\infty$.

11. Prove that if $x_n \leq y_n$ for all n, then

$$\underline{\lim}\, x_n \leq \underline{\lim}\, y_n \text{ and } \overline{\lim}\, x_n \leq \overline{\lim}\, y_n.$$

Chapter 3

The Geometry of n Dimensions
*Vector Spaces

§1. Euclidean n-Space, E^n

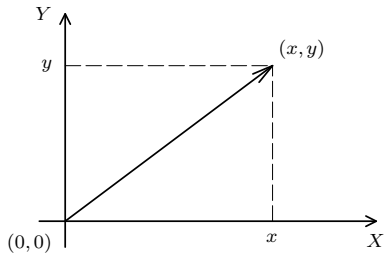


FIGURE 12

The reader is certainly familiar with the representation of ordered pairs of real numbers (x, y) as points in the xy-plane. Because of this representation, such pairs are often called "*points*" *of the Cartesian plane* (each pair being regarded as *one* "point"). The set of all such pairs is, by definition, the Cartesian product (or cross product) $E^1 \times E^1$, also briefly denoted by E^2. An ordered pair $(x, y) \in E^2$ can also be graphically represented as a *directed line segment* ("vector") passing from the origin $(0, 0)$ to (x, y) (see Figure 12). Therefore, such pairs are also called "*vectors*" in E^2.

Quite similarly, *ordered triples* (x, y, z) of real numbers are called "points" or "vectors" of the three-dimensional space $E^3 = E^1 \times E^1 \times E^1$. Nothing prevents us also from considering the set E^n of all ordered *n-tuples* of real numbers (with n fixed). Though in n dimensions there is no actual geometric representation, it is convenient to use the geometric *language* in this case, too. Thus every ordered n-tuple of real numbers

$$(x_1, x_2, \ldots, x_n)$$

will also be called a "*point*" or "*vector*" in E^n, and the single numbers

$$x_1, x_2, \ldots, x_n$$

of which it is composed are called its *coordinates* or *components*. E^n itself is called *n-dimensional Euclidean space*, briefly, "*n-space*". A point in E^n will

often be denoted by a *single* letter (preferably with a bar or arrow above it), and then its n coordinates will be denoted by *the same* letter, with corresponding subscripts (but without the bar or arrow). Thus we write

$$\vec{x} = (x_1, x_2, \ldots, x_n),\ \bar{u} = (u_1, u_2, \ldots, u_n),\ \text{etc.};$$

the notation $\bar{x} = (0, -1, 2, 4)$ means that $\bar{x}$ is a point (vector) in E^4, with coordinates 0, -1, 2, and 4 (in this order). In E^2 and E^3, we shall also sometimes use x, y, z to denote the coordinates; e.g., $\vec{v} = (x, y, z) \in E^3$, or $\bar{u} = (x, y) \in E^2$. It should be well noted that the term "point" or "vector" means the *n-tuple*, and not its graphical representation ("dot" or "line segment"); a drawing may not be used at all. The formula $\bar{x} \in E^n$ means that $\bar{x}$ is a point in E^n, i.e., an n-tuple, namely $(x_1, x_2, \ldots, x_n)$.

As we know, two ordered n-tuples are equal only if the *corresponding* coordinates are the same. Thus *two vectors (points) $\vec{x}$ and $\vec{y}$ in E^n are equal iff they have the same corresponding components*, i.e., if

$$x_1 = y_1,\ x_2 = y_2,\ \ldots,\ x_n = y_n,$$

but not if the components occur in different order; e.g., $(4, 2, 1) \neq (2, 1, 4)$. **Note:** *One vector equation is equivalent to n coordinate equations.*

The point whose coordinates are all 0 is called the *origin* or the *zero-vector*, denoted by $\vec{0}$ or $\bar{0}$. Thus $\vec{0} = (0, 0, \ldots, 0)$ (n times). The vector whose k-th coordinate is 1 and whose remaining $n-1$ coordinates are 0 is called the *k-th basic unit vector*, denoted by $\vec{e}_k$; there are exactly n such vectors, namely,

$$\vec{e}_1 = (1, 0, 0, \ldots, 0),\ \vec{e}_2 = (0, 1, 0, \ldots, 0),\ \ldots,\ \vec{e}_n = (0, 0, \ldots, 0, 1).$$

In E^2, we often denote these vectors by $\vec{\imath}$ and $\vec{\jmath}$; in E^3, we denote them by $\vec{\imath}$, $\vec{\jmath}$, and $\vec{k}$, respectively.

The term "vector" (rather than "point") is preferably used when certain *operations* are involved, which we shall define next; single real numbers are then called *scalars*. **Note:** *No scalar can be equal to a vector* in E^n (since the latter is an *n-tuple*), except if $n = 1$ (i.e., if we consider E^1 itself as our "space"). Also note that *the n components of a vector in E^n are scalars*, not vectors. Sometimes we write $\overrightarrow{0x}$ for a vector $\vec{x}$ (especially when we think of $\vec{x}$ as represented by a directed line segment); $\overrightarrow{0x}$ is often called the "*position vector*" of the "point" $\bar{x}$. In our theory, it is just another name for the vector (point) $\vec{x}$ itself.

Definition 1.

Given two vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$ in E^n, we define their *sum* and *difference* to be the vector whose coordinates are obtained by adding or subtracting, respectively, the corresponding coordinates of x and y; thus

$$\vec{x} \pm \vec{y} = (x_1 \pm y_1, x_2 \pm y_2, \ldots, x_n \pm y_n).$$

Similarly for the sum of three or more vectors. Instead of $\vec{0} - \vec{x}$ (where $\vec{0}$ is the zero-vector), we simply write $-\vec{x}$, and we call $-\vec{x}$ the *additive inverse of $\vec{x}$*, or the *vector inverse to $\bar{x}$*. The reader will note that this definition agrees with the familiar geometric rule of constructing the sum of two vectors, in E^2 or E^3, as the diagonal of the parallelogram whose sides are these vectors, represented as directed line segments. Imitating the usual geometric terminology, we shall also call $\vec{x} - \vec{y}$ the "*vector passing from the point $\vec{y}$ to the point $\vec{x}$*" and denote it also by $\overrightarrow{yx}$. Thus $\overrightarrow{yx} = \vec{x} - \vec{y}$, by definition. In particular, this agrees with our notation $\vec{x} = \overrightarrow{0x} = \vec{x} - \vec{0}$.

By our definitions,

$$-\vec{x} = (0 - x_1, 0 - x_2, \ldots, 0 - x_n) = (-x_1, -x_2, \ldots, -x_n).$$

Thus *the coordinates of $-\vec{x}$ are exactly the additive inverses of the corresponding coordinates of $\vec{x}$.*

Definition 2.

Given a vector $\vec{x} = (x_1, \ldots, x_n)$ in E^n and a scalar $a \in E^1$, we define the *product of a by $\vec{x}$* to be the vector

$$a\vec{x} = (ax_1, ax_2, \ldots, ax_n),$$

i.e., the vector whose coordinates are products of a by the corresponding coordinates of $\vec{x}$.

Instead of $\left(\frac{1}{a}\right)\vec{x}$ we sometimes write $\frac{\vec{x}}{a}$ (here a must be a scalar $\neq 0$).

Caution: We have as yet no definition for a product of *two vectors*, only for the product of a *scalar* by a vector. Such products are also called *scalar multiples* of the given vector $\vec{x}$.

Examples.

If $\vec{u} = (0, -1, 4, 2)$, $\vec{v} = (2, 2, -3, 1)$, and $\vec{w} = (1, 5, 4, 2)$ are vectors in E^4, then

(1) $\vec{u} + \vec{v} + \vec{w} = (3, 6, 5, 5)$, $\vec{u} - \vec{w} = (-1, -6, 0, 0)$;

(2) $2\vec{u} = (0, -2, 8, 4)$, $1\vec{v} = (2, 2, -3, 1) = \vec{v}$;

(3) $3\vec{e}_1 = 3(1, 0, 0, 0) = (3, 0, 0, 0)$;

(4) $5\vec{e}_2 = (0, 5, 0, 0)$, $\frac{1}{2}\vec{u} = (0, -\frac{1}{2}, 2, 1)$;

(5) $3\vec{e}_1 + 2\vec{e}_2 - 5\vec{e}_3 + \vec{e}_4 = (3, 2, -5, 1)$, $3\vec{u} - 2\vec{v} + 5\vec{w} = (1, 18, 38, 14)$;

(6) $0\vec{u} = 0\vec{v} = 0\vec{w} = (0, 0, 0, 0) = \vec{0}$;

(7) $(-1)\vec{u} = (0,\ 1,\ -4,\ -2) = -\vec{u}$;

(8) $\vec{u} + (-\vec{u}) = (0,\ 0,\ 0,\ 0) = \vec{0}$.

Theorem 1. *For any vectors $\vec{u}$, $\vec{v}$, $\vec{w}$ in E^n and any scalars $a, b \in E^1$, we have the following:*

(a) $\vec{u} + \vec{v}$ *and* $a\vec{v}$ *are vectors in* E^n (closure laws);

(b) $\vec{u} + \vec{v} = \vec{v} + \vec{u}$ (commutativity of vector addition);

(c) $\vec{u} + (\vec{v} + \vec{w}) = (\vec{u} + \vec{v}) + \vec{w}$ (associativity of addition);

(d) $\vec{u} + \vec{0} = \vec{0} + \vec{u} = \vec{u}$ (i.e., $\vec{0}$ is the neutral element of vector addition);

(e) $\vec{u} + (-\vec{u}) = \vec{0}$ ($-\vec{u}$ is the additive inverse of $\vec{u}$);

(f) $a(\vec{u} + \vec{v}) = a\vec{u} + a\vec{v}$; $(a + b)\vec{u} = a\vec{u} + b\vec{u}$ (distributive laws);

(g) $(ab)\vec{u} = a(b\vec{u})$;

(h) $1\vec{u} = \vec{u}$.

Proof. Assertion (a) is immediate from Definitions 1 and 2. The remaining assertions easily follow from the corresponding properties of real numbers. For example, to prove (b), let $\vec{u} = (u_1, \ldots, u_n)$, $\vec{v} = (v_1, \ldots, v_n)$. Then, by definition, we have

$$\vec{u} + \vec{v} = (u_1 + v_1,\ u_2 + v_2,\ \ldots,\ u_n + v_n)$$

and

$$\vec{v} + \vec{u} = (v_1 + u_1,\ v_2 + u_2,\ \ldots,\ v_n + u_n).$$

But the right sides in both equations coincide because of the commutativity of addition in E^1. Thus $\vec{u} + \vec{v} = \vec{v} + \vec{u}$, as required; similarly for the remaining assertions, which we leave to the reader as an exercise, along with the proofs of the next two corollaries. □

Corollary 1. $(\forall \vec{v} \in E^n)\ 0\vec{v} = \vec{0}$; *and* $(\forall a \in E^1)\ a\vec{0} = \vec{0}$.

Corollary 2. $(\forall \vec{v},\ \vec{w} \in E^n)\ (-1)\vec{v} = -\vec{v}$, *and* $\vec{v} + (-\vec{w}) = \vec{v} - \vec{w}$.

Theorem 2. *If $\vec{v} = (v_1, \ldots, v_n)$ is a vector in E^n, then*

$$\vec{v} = v_1\vec{e}_1 + v_2\vec{e}_2 + \cdots + v_n\vec{e}_n = \sum_{k=1}^{n} v_k\vec{e}_k,$$

where the $\vec{e}_k$ are the basic unit vectors in E^n. Moreover, if $\vec{v} = \sum_{k=1}^{n} a_k e_k$ for some scalars a_k, then necessarily $a_k = v_k$, $k = 1, 2, \ldots, n$.

Proof. By definition,

$$\vec{e}_1 = (1, 0, 0, \ldots, 0),\ \vec{e}_2 = (0, 1, \ldots, 0),\ \ldots,\ \vec{e}_n = (0, \ldots, 0, 1).$$

Thus

$$v_1\vec{e}_1 = (v_1, 0, \ldots, 0),\ v_2\vec{e}_2 = (0, v_2, \ldots, 0),\ \ldots,\ v_n\vec{e}_n = (0, 0, \ldots, v_n).$$

(Observe that the v_k are *scalars*; the $\vec{e}_k$ are *vectors*.)

Adding up componentwise, we obtain

$$\sum_{k=1}^{n} v_k\vec{e}_k = v_1\vec{e}_1 + v_2\vec{e}_2 + \cdots + v_n\vec{e}_n = (v_1,\ v_2,\ \ldots,\ v_n) = \vec{v},$$

as asserted. Moreover, for any other scalars $a_1, \ldots, a_n$, exactly the same procedure shows that

$$\sum_{k=1}^{n} a_k\vec{e}_k = (a_1,\ a_2,\ \ldots,\ a_n).$$

Thus, if $\vec{v} = \sum_{k=1}^{n} a_k\vec{e}_k$, then $\vec{v} = (a_1, \ldots, a_n)$. Since also $\vec{v} = (v_1, \ldots, v_n)$, the two n-tuples must *coincide*, i.e., $a_k = v_k$, $k = 1, 2, \ldots, n$, and all is proved. □

Note 1. Any sum of the form

$$\sum_{k=1}^{m} a_k\vec{x}_k \quad (a_k \in E^1,\ \vec{x}_k \in E^n)$$

is called a *linear combination* of the vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ (their number must be definite but otherwise arbitrary). Thus Theorem 2 shows that *any vector $\vec{v} \in E^n$ can be expressed, in a unique way, as a linear combination of the n basic unit vectors $\vec{e}_k$* (the coefficients a_k being necessarily the *components* of $\vec{v}$).

Note 2. As we have noted, in E^3 the basic unit vectors are often denoted by $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ and the coordinates by x, y, z. Then, by Theorem 2,

$$\vec{v} = (x,\ y,\ z) = x\vec{\imath} + y\vec{\jmath} + z\vec{k},$$

and this representation of $\vec{v}$ is *unique*. Thus the right side sum may be treated as a standard notation for a vector, instead of $(x,\ y,\ z)$. It should, however, be well-noted that this sum represents *an ordered triple*, namely, $(x,\ y,\ z)$.

Note 3. From our definitions and Theorem 1, the n-space E^n has emerged as a set of elements (called "vectors" or "points") for which two operations are defined, namely, addition of vectors and multiplication of a vector by a scalar (real number). There also are many other sets (not necessarily sets of n-tuples) for which two such operations are defined in some manner. Any set with two such operations is called a *real vector space* if these operations obey all laws specified in Theorem 1. E^1 is called its *field of scalars*. Thus E^n is a real vector space under the operations defined above.

Caution: We shall not define any *inequalities* ($<$) for vectors in E^n. Thus, expressions like $\bar{x} < \bar{y}$ will not be used and should be carefully avoided except if $n = 1$, i.e., if the "vectors" under consideration are simply real numbers (elements of E^1).

Despite the two operations defined in E^n, the n-space *is not a field* (except in the case of E^1), mainly because the multiplication of *a vector by a vector* is not defined in E^n. Scalar multiples are not products of two *vectors*, even though some of their properties resemble those of products of real numbers. There also is no such thing as a "neutral element of vector multiplication" (though there is a neutral element of vector *addition*, namely, $\vec{0}$). In the next section we shall define certain products ("inner products") of vectors; but even so, E^n will not become a field, because these products do not satisfy the field axioms in full. Only for E^2 shall we later define a vector multiplication that will satisfy these axioms, and so E^2 will become a field.

Note 4. As we have seen in Theorem 2, sometimes we have to number several *vectors* by affixing appropriate subscripts; e.g., $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ or $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$. In this case, the coordinates of these vectors are denoted by attaching a *second* subscript. For example, the coordinates of $\vec{x}_1$ are $x_{11}, x_{12}, \ldots, x_{1n}$. Similarly, $\vec{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, etc.

Problems on Vectors in E^n

1. Find the expression $2\vec{u} - \vec{v} - 3\vec{w} + 5\vec{x}$, given that

(a) $\vec{u} = (-1, 2, 0, -7)$, $\vec{v} = (0, 0, -1, -2)$, $\vec{w} = (2, 4, -3, -3)$, $\vec{x} = (0, 1, 0, 1)$;

(b) $\vec{u} = (2, 2, 2)$, $\vec{v} = (-3, 4, 1)$, $\vec{w} = \vec{0}$, $\vec{x} = (5, -7, 0)$;

(c) $\vec{u} = 3\vec{\imath} + \vec{\jmath} - 2\vec{k}$, $\vec{v} = -4\vec{\imath} + 2\vec{\jmath} - \vec{k}$, $\vec{w} = 2\vec{\imath} + \vec{\jmath}$, $\vec{x} = -3\vec{\jmath} + 2\vec{k}$;

(d) $\vec{u} = (2, 1, -1, 0)$, $\vec{v} = (0, -5, 6, 6)$, $\vec{w} = (3, -2, 4, 8)$, $\vec{x} = (3, 3, 3, 3)$.

(In part (c), first rewrite the given vectors as *triples*.)

2. Complete the proof of Theorem 1.

3. Prove Corollaries 1 and 2 in two ways:

(a) using definitions only (in terms of coordinates);

(b) using the laws of Theorem 1 (without coordinates) and assuming $\vec{v} - \vec{w} = \vec{v} + (-\vec{w})$ as a *definition*.

4. In Problem 1, parts (a), (b), and (d), express the given vectors as linear combinations of the basic unit vectors, and compute the required expression $2\vec{u} - \vec{v} - 3\vec{w} + 5\vec{x}$ directly in terms of these unit vectors. Moreover, express $\vec{x}$ as a linear combination of $\vec{u}$, $\vec{v}$, $\vec{w}$, if possible.

5. Find (if possible) four scalars a, b, c, and d such that $\vec{y} = a\vec{u} + b\vec{v} + c\vec{w} + d\vec{x}$, where $\vec{u}$, $\vec{v}$, $\vec{w}$, $\vec{x}$ are as in Problem 1(a), if

(a) $\vec{y} = \vec{e}_1$; (b) $\vec{y} = \vec{e}_2$; (c) $\vec{y} = \vec{e}_3$;
(d) $\vec{y} = (-2, 4, 0, 1)$; (e) $\vec{y} = \vec{e}_4$.

6. Do Problem 5 with $\vec{u}$, $\vec{v}$, $\vec{w}$, $\vec{x}$ as in Problem 1(d).

7. Set up and solve for E^3 a problem analogous to Problem 5, working with the three vectors $\vec{u}$, $\vec{v}$, $\vec{x}$ of Problem 1(b). Do the same for $\vec{u}$, $\vec{v}$, $\vec{x}$ of 1(c).

8. A finite set of vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_m$ in E^n is said to be *linearly dependent* if there are scalars $a_1, a_2, \ldots, a_m$, *not all zero*, such that

$$\sum_{k=1}^{m} a_k \vec{v}_k = \vec{0};$$

if no such scalars exist, the vectors are *linearly independent* (this means that

$$\sum_{k=1}^{m} a_k \vec{v}_k$$

cannot vanish unless all a_k are 0). Prove that the following sets of vectors are linearly independent:

(a) the basic unit vectors in E^3;

(b) same for E^n;

(c) the vectors $(1, 2, -3, 4)$, $(2, 3, 0, 0)$ in E^4;

(d) the vectors $(2, 0, 0)$, $(4, -1, 3)$, and $(0, 4, 1)$ in E^3.

Which of the sets of vectors given in Problem 1 are linearly dependent and which are not? (Give a proof!)

§2. Inner Products. Absolute Values. Distances

We shall now define some new operations on vectors in E^n.

Definition 1.

The *inner product* or *dot product* $\vec{u} \cdot \vec{v}$ of two vectors $\vec{u} = (u_1, u_2, \ldots, u_n)$ and $\vec{v} = (v_1, v_2, \ldots, v_n)$ in E^n is defined as follows:

$$\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n = \sum_{k=1}^{n} u_k v_k.$$

Note that the dot product is a *scalar* (real number), not a vector. Therefore, the dot product is sometimes called the *scalar* product of two vectors.[1]

Example.

Let $\vec{u} = (3, 1, -9, 4)$, $\vec{v} = (-1, 3, 1, 0)$. Then

$$\vec{u} \cdot \vec{v} = 3 \cdot (-1) + 1 \cdot 3 + (-9) \cdot 1 + 4 \cdot 0 = -9.$$

Definition 2.

The *absolute value* (or *length*, or *norm*, or *magnitude*, or *modulus*), $|\vec{v}|$, of a vector $\vec{v} = (v_1, v_2, \dots, v_n)$ in E^n is the scalar defined by

$$|\vec{v}| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{k=1}^{n} v_k^2},$$

i.e., it is the *nonnegative* value of the square root of $\sum_{k=1}^{n} v_k^2$.

Example.

Let $\vec{v} = (3, 4, 0) \in E^3$. Then $|\vec{v}| = \sqrt{9 + 16 + 0} = 5$.

Note 1. In E^1, all "vectors" are simply real numbers, and v has only one component, namely, itself. Thus, by this definition, $|v| = \sqrt{v^2}$; the root equals v if $v \geq 0$ and $-v$ if $v < 0$ (since we always take the *nonnegative* value). Thus it equals the absolute value of v *as defined previously*, for real numbers, so *the two definitions agree*.

Note 2. Geometrically (in E^1, E^2 and E^3), $|\vec{v}|$ is *the length of the line segment joining the origin with the point* $\vec{v}$. For example, if $\vec{v} = (x, y) \in E^2$ (see Figure 13) then $|\vec{v}| = \sqrt{x^2 + y^2}$ is exactly that distance from $\vec{0}$ to $\vec{v}$, as is known by elementary geometry. Similarly for E^3, where $|\vec{v}| = \sqrt{x^2 + y^2 + z^2}$.

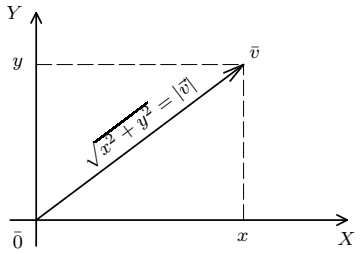


FIGURE 13

Note 3. By Definitions 1 and 2, we have

$$\vec{u} \cdot \vec{u} = \sum_{k=1}^{n} u_k u_k = \sum_{k=1}^{n} u_k^2 = |\vec{u}|^2;$$

hence

$$\sqrt{\vec{u} \cdot \vec{u}} = \sqrt{\sum_{k=1}^{n} u_k^2} = |\vec{u}|.$$

[1] Some authors also use the notation $(\vec{u}, \vec{v})$ or $[\vec{u}, \vec{v}]$ instead of $\vec{u} \cdot \vec{v}$. We shall not use this terminology.

This could serve as a *definition* of the absolute value, $|\vec{u}|$, equivalent to Definition 2. We shall use it below.

Theorem 1. *For any vectors* $\vec{u}$, $\vec{v}$, $\vec{w} \in E^n$ *and scalars* $a, b \in E^1$, *we have*

(a) $\vec{u} \cdot \vec{u} \geq 0$; *and* $\vec{u} \cdot \vec{u} > 0$ *iff* $\vec{u} \neq \vec{0}$;

(b) $(a\vec{u}) \cdot (b\vec{v}) = ab(\vec{u} \cdot \vec{v})$;

(c) $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$ (commutativity of inner products);

(d) $(\vec{u} + \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} + \vec{v} \cdot \vec{w}$ (distributive law).

The proof is immediate from our definitions. (One only has to express $\vec{u}$, $\vec{v}$, and $\vec{w}$ in terms of their coordinates and proceed as in Theorem 1 of §1.) We leave it to the reader. Note that (b) implies that $\vec{u} \cdot \vec{0} = 0$ (put $a = 1$ and $b = 0$), and $a(\vec{u} \cdot \vec{v}) = (a\vec{u}) \cdot \vec{v}$.

Definition 3.

Two vectors $\vec{u}$ and $\vec{v}$ are said to be *parallel* or *collinear* iff one of them is a scalar multiple of the other, i.e.,

$$\vec{u} = t\vec{v} \quad \text{or} \quad \vec{v} = t\vec{u}$$

for some scalar $t \in E^1$. Notation: $\vec{u} \parallel \vec{v}$.

Geometrically (if $\vec{u}$ and $\vec{v}$ are represented as directed line segments), $\vec{u}$ and $\vec{v}$ have *the same direction* (if $t > 0$) or *opposite directions* (if $t < 0$).

Note. $\vec{0} \parallel \vec{u}$ *always* since $\vec{0} = 0\vec{u}$, $(t = 0)$.

Theorem 2. *For any vectors* $\vec{u}$, $\vec{v} \in E^n$ *and any scalar* $a \in E^1$, *we have*

(a′) $|\vec{u}| \geq 0$; *and* $|\vec{u}| = 0$ *iff* $\vec{u} = \vec{0}$;

(b′) $|a\vec{u}| = |a|\,|\vec{u}|$;

(c′) $|\vec{u} \cdot \vec{v}| \leq |\vec{u}|\,|\vec{v}|$ (Cauchy–Schwarz inequality) *and* $|\vec{u} \cdot \vec{v}| = |\vec{u}|\,|\vec{v}|$ *iff* $\vec{u} \parallel \vec{v}$;

(d′) $|\vec{u} + \vec{v}| \leq |\vec{u}| + |\vec{v}|$ *and* $\big||\vec{u}| - |\vec{v}|\big| \leq |\vec{u} - \vec{v}|$ (triangle inequalities).

Proof. Property (a′) follows from Theorem 1(a) since $|\vec{u}|^2 = \vec{u} \cdot \vec{u}$, by Note 3 to Definition 2.

For (b′), we use Theorem 1(b) to obtain

$$(a\vec{u}) \cdot (a\vec{u}) = a^2(\vec{u} \cdot \vec{u}) = a^2|\vec{u}|^2 \quad (\text{since } \vec{u} \cdot \vec{u} = |\vec{u}|^2).$$

Also, $(a\vec{u}) \cdot (a\vec{u}) = |a\vec{u}|^2$. Hence $|a\vec{u}|^2 = a^2|\vec{u}|^2$, and (b′) follows.

(c′) If $\vec{u} \parallel \vec{v}$, then $\vec{u} = t\vec{v}$ or $\vec{v} = t\vec{u}$ (Definition 3); say, $\vec{u} = t\vec{v}$. Then, by (b′) and Theorem 1(b),

$$|\vec{u} \cdot \vec{v}| = |t\vec{v} \cdot \vec{v}| = |t|(\vec{v} \cdot \vec{v}) = |t|\,|\vec{v}|^2 = |t|\,|\vec{v}|\,|\vec{v}| = |t\vec{v}|\,|\vec{v}| = |\vec{u}|\,|\vec{v}|.$$

Thus, $\vec{u} \parallel \vec{v}$ implies the *equality* $|\vec{u} \cdot \vec{v}| = |\vec{u}|\,|\vec{v}|$.

Now suppose $\vec{u}$ and $\vec{v}$ are *not* parallel. Then $\vec{v} = t\vec{u}$ for *no* $t \in E^1$. Hence $(\forall t \in E^1)\ |t\vec{u} - \vec{v}|^2 \neq 0$. But, by Definition 2,

$$|t\vec{u} - \vec{v}|^2 = \sum_{k=1}^{n} (tu_k - v_k)^2.$$

Thus,

$$0 \neq |t\vec{u} - \vec{v}|^2 = \sum_{k=1}^{n} (tu_k - v_k)^2 = t^2 \sum_{k=1}^{n} u_k^2 - 2t \sum_{k=1}^{n} u_k v_k + \sum_{k=1}^{n} v_k^2, \quad (t \in E^1).$$

Setting, for brevity,

$$A = \sum_{k=1}^{n} u_k^2,\ B = 2\sum_{k=1}^{n} u_k v_k,\ \text{and } C = \sum_{k=1}^{n} v_k^2,$$

we see that the quadratic equation $0 = At^2 - Bt + C$ has *no* real solutions for t. Thus, by elementary algebra, its discriminant $B^2 - 4AC$ must be *negative*. Substituting the values of A, B, C in $B^2 - 4AC < 0$ and dividing by 4, we get

$$\left(\sum_{k=1}^{n} u_k v_k\right)^2 < \left(\sum_{k=1}^{n} u_k^2\right)\left(\sum_{k=1}^{n} v_k^2\right).$$

By Definitions 1 and 2, this means that $|\vec{u} \cdot \vec{v}|^2 < |\vec{u}|^2 |\vec{v}|^2$, or $|\vec{u} \cdot \vec{v}| < |\vec{u}|\,|\vec{v}|$.

We have shown that $|\vec{u} \cdot \vec{v}| = |\vec{u}|\,|\vec{v}|$ or $|\vec{u} \cdot \vec{v}| < |\vec{u}|\,|\vec{v}|$, *according to whether* $\vec{u}$ *is or is not parallel to* $\vec{v}$. Thus assertion (c′) is proved.

(d′) Expand $|\vec{u} + \vec{v}|^2$ using Theorem 1(d) and Note 3 to get

$$|\vec{u} + \vec{v}|^2 = (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) = \vec{u} \cdot \vec{u} + 2\vec{u} \cdot \vec{v} + \vec{v} \cdot \vec{v} = |\vec{u}|^2 + 2\vec{u} \cdot \vec{v} + |\vec{v}|^2.$$

As $\vec{u} \cdot \vec{v} \leq |\vec{u} \cdot \vec{v}| \leq |\vec{u}|\,|\vec{v}|$ (by (c′)), this yields

$$|\vec{u} + \vec{v}|^2 \leq |\vec{u}|^2 + 2|\vec{u}|\,|\vec{v}| + |\vec{v}|^2 = (|\vec{u}| + |\vec{v}|)^2,$$

proving the first formula in (d′). The second formula follows from it exactly as in Chapter 2, §4, Corollary 6. Thus all is proved. □

Note 4. In E^2 and E^3, the triangle inequalities have a simple geometric interpretation. Represent the vectors $\vec{u}$ and $\vec{v}$ as (directed) sides in a triangle. Then $\vec{u} + \vec{v}$ represents geometrically the third side (see Figure 14). The absolute values $|\vec{u}|$, $|\vec{v}|$, and $|\vec{u}+\vec{v}|$ are the lengths of the sides. Thus the first formula (d′) states that a side of a triangle never exceeds the sum of the other two sides, while the second formula (d′) says that the difference of two sides never exceeds the third side. (This explains the name "triangle inequalities".) If $\vec{u} \parallel \vec{v}$, the triangle "collapses", and inequalities become *equalities* (see Problem 7).

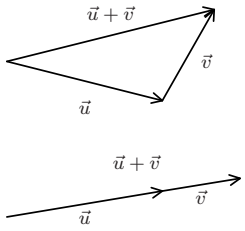


Figure 14

From elementary geometry in E^2 and E^3, the reader is certainly familiar with the formulas for the distance between two points $\bar{u}$ and $\bar{v}$, in terms of their coordinates. Denoting this distance by $\rho(\bar{u}, \bar{v})$, we have in E^2

$$\rho(\bar{u}, \bar{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2}\,;$$

and in E^3

$$\rho(\bar{u}, \bar{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2}\,.$$

Note that the *differences* $u_k - v_k$ *are the coordinates of* $\bar{u} - \bar{v}$. Hence, by Definition 2, the *square roots given above equal exactly the absolute value of the vector* $\bar{u} - \bar{v}$, so that $\rho(\bar{u}, \bar{v}) = |\bar{u} - \bar{v}|$, in both E^2 and E^3. It is natural to define distances in E^n in a similar manner, as we shall do now.

Definition 4.

The *distance* $\rho(\bar{u}, \bar{v})$ between two points $\bar{u} = (u_1, \ldots, u_n)$ and $\bar{v} = (v_1, \ldots, v_n)$ in E^n is the scalar defined by

$$\rho(\bar{u}, \bar{v}) = |\bar{u} - \bar{v}| = \sqrt{\sum_{k=1}^{n} (u_k - v_k)^2} = \sqrt{(\bar{u} - \bar{v}) \cdot (\bar{u} - \bar{v})}\,.$$

Note 5. When speaking of *distances*, we shall use the term "*point*" rather than "*vector*", and the notation $\bar{u}$ rather than $\vec{u}$. As previously noted, we call $\bar{u} - \bar{v} = \overrightarrow{vu}$ the "vector passing from the point $\bar{v}$ to the point $\bar{u}$" or, briefly, "the vector from $\bar{v}$ to $\bar{u}$" (in this order), as is suggested by Figure 15. With this terminology and notation, we have

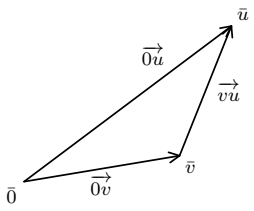


Figure 15

$$\rho(\bar{u}, \bar{v}) = |\overrightarrow{vu}| = |\bar{u} - \bar{v}|;$$

i.e., *the distance* $\rho(\bar{u}, \bar{v})$ *is the length of the vector from* $\bar{v}$ *to* $\bar{u}$.

Theorem 3. *For any points* $\bar{u}$, $\bar{v}$, $\bar{w} \in E^n$, *we have*

(i) $\rho(\bar{u}, \bar{v}) \geq 0$; *and* $\rho(\bar{u}, \bar{v}) = 0$ *iff* $\bar{u} = \bar{v}$;

(ii) $\rho(\bar{u}, \bar{v}) = \rho(\bar{v}, \bar{u})$ (symmetry law);

(iii) $\rho(\bar{u}, \bar{w}) \leq \rho(\bar{u}, \bar{v}) + \rho(\bar{v}, \bar{w})$ (triangle inequality).

Proof. (i) Since $\rho(\bar{u},\bar{v}) = |\bar{u}-\bar{v}|$, we have, by Theorem 2(a′),

$$\rho(\bar{u},\bar{v}) = |\bar{u}-\bar{v}| \geq 0.$$

Also, $|\bar{u}-\bar{v}| \neq 0$ iff $\bar{u}-\bar{v} \neq 0$, i.e., iff $\bar{u} \neq \bar{v}$. Hence $\rho(\bar{u},\bar{v}) \neq 0$ iff $u \neq v$; i.e., $\rho(\bar{u},\bar{v}) = 0$ iff $\bar{u} = \bar{v}$, as asserted.

(ii) By Theorem 2(b′),

$$|\bar{u}-\bar{v}| = |(-1)(\bar{u}-\bar{v})| = |\bar{v}-\bar{u}|.$$

As $|\bar{u}-\bar{v}| = \rho(\bar{u},\bar{v})$, this means that $\rho(\bar{u},\bar{v}) = \rho(\bar{v},\bar{u})$, as required.

(iii) By definition,

$$\rho(\bar{u},\bar{v}) + \rho(\bar{v},\bar{w}) = |\bar{u}-\bar{v}| + |\bar{v}-\bar{w}|;$$

and by the triangle inequality for absolute values,

$$|\bar{u}-\bar{v}| + |\bar{v}-\bar{w}| \geq |\bar{u}-\bar{w}| = \rho(\bar{u},\bar{w}).$$

Hence $\rho(\bar{u},\bar{v}) + \rho(\bar{v},\bar{w}) \geq \rho(\bar{u},\bar{w})$, and all is proved. □

Note 6. We also have $|\rho(\bar{u},\bar{v}) - \rho(\bar{w},\bar{v})| \leq \rho(\bar{u},\bar{w})$. The proof is left to the reader as an exercise.

Problems on Vectors in E^n (continued)

1. Complete the proofs of Theorems 1 and 2 (last part) and Note 6.

2. Prove Theorem 2(a′)(b′) from our definitions, without using Theorem 1.

3. Given the vectors (points) $\vec{u}$, $\vec{v}$, $\vec{w}$, $\vec{x}$ as in Problem 1 of §1, compute their absolute values, mutual distances and dot products. (Treat the cases (a), (b), (c), and (d) *separately*.) Take any three of these vectors and verify by direct computation that they satisfy the formulas of Theorems 1 and 2. Are any two of these vectors parallel?

4. Slightly modify the proof of Theorem 2(c′) to obtain the *stronger* result

$$\left(\sum_{k=1}^{n} |u_k v_k|\right)^2 \leq \left(\sum_{k=1}^{n} u_k^2\right)\left(\sum_{k=1}^{n} v_k^2\right).$$

Why is this *stronger* than the ordinary Cauchy–Schwarz inequality?

5. Give another proof of the Cauchy–Schwarz inequality, $|\vec{u}\cdot\vec{v}| \leq |\vec{u}|\,|\vec{v}|$.

[Outline: If $|\vec{u}| = 0$ or $|\vec{v}| = 0$, this reduces to the trivial $0 \leq 0$. Thus assume $|\vec{u}| > 0$, $|\vec{v}| > 0$, and set $a = |\vec{v}|/|\vec{u}|$; so $a > 0$ and $a|\vec{u}| = |\vec{v}|$. Deduce that

$$a^2(\vec{u}\cdot\vec{u}) = a^2|\vec{u}|^2 = a|\vec{u}|\,|\vec{v}| = |\vec{v}|^2 = \vec{v}\cdot\vec{v}. \tag{i}$$

Now consider $(a\vec{u}\pm\vec{v})\cdot(a\vec{u}\pm\vec{v}) \geq 0$ (Theorem 1(a)). By Theorem 1(d)(b), expanding in the usual way, obtain

$$0 \leq (a\vec{u}\pm v)\cdot(a\vec{u}\pm\vec{v}) = a^2\vec{u}\cdot\vec{u} + \vec{v}\cdot\vec{v} \pm 2a\vec{u}\cdot\vec{v}.$$

Hence, by step (i),

$$0 \leq a|\vec{u}|\,|\vec{v}| + a|\vec{u}|\,|\vec{v}| \pm 2a(\vec{u}\cdot\vec{v}) = 2a|\vec{u}|\,|\vec{v}| \pm 2a(\vec{u}\cdot\vec{v});$$

or, transposing, $\pm 2a(\vec{u}\cdot\vec{v}) \leq 2a|\vec{u}|\,|\vec{v}|$. Divide by $2a$ to obtain the result.]

6. If $\vec{v} \neq \vec{0}$, prove that $\vec{u} \parallel \vec{v}$ iff

$$\frac{u_1}{v_1} = \frac{u_2}{v_2} = \cdots = \frac{u_n}{v_n} = t,$$

for some $t \in E^1$, where "$u_k/v_k = t$" is to be replaced by "$u_k = 0$" if $v_k = 0$.

7. Prove that

(i) $|\vec{u}+\vec{v}| = |\vec{u}| + |\vec{v}|$ iff $\vec{u} = t\vec{v}$ or $\vec{v} = t\vec{u}$ for some $t \geq 0$;

(ii) $|\vec{u}-\vec{v}| = |\vec{u}| + |\vec{v}|$ iff $\vec{u} = t\vec{v}$ or $\vec{v} = t\vec{u}$ for some $t \leq 0$.

[Hint: For the "only if", proceed as in the proof of Theorem 2(d′), using the "equality" part of Theorem 2(c′).]

***8.** Use induction on n to prove the *Lagrange identity* (valid in any field):

$$\left(\sum_{k=1}^{n} u_k^2\right)\left(\sum_{k=1}^{n} v_k^2\right) - \left(\sum_{k=1}^{n} u_k v_k\right)^2 = \sum_{1\leq i\leq k\leq n} (u_i v_k - u_k v_i)^2,$$

where the right-hand sum contains all terms for which $1 \leq i \leq k \leq n$ (only).

***9.** Using the results of Problems 6 and 8, find a new proof of Theorem 2(c′).

§3. Angles and Directions

The inner product $\vec{u}\cdot\vec{v}$ of two vectors, as defined in §2, has a simple geometric interpretation (in E^2 and E^3), when the vectors are represented as directed line segments: *it equals the product of the lengths of $\vec{u}$ and $\vec{v}$ multiplied by the cosine of the angle between $\vec{u}$ and $\vec{v}$*,

$$\vec{u}\cdot\vec{v} = |\vec{u}|\,|\vec{v}|\cos\langle\vec{u},\vec{v}\rangle,$$

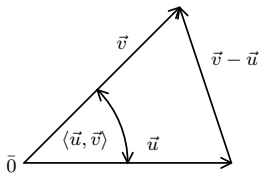


FIGURE 16

where $\langle\vec{u},\vec{v}\rangle$ denotes that angle. Indeed (see Figure 16), by the law of cosines,

$$|\vec{u}|^2 + |\vec{v}|^2 - 2|\vec{u}|\,|\vec{v}|\cos\langle\vec{u},\vec{v}\rangle = |\vec{v}-\vec{u}|^2.$$

As $|\vec{u}|^2 = \vec{u}\cdot\vec{u}$, $|\vec{v}|^2 = \vec{v}\cdot\vec{v}$, etc., we obtain

$$\begin{aligned}\vec{u}\cdot\vec{u} + \vec{v}\cdot\vec{v} - 2|\vec{u}|\,|\vec{v}|\cos\langle\vec{u},\vec{v}\rangle &= |\vec{v}-\vec{u}|^2\\ &= (\vec{v}-\vec{u})\cdot(\vec{v}-\vec{u})\\ &= \vec{v}\cdot\vec{v} + \vec{u}\cdot\vec{u} - 2\vec{u}\cdot\vec{v},\end{aligned}$$

by the distributive law. Cancelling and reducing, we get

$$\vec{u}\cdot\vec{v} = |\vec{u}|\,|\vec{v}|\cos\langle\vec{u},\vec{v}\rangle,$$

as asserted. If $\vec{u}\neq\vec{0}$ and $\vec{v}\neq\vec{0}$, we also obtain

$$\cos\langle\vec{u},\vec{v}\rangle = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\,|\vec{v}|}.$$

It is natural to accept this as a *definition* of an angle in E^n as well.

Definition 1.

Given two vectors $\vec{u}\neq\vec{0}$ and $\vec{v}\neq\vec{0}$ in E^n, we define the (undirected) *angle* between them, denoted $\langle\vec{u},\vec{v}\rangle$, as the main value of

$$\arccos\frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\,|\vec{v}|},$$

i.e., the (unique) number between 0 and π such that

$$\cos\langle\vec{u},\vec{v}\rangle = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\,|\vec{v}|} \quad (\vec{u}\neq\vec{0},\ \vec{v}\neq\vec{0}). \tag{1}$$

Note 1. Throughout this and some other sections, we assume the notions and laws of elementary trigonometry to be known. Actually, however, what will be needed are only the *cosines* of the angles, and we may treat formula (1) as a *definition*, even without speaking of the "angle" itself. It is only for the sake of geometric interpretation that we speak of "angles", "cosines", "perpendicularity", etc., and sometimes express "angles" in degrees instead of radians.

Note 2. By the Cauchy–Schwarz inequality, we always have $|\vec{u}\cdot\vec{v}| \le |\vec{u}|\,|\vec{v}|$. Hence the fraction $(\vec{u}\cdot\vec{v})/(|\vec{u}|\,|\vec{v}|)$ in formula (1) never exceeds 1 in absolute value, so that an angle with $\cos\langle\vec{u},\vec{v}\rangle = (\vec{u}\cdot\vec{v})/(|\vec{u}|\,|\vec{v}|)$ does exist. However, *it is not defined if $\vec{u}=\vec{0}$ or $\vec{v}=\vec{0}$.*

Definition 2.

Two vectors $\vec{u}$ and $\vec{v}$ in E^n are said to be *orthogonal* or *perpendicular* if $\vec{u}\cdot\vec{v}=0$; or, in terms of coordinates,

$$\sum_{k=1}^{n} u_k v_k = 0.$$

We then write $\vec{u}\perp\vec{v}$.

This notion is defined also if $\vec{u}=\vec{0}$ or $\vec{v}=\vec{0}$. In particular, $\vec{0}\perp\vec{v}$ for *every* $\vec{v}\in E^n$; and $\vec{e}_k\perp\vec{e}_i$ $(k\neq i)$ for the basic unit vectors. (Verify!) If, however, $\vec{u}\neq\vec{0}$ and $\vec{v}\neq\vec{0}$, then $\vec{u}\perp\vec{v}$ also means that

$$\cos\langle\vec{u},\vec{v}\rangle = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\,|\vec{v}|} = 0, \text{ i.e., } \langle\vec{u},\vec{v}\rangle = \frac{\pi}{2}.$$

Of special importance are the n angles which a given vector $\vec{v}\neq\vec{0}$ forms with the basic unit vectors $\vec{e}_1,\dots,\vec{e}_n$, i.e., the angles $\langle\vec{v},\vec{e}_k\rangle$, $k=1,\dots,n$. They are called the *direction angles* of $\vec{v}$, and their cosines are called the *direction cosines* of $\vec{v}$. Thus every vector $\vec{v}\neq\vec{0}$ in E^n has exactly n direction cosines. Geometrically (in E^2 and E^3), the direction angles are those between $\vec{v}$ and the positive directions of the coordinate axes $(\vec{\imath},\vec{\jmath},\vec{k})$. We now obtain the following result.

Corollary 1. *For any vector $\vec{v}=(v_1,\dots,v_n)\neq\vec{0}$ in E^n, the following is true*:

(a) *We have*

$$\cos\langle\vec{v},\vec{e}_k\rangle = \frac{v_k}{|\vec{v}|}, \quad k=1,\dots,n;$$

i.e., the direction cosines of $\vec{v}$ are obtained by dividing its coordinates v_k by the length $|\vec{v}|$ of $\vec{v}$.

(b) *The sum of the squares of the direction cosines of $\vec{v}$ always equals* 1:

$$\sum_{k=1}^{n}\cos^2\langle\vec{v},\vec{e}_k\rangle = 1. \tag{2}$$

Proof. By definition, all coordinates of $\vec{e}_k$ are 0 except the k-th, which is 1. Thus, computing the length of $\vec{e}_k$, we obtain $|\vec{e}_k|=1$. Similarly, the dot product $\vec{v}\cdot\vec{e}_k$ equals v_k (the k-th coordinates of $\vec{v}$) because, by definition, it is a sum in which all terms but one, $v_k\times 1$, are equal to 0. Substituting this in formula (1), we have

$$\cos\langle\vec{v},\vec{e}_k\rangle = \frac{\vec{v}\cdot\vec{e}_k}{|\vec{v}|\,|\vec{e}_k|} = \frac{v_k}{|\vec{v}|},$$

proving assertion (a).

Part (b) is obtained by substituting this in (2) and noting that $\sum_{k=1}^{n} v_k^2 = |\vec{v}|^2$; we leave the details to the reader. □

Note 3. In E^3, the direction angles of $\vec{v}$ are often denoted by α, β, γ. Then formula (2) simplifies to

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1.$$

Definition 3.

By a *unit vector* or a *direction* in E^n is meant any vector of length $|\vec{v}| = 1$. Such are, e.g., the n basic unit vectors $\vec{e}_k$ (see above).

By dividing any vector $\vec{v} \neq \vec{0}$ by its own magnitude $|\vec{v}| \neq 0$, *we always obtain a unit vector* (called the *unit* of $\vec{v}$, or the *direction* of $\vec{v}$, or the *normalized vector* of $\vec{v}$). Indeed, the resulting $\vec{u} = \vec{v}/|\vec{v}|$ has length 1 since, by Theorem 2(b′) of §2,

$$\left|\frac{1}{|\vec{v}|}\vec{v}\right| = \frac{1}{|\vec{v}|}|\vec{v}| = 1.$$

To *normalize* a vector $\vec{v} \neq \vec{0}$ means to divide it by its own magnitude $|\vec{v}|$, i.e., to multiply by $1/|\vec{v}|$. Of course, this is only possible if $\vec{v} \neq \vec{0}$.

We also obtain the following result.

Corollary 2. *The direction cosines of any vector $\vec{v} \neq \vec{0}$ in E^n are equal to the corresponding components of its unit $\vec{v}/|\vec{v}|$. Hence, if $|\vec{v}| = 1$, these cosines are simply the components of $\vec{v}$. (It also follows that the components of a unit vector never exceed 1 in absolute value.)*

Indeed, the coordinates of $\vec{v}/|\vec{v}|$, by definition, are obtained by dividing those of $\vec{v}$ by the scalar $|\vec{v}|$. But, by Corollary 1(a), so also are obtained the direction cosines of $\vec{v}$. Thus our assertion follows.

Examples.

Take two vectors in E^4: $\vec{u} = (1, -2, 0, -1)$ and $\vec{v} = (0, 3, 2, -2)$. Then

$$|\vec{u}| = \sqrt{1^2 + (-2)^2 + 0^2 + (-1)^2} = \sqrt{6};$$

similarly $|\vec{v}| = \sqrt{17}$. Since $\vec{u} \neq \vec{0}$ and $\vec{v} \neq \vec{0}$, the angle $\langle \vec{u}, \vec{v} \rangle$ exists and, by definition,

$$\cos\langle \vec{u}, \vec{v} \rangle = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}|\,|\vec{v}|} = \frac{-3}{\sqrt{6 \cdot 17}} = \frac{-3}{\sqrt{102}}.$$

To obtain the direction cosines of $\vec{u}$, we normalize it:

$$\frac{\vec{u}}{|\vec{u}|} = \frac{(1, -2, 0, -1)}{\sqrt{6}} = \left(\frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, 0, \frac{-1}{\sqrt{6}}\right).$$

These four numbers are the required cosines, by Corollary 2.

We leave to the reader the proof of the following proposition.

Corollary 3. *The direction cosines of a vector $\vec{v} \neq \vec{0}$ in E^n do not change if $\vec{v}$ is multiplied by a scalar $a > 0$; they change sign only if $a < 0$. Hence the direction cosines of $-\vec{v}$ are those of $\vec{v}$ with opposite signs.*

Note 4. The notions of *angle* and *unit vector* were defined by using inner products and absolute values. Thus one can define them, in exactly the same manner, not only in E^n but also in other vector spaces (see Note 3, §1) *in which inner products (satisfying Theorem 1 of §2) are defined*. Such vector spaces are called *Euclidean*. For more details, see §9.

§4. Lines and Line Segments

The term "*line*" shall always mean a line *extending indefinitely* (never a *line segment*, which is only a part of a line).

To obtain all points of a straight line in E^2 or E^3, we take a "vector" $\vec{u} = \overrightarrow{ab}$ (joining two given points $\bar{a}$ and $\bar{b}$ on the line) and then, so to say, "stretch" it indefinitely in both directions, i.e., multiply $\vec{u}$ by all possible scalars $t \in E^1$ (positive, negative, and 0). Now, by definition,

$$\vec{u} = \overrightarrow{ab} = \bar{b} - \bar{a} = \overrightarrow{0b} - \overrightarrow{0a}$$

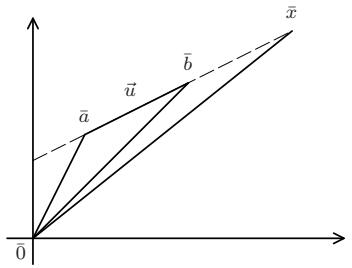


Figure 17

(see Figure 17). The "position vector" $\overrightarrow{0x}$ of any point $\bar{x}$ on the line $\overline{ab}$ is, geometrically, the sum of $\overrightarrow{0a}$ and $\overrightarrow{ax}$: $\overrightarrow{0x} = \overrightarrow{0a} + \overrightarrow{ax}$. Here the vector $\overrightarrow{ax}$ is *a scalar multiple of* $\overrightarrow{ab} = \vec{u}$; specifically, $\overrightarrow{ax} = t\vec{u}$, where

$$t = \frac{|\overrightarrow{ax}|}{|\vec{u}|} \quad \text{or} \quad t = -\frac{|\overrightarrow{ax}|}{|\vec{u}|}$$

according to whether the vectors $\overrightarrow{ax}$ and $\vec{u}$ have the same or opposite directions. Thus we have

$$\bar{x} = \overrightarrow{0x} = \overrightarrow{0a} + \overrightarrow{ax} = \bar{a} + t\vec{u}.$$

Conversely, every point of that form (for any $t \in E^1$) lies on the line $\overline{ab}$. Thus *the line $\overline{ab}$ in E^2 or E^3 is exactly the set of all points $\bar{x}$ of the form*

$$\bar{x} = \bar{a} + t\vec{u} = \bar{a} + t(\bar{b} - \bar{a}), \quad t \in E^1.$$

By varying t, we obtain all points of $\overline{ab}$.

It is natural to accept this as a *definition* of a line in E^n.

Definition 1.

The *line* passing through two given points $\bar{a}$, $\bar{b} \in E^n$ ($\bar{a} \neq \bar{b}$) (equivalently, the line passing through $\bar{a}$ in the direction of a vector $\vec{u} = \overrightarrow{ab} = \bar{b} - \bar{a}$) is

the set of all points $\bar{x} \in E^n$ of the form

$$\bar{x} = \bar{a} + t\vec{u} = \bar{a} + t(\bar{b} - \bar{a}),$$

where t is a variable which takes on all real values (we call it a *real parameter*). In symbols,

$$\text{Line } \overline{ab} = \{\bar{x} \in E^n \mid \bar{x} = \bar{a} + t\vec{u} \text{ for some } t \in E^1\}; \quad \vec{u} = \bar{b} - \bar{a} = \overrightarrow{ab} \neq \vec{0}. \quad (1)$$

Briefly, we call it "the line $\bar{x} = \bar{a} + t\vec{u}$" or "the line $\bar{x} = \bar{a} + t(\bar{b} - \bar{a})$"; instead, we may write $\bar{x} = (1 - t)\bar{a} + t\bar{b}$ (rearranging brackets). The formula $\bar{x} = \bar{a} + t\vec{u}$ (respectively, $\bar{x} = \bar{a} + t(\bar{b} - \bar{a})$) is called *the equation* of the line (more precisely, its *parametric* equation). In the first case, we say that the line is given by a point $\bar{a}$ and a direction $\vec{u}$; in the second case, it is determined by two of its points, $\bar{a}$ and $\bar{b}$. In terms of the *coordinates* of $\bar{x}$, $\bar{a}$ and $\vec{u}$ (or $\bar{b}$), the parametric equation is equivalent to n simultaneous equations (called the *parametric coordinate equations* of the line):

$$x_k = a_k + tu_k = a_k + t(b_k - a_k), \quad k = 1, 2, \ldots, n. \quad (2)$$

It is a great advantage of the vector notation that *one* vector equation replaces n coordinate equations.

Now, since the vector $\vec{u}$ (used to form the line) is anyway being multiplied by arbitrary scalars t, it is clear that the line (1) will not gain or lose any of its points if $\vec{u}$ is replaced by some scalar multiple $c\vec{u}$ ($c \neq 0$). *In particular, we may replace $\vec{u}$ by its unit $\vec{u}/|\vec{u}|$* (taking $c = 1/|\vec{u}|$). Thus we may always assume (if desirable) that $\vec{u}$ is a unit vector itself. In this case the equation $\bar{x} = \bar{a} + t\vec{u}$ (and the equations (2)) are said to be *normal*. To *normalize* an equation of a line means to replace $\vec{u}$ by $\vec{u}/|\vec{u}|$. Since c may also be negative, the line (1) does not change if we replace $\vec{u}$ by $-\vec{u}$; thus the direction of a line is not uniquely determined: we always have two choices of the unit vector $\vec{u}$. If, however, a particular $\vec{u}$ is *prescribed* in advance, we speak of a *directed* line. The coordinates of the direction vector $\vec{u}$ (or any of its scalar multiples $c\vec{u}$) are called a *set of n direction numbers* for the line (1); of course, there are infinitely many such sets corresponding to different values of c. In particular, the *direction cosines* of $\vec{u}$ (i.e., the components of the unit vector $\vec{u}/|\vec{u}|$) are called a set of *direction cosines of the line*. (There are precisely two such sets, namely the direction cosines of $\vec{u}$ and those of $-\vec{u}$.)

In addition to changing the vector $\vec{u}$, we may also alter the parameter t. Indeed, since t is anyway supposed to take on all real values, nothing will change if we replace it by some other variable expression θ *which likewise runs over all real values*, e.g., by $\theta = 1 - t$. Thus, every line has infinitely many parametric equations, depending on the choice of the parameter. We can also entirely eliminate the parameter from equations (2) by rewriting them as follows

(assuming that $b_k - a_k \neq 0$), and then dropping "t" on the right, if desirable:

$$\frac{x_1 - a_1}{b_1 - a_1} = \frac{x_2 - a_2}{b_2 - a_2} = \cdots = \frac{x_n - a_n}{b_n - a_n} = t. \quad (3)$$

One can write the equations in that form even if some of the denominators vanish. It is then understood that the corresponding numerators are to be equated to 0, e.g., $x_k - a_k = 0$, and this equation replaces the (senseless) equation involving the fraction with the vanishing denominator. Note that the x_k in (3) and (2) are *variables*.

Dropping t in (3), we are left with $n - 1$ equations between n fractions involving only the (fixed) coordinates of $\bar{a}$ and $\bar{b}$ and the (variable) coordinates of $\bar{x}$. A point $\bar{x}$ then belongs to the line $\overline{ab}$ if and only if its coordinates satisfy these $n - 1$ equations (called the *nonparametric* equations of a line through two given points). If, instead, the line is given in terms of one point $\bar{a}$ and a direction vector $\vec{u} = \bar{b} - \bar{a}$, then, replacing $b_k - a_k$ by u_k, we get

$$\frac{x_1 - a_1}{u_1} = \frac{x_2 - a_2}{u_2} = \cdots = \frac{x_n - a_n}{u_k}. \quad (4)$$

Here the u_k form a set of direction numbers. Normalizing (i.e., dividing the u_k by $|\vec{u}| = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}$), we get a set of *direction cosines* of $\overline{ab}$.

If $\vec{u}$ and $\vec{v}$ are the direction vectors of two lines, we also call $\langle \vec{u}, \vec{v} \rangle$ (as defined in §3) the *angle between the two lines*. This angle is uniquely determined if the lines are directed; otherwise, by changing the sign of $\vec{u}$ or $\vec{v}$, one can also change the sign of $\cos\langle \vec{u}, \vec{v} \rangle$. (Verify this!) Thus one obtains *two* angles, α and $\pi - \alpha$. Two lines are said to be *perpendicular* or *orthogonal* if $\vec{u} \perp \vec{v}$, i.e., if

$$\vec{u} \cdot \vec{v} = \sum_{k=1}^{n} u_k v_k = 0.$$

They are said to be *parallel* if one of $\vec{u}$ and $\vec{v}$ is a scalar multiple of the other, say $\vec{u} = c\vec{v}$; in this case, we also say that the vectors $\vec{u}$ and $\vec{v}$ are *collinear* (see Definition 3 in §2).

Note 1. More precisely, we say that $\vec{u}$ and $\vec{v}$ are *vector-collinear* to mean that $\vec{u} = c\vec{v}$ or $\vec{v} = c\vec{u}$. On the other hand, it is customary to say that three *points* $\bar{a}$, $\bar{b}$, $\bar{c}$ are collinear iff they lie on *one and the same* line (a different notion!).

If, in the parametric equation $\bar{x} = \bar{a} + t\vec{u}$, or

$$\bar{x} = \bar{a} + t(\bar{b} - \bar{a}) = (1 - t)\bar{a} + t\bar{b},$$

we let t vary not over all of E^1 but only over some subset of E^1, then we obtain only a part of the line $\overline{ab}$. In particular, by letting t vary over some *interval* in E^1, we obtain what is called a *line segment* in E^n. (We reserve the name "*interval*" for another kind of sets, to be defined in §7. In E^1, both kinds of

sets coincide with ordinary intervals.) Exactly as in E^1, we have four types of such line segments. We define them below.

Definition 2.

Given two points $\bar{a}$ and $\bar{b}$ in E^n, we define the *open line segment from $\bar{a}$ to $\bar{b}$*, denoted $L(\bar{a},\bar{b})$, as the set of all points $\bar{x}\in E^n$ of the form

$$\bar{x}=\bar{a}+t(\bar{b}-\bar{a})=(1-t)\bar{a}+t\bar{b},$$

where t varies over the interval $(0,1)\subset E^1$, i.e., $0<t<1$. In symbols,

$$L(a,b)=\{\bar{x}\in E^n \mid \bar{x}=\bar{a}+t(\bar{b}-\bar{a}) \text{ for some } t\in(0,1)\}.$$

This is also briefly written as $L(a,b)=\{\bar{a}+t(\bar{b}-\bar{a}) \mid 0<t<1\}$, i.e., "the set of all points $\bar{a}+t(\bar{b}-\bar{a})$ for $0<t<1$."

Similarly, the *closed line segment* $L[\bar{a},\bar{b}]$ is

$$L[\bar{a},\bar{b}]=\{\bar{a}+t(\bar{b}-\bar{a}) \mid 0\le t\le 1\};$$

the *half-open* line segment is

$$L(\bar{a},\bar{b}]=\{\bar{a}+t(\bar{b}-\bar{a}) \mid 0<t\le 1\},$$

and the *half-closed* one is

$$L[\bar{a},\bar{b})=\{\bar{a}+t(\bar{b}-\bar{a}) \mid 0\le t<1\}.$$

In all cases, $\bar{a}$ and $\bar{b}$ are called the *endpoints* of the line segment, and $|\bar{b}-\bar{a}|$ is called its *length*.

Note 2. (i) The line segments are also defined in case $\bar{a}=\bar{b}$ ("degenerate case"). (ii) Setting $t=0$ or $t=1$, we obtain the endpoints $\bar{a}$ and $\bar{b}$, respectively. The other points are obtained as t varies between 0 and 1.

Examples.

Take three points in E^3: $\bar{a}=(0,-1,2)$, $\bar{b}=(1,1,1)$, $\bar{c}=(3,1,-1)$. Then the line $\overline{ab}$ has the parametric equation $\bar{x}=\bar{a}+t(\bar{b}-\bar{a})$; or, in coordinates,

$$x_1=0+t(1-0)=t,\ x_1=-1+2t,\ x_3=2-t;$$

or, writing (x,y,z) for (x_1,x_2,x_3),

$$x=t,\ y=-1+2t,\ z=2-t.$$

Eliminating t (as in formula (3)), we obtain

$$\frac{x}{1}=\frac{y+1}{2}=\frac{z-2}{-1};\ \text{or, normalizing, } \frac{x}{1/\sqrt{6}}=\frac{y+1}{2/\sqrt{6}}=\frac{z-2}{-1/\sqrt{6}},$$

where $(1,2,-1)$ is a set of *direction numbers* (coordinates of the vector $\vec{u}=\overrightarrow{ab}=\bar{b}-\bar{a}$), while

$$\Big(\frac{1}{\sqrt{6}},\frac{2}{\sqrt{6}},\frac{-1}{\sqrt{6}}\Big)$$

is a set of *direction cosines* (coordinates of the unit vector $\vec{u}/|\vec{u}|$). A set of direction numbers for the line $\overline{bc}$ is obtained from the vector $\vec{v}=\overrightarrow{bc}=\bar{c}-\bar{b}=(2,0,-2)$; the direction cosines are $(2/\sqrt{8},0,-2/\sqrt{8})$. Using formula (4), we obtain the coordinate equations in the symbolic form (not normalized)

$$\frac{x-1}{2}=\frac{y-1}{0}=\frac{z-1}{-2};\ \text{i.e., } \frac{x-1}{2}=\frac{z-1}{-2}\ \text{and } y-1=0.$$

The angle between $\overrightarrow{ab}$ and $\overrightarrow{bc}$ is given by

$$\cos\langle\vec{u},\vec{v}\rangle=\frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\,|\vec{v}|}=\frac{4}{\sqrt{48}}=\frac{1}{\sqrt{3}}.$$

***Note 3.** *Any line*, $\bar{x}=\bar{a}+t\vec{u}$, in E^n is an *isomorphic copy* of E^1, in the sense of §14 of Chapter 2. Indeed, let us define a mapping f on E^1 by setting $(\forall t\in E^1)\ f(t)=\bar{a}+t\vec{u}$ (with $\bar{a}$ and $\vec{u}$ fixed) and let L denote the given line. Clearly, as t varies over E^1, $f(t)$ varies over L; thus f is a map of E^1 *onto* L. This map is also easily proved to be one-to-one, and it becomes an ordered-field–isomorphism if operations and inequalities in L are defined as follows: Let $\bar{x}=f(t)$, $\bar{x}'=f(t')$; then, by definition,

$$x+x'=f(t+t'),\ \bar{x}\bar{x}'=f(tt'),\ \text{and } \bar{x}<\bar{x}'\ \text{iff } t<t'$$

(cf. Problem 10 below).

Problems on Lines, Angles, and Directions in E^n

1. Prove in detail Corollary 1(b) and Corollary 3 of §3. Also show that the angle $\langle\vec{u},\vec{v}\rangle$ does not change if $\vec{u}$ and $\vec{v}$ are multiplied by some scalars of the same sign. What if the scalars are of different signs?

2. Prove geometrically (in E^3) that the dot product $\vec{v}\cdot\vec{u}$, where $\vec{u}$ is a *unit vector*, is the orthogonal (directed) projection of $\vec{v}$ on the directed line $\bar{x}=\bar{a}+t\vec{u}$ (where $\bar{a}$ is arbitrary but fixed). Define analogously projections of vectors on directed lines in E^n.

3. Find the mutual angles between the vectors $\vec{u}$, $\vec{v}$, and $\vec{w}$ specified in Problem 1 of §1 (do cases (a)–(d) separately). Also normalize these vectors and find their direction cosines. Verify by actual computation, in at least one case, that Formula (b) of Corollary 1 in §3 holds. Are any two of the vectors perpendicular?

4. Let $\vec{u}$, $\vec{v} \in E^3$, and let

$$\vec{w} = (u_2v_3 - u_3v_2,\ u_3v_1 - u_1v_3,\ u_1v_2 - u_2v_1).$$

Show that $\vec{u} \perp \vec{w}$ and $\vec{v} \perp \vec{w}$.

Note: The vector $\vec{w}$ so defined is called the *cross product* of $\vec{u}$ and $\vec{v}$ and is denoted by $\vec{u} \times \vec{v}$ or symbolically by the "determinant"

$$\begin{vmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},$$

where $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$ are the basic unit vectors in E^3. Show that

$$\vec{u} \times \vec{v} = -(\vec{v} \times \vec{u})$$

and that in general

$$(\vec{u} \times \vec{v}) \times \vec{x} \neq \vec{u} \times (\vec{v} \times \vec{x}).$$

(Give a counterexample!) Also prove that two lines $\bar{x} = \bar{a} + t\vec{u}$ and $\bar{x} = \bar{b} + t\vec{v}$ in E^3 are parallel iff $\vec{u} \times \vec{v} = 0$. (Note that cross products are defined *only in* E^3.)

5. Find a vector (unit) in E^3, with positive coordinates, which forms equal angles with the axes (i.e., with the basic unit vectors). Solve a similar problem in E^4.

6. Given three points in E^4: $\bar{a} = (0, 0, -1, 2)$, $\bar{b} = (2, 4, -3, -1)$, $\bar{c} = (5, 4, 2, 0)$. Find the angles of the triangle $\bar{a}\,\bar{b}\,\bar{c}$ and the equations of its sides, in nonparametric form. Normalize the equations. For each side give a set of direction numbers and direction cosines.

6'. Let $\bar{b}$ be any point on the line $\bar{x} = \bar{a} + t\vec{u}$. Show that this line coincides with the line $\bar{x} = \bar{b} + \theta\vec{u}$.
[Hint: Let $\bar{b} = \bar{a} + t_0\vec{u}$. Find θ.]

7. A *globe* (solid sphere) in E^n, with center $\bar{p}$ and radius $\epsilon > 0$, is by definition the set

$$\{\bar{x} \in E^n \mid \rho(\bar{x}, \bar{p}) < \epsilon\},$$

denoted $G_{\bar{p}}(\epsilon)$. Show that if $\bar{a}$, $\bar{b} \in G_{\bar{p}}(\epsilon)$, then also $L[\bar{a}, \bar{b}] \subseteq G_{\bar{p}}(\epsilon)$. Prove the same property (called *convexity*) also for the *closed globe*

$$\overline{G}_{\bar{p}}(\epsilon) = \{\bar{x} \in E^n \mid \rho(\bar{x}, \bar{p}) \leq \epsilon\}.$$

Disprove it for the nonsolid *sphere*

$$S_{\bar{p}}(\epsilon) = \{\bar{x} \in E^n \mid \rho(\bar{x}, \bar{p}) = \epsilon\}.$$

[Hint: Take a line *through* $\bar{p}$; say, $\bar{x} = \bar{p} + t\vec{e}_1$. Let $-\epsilon \leq t \leq \epsilon$.]

8. In Problem 6 find the nonparametric equations of the lines through each vertex parallel to the opposite side of the triangle $\bar{a}\,\bar{b}\,\bar{c}$. Find also the points of intersection of these three lines.

9. Prove that if a vector $\vec{v}$ in E^n is perpendicular to each of the n basic unit vectors, i.e., $\vec{v} \cdot \vec{e}_k = 0$, $k = 1, 2, \ldots, n$, then necessarily $\vec{v} = \vec{0}$. Infer that if $\vec{v} \cdot \vec{x} = 0$ for *all* $\bar{x}$, then $\vec{v} = \vec{0}$.

10. Prove that the map f defined in Note 3 of §4 is one-to-one.
[Hint: Show that $t \neq t' \implies |f(t) - f(t')| = \rho(f(t), f(t')) \neq 0$.]
Next, verify that the line L is an ordered field, with zero element $f(0) = a$ and unity $f(1)$, under operations and ordering as defined in Note 3, and that $f(t + t') = f(t) + f(t')$ and $f(tt') = f(t) \cdot f(t')$, by definition. *(Hence infer that f is an isomorphism between the fields E^1 and L.)

11. (i) Given a point $\bar{p} \in E^n$ and a line $\bar{x} = \bar{a} + t\vec{u}$ ($|\vec{u}| = 1$), find the *orthogonal projection* of $\bar{p}$ on the line, i.e., a point $\bar{x}_0 = \bar{a} + t_0\vec{u}$ such that $\overrightarrow{x_0p} \perp \vec{u}$.
[Hint: By Problem 2, $t_0 = (\bar{p} - \bar{a}) \cdot \vec{u}$; verify that $(\bar{p} - \bar{x}_0) \cdot \vec{u} = 0$ if $\bar{x}_0 = \bar{a} + t_0\vec{u}$.]
(ii) Show that

$$\rho(\bar{p}, \bar{x}_0) = |\bar{p} - \bar{x}_0| = \sqrt{|\bar{p} - \bar{a}|^2 - t_0^2} = |\bar{p} - \bar{a}|\,|\sin\alpha|,$$

where $\alpha = \langle \vec{u},\ \bar{p} - \bar{a} \rangle$.
[Hint: Use the formulas

$$|\bar{p} - \bar{x}_0|^2 = (\bar{p} - \bar{x}_0) \cdot (\bar{p} - \bar{x}_0)$$

and

$$|\sin\alpha| = \sqrt{1 - \cos^2\alpha}.]$$

(iii) Noting that $\bar{a}$ is an *arbitrary* point on the line, infer that $\rho(\bar{p}, \bar{x}_0)$ is the *least* distance from $\bar{p}$ to a point $\bar{a}$ on the line.

12. Find the three altitudes of the triangle $\bar{a}\,\bar{b}\,\bar{c}$ of Problem 6. (Use Problem 11.)

13. Given two *nonparallel* lines in E^n: $\bar{x} = \bar{a} + t\vec{u}$ and $\bar{y} = \bar{b} + \theta\vec{v}$, where t, θ are real parameters and $|\vec{u}| = |\vec{v}| = 1$. Find two points $\bar{x}$ and $\bar{y}$ on these lines such that $(\bar{x} - \bar{y}) \perp \vec{u}$ and simultaneously $(\bar{x} - \bar{y}) \perp \vec{v}$. Infer from Problem 11 that, for these points, $\rho(\bar{x}, \bar{y})$ is the shortest distance between a point on one line and a point on the other line.
[Hint: We have to satisfy the simultaneous equations in two unknowns:

$$(\bar{x} - \bar{y}) \cdot \vec{u} = 0 \quad \text{and} \quad (\bar{x} - \bar{y}) \cdot \vec{v} = 0.$$

Substitute $\bar{x} = \bar{a} + t\vec{u}$ and $\bar{y} = \bar{b} + \theta\vec{v}$, and transform the two equations into

$$(\bar{a} - \bar{b}) \cdot \vec{u} + t - \theta(\vec{u} \cdot \vec{v}) = 0 \text{ and } (\bar{a} - \bar{b}) \cdot \vec{v} - \theta + t(\vec{u} \cdot \vec{v}) = 0.$$

Solve for t, θ.]

§5. Hyperplanes in E^n. *Linear Functionals on E^n

I. A plane in E^3 can be geometrically described as follows. Fix a point $\bar{a}$ of the plane and a vector $\vec{u} = \overrightarrow{ab}$ perpendicular to the plane (imagine a pencil standing vertically at $\bar{a}$ on the horizontal plane of the table). Then a point $\bar{x}$ lies on the plane iff $\vec{u} \perp \overrightarrow{ax}$ (the pencil $\vec{u}$ is perpendicular to the line $\overline{ax}$ drawn on the table). It is natural to accept this as a definition in E^n as well (here "planes" are also called "*hyperplanes*").

Definition 1.

By a *hyperplane* (*briefly, plane*) *through a given point* $\bar{a} \in E^n$, *perpendicular to a fixed vector* $\vec{u} \neq \vec{0}$, we mean the set of all points $\bar{x} \in E^n$ such that $\vec{u}$ is perpendicular to $\overrightarrow{ax}$. In symbols, it is the set

$$\{\bar{x} \in E^n \mid \vec{u} \perp \overrightarrow{ax}\}.$$

The vector $\vec{u}$ is called a *normal vector* of the plane (not to be confused with "*normalized* vector"). **Note:** $\vec{u} \neq \vec{0}$.

Since

$$\overrightarrow{ax} = \bar{x} - \bar{a} = (x_1 - a_1,\, x_2 - a_2,\, \ldots,\, x_n - a_n),$$

the formula $\vec{u} \perp \overrightarrow{ax}$, or (by definition) $\vec{u} \cdot (\overrightarrow{ax}) = 0$, can also be written as $\vec{u} \cdot (\bar{x} - \bar{a}) = 0$; or, in terms of coordinates,

$$\sum_{k=1}^{n} u_k(x_k - a_k) = 0, \text{ where } \vec{u} \neq 0 \text{ (i.e., } \textit{not all } u_k \textit{ vanish}). \tag{1}$$

Formula (1) is called the *coordinate equation* of the plane, while the formula $\vec{u} \cdot (\bar{x} - \bar{a}) = 0$ is its *vector equation*. We briefly refer to the plane by giving its equation; e.g.,

$$\text{"the plane } \sum_{k=1}^{n} u_k(x_k - a_k) = 0\text{"}$$

(with the numbers u_k and a_k as specified). The plane consists of exactly the points $\bar{x}$ whose coordinates satisfy the equation of the plane. Removing brackets in (1) and transposing the constant terms, we obtain

$$u_1x_1 + u_2x_2 + \cdots + u_nx_n = c \quad \Big(\text{where } c = \sum_{k=1}^{n} u_ka_k\Big). \tag{2}$$

Algebraically, this is a *linear equation* in the variables x_k, with given coefficients u_k (not all 0) and constant term c. Thus *every hyperplane in* E^n *has a linear coordinate equation*, i.e., one of the form (2). Conversely, given any

equation of that form, with at least one of the u_k (say, u_1) not zero, we can rewrite it in the form

$$u_1\Big(x_1 - \frac{c}{u_1}\Big) + u_2x_2 + \cdots + u_nx_n = 0.$$

Then, setting $a_1 = \dfrac{c}{u_1}$ and $a_k = 0$ for $k \geq 2$, we obtain from it an equation of the form (1), representing a hyperplane through

$$\bar{a} = \Big(\frac{c}{u_1},\, 0,\, \ldots,\, 0\Big),$$

perpendicular to $\vec{u} = (u_1,\, \ldots,\, u_n)$. Thus we have proved the following proposition.

Theorem 1. *A set* $A \subset E^n$ *is a hyperplane iff* A *is exactly the set of all points* $\bar{x} = (x_1,\, \ldots,\, x_n)$ *satisfying some equation of the form* (2), *with at least one of the coefficients* u_k *not* 0. *These coefficients are the components of a vector* $\vec{u} = (u_1,\, \ldots,\, u_n)$ *normal to the plane.*

In this connection, (2) is called the *general* equation of a hyperplane. Clearly, we obtain an equivalent equation (representing the same point set) if we multiply both sides of (2) by a nonzero scalar q. Then u_k is replaced by qu_k, i.e., $\vec{u}$ is replaced by $q\vec{u}$. This shows that we may replace the normal vector $\vec{u}$ by any scalar multiple $q\vec{u}$ ($q \neq 0$), without changing the hyperplane. In particular, setting $q = 1/|\vec{u}|$, we replace $\vec{u}$ by its unit $\vec{u}/|\vec{u}|$ and get

$$\frac{1}{|\vec{u}|}(u_1x_1 + \cdots + u_nx_n) = \frac{c}{|\vec{u}|}, \text{ with } |\vec{u}| = \sqrt{\sum_{k=1}^{n} u_k^2}. \tag{3}$$

This is called the *normalized* or *normal* equation of the hyperplane. Actually, there are *two* normal equations since we may also replace $\vec{u}$ by $-\vec{u}$, changing all signs in (3), i.e., changing the direction of $\vec{u}$. If, however, the direction is *prescribed*, we speak of a *directed* hyperplane.

If all but one coefficients u_k vanish, then $\vec{u}$ becomes a scalar multiple of the corresponding basic unit vector $\vec{e}_k$; the plane is then perpendicular to $\vec{e}_k$, and we say that it is "*perpendicular to the k-th axis*". Equation (2) then turns into $u_kx_k = c$ or $x_k = c_k$, where $c_k = c/u_k$; e.g., $x_1 = 5$ is the equation of a plane perpendicular to $\vec{e}_1$. It consists of all $\vec{x} \in E^n$, with $x_1 = 5$. Imitating geometry in E^3, we also define the following:

The *angle* between two hyperplanes with normal vectors $\vec{u}$ and $\vec{v}$ is, by definition, the angle $\langle \vec{u}, \vec{v} \rangle$ between these vectors. Actually, unless the hyperplanes are directed, there are *two* angles: $\langle \vec{u}, \vec{v} \rangle$ and $\langle -\vec{u}, \vec{v} \rangle$. In particular, the hyperplanes are *perpendicular* to each other iff $\vec{u} \perp \vec{v}$ and *parallel* to each other iff $\vec{u} = q\vec{v}$ or $\vec{v} = q\vec{u}$ for some $q \in E^1$ (i.e., if $\vec{u}$ and $\vec{v}$ are collinear). The angle between a hyperplane (with normal vector $\vec{u}$) and a line with direction vector $\vec{v}$ is, by definition, the *complement* of $\langle \vec{u}, \vec{v} \rangle$. It may be defined as the angle α

whose cosine equals $\sin\langle \vec{u}, \vec{v}\rangle = \pm\sqrt{1-\cos^2\langle \vec{u}, \vec{v}\rangle}$. (Clearly, there are *two* such angles.) Accordingly, the plane and the line are said to be *parallel* if $\vec{u} \perp \vec{v}$ and *perpendicular* if $\vec{u} \parallel \vec{v}$. A set of points in E^n is said to be *coplanar* if it is contained in some hyperplane. A set of vectors in E^n is *vector-coplanar* iff these vectors are perpendicular to some fixed vector $\vec{u} \in E^n$; so are, e.g., any $n-1$ of the basic unit vectors $\vec{e}_k$, because all of them are perpendicular to the remaining $\vec{e}_k$.

***II.** Consider again the left side of equation (2), *without the constant term* c:

$$\sum_{k=1}^{n} u_k x_k,$$

or, in vector form, $\bar{u} \cdot \bar{x}$. Let us define a map $f\colon E^n \to E^1$, setting $(\forall \bar{x} \in E^n)$ $f(\bar{x}) = \bar{u} \cdot \bar{x}$, with $\bar{u}$ fixed. By properties of dot products (Theorem 1 of §2), we have, for any $\bar{x}$, $\bar{y} \in E^n$ and $a \in E^1$,

$$\bar{u} \cdot (\bar{x} + \bar{y}) = \bar{u} \cdot \bar{x} + \bar{u} \cdot \bar{y} \text{ and } \bar{u} \cdot (a\bar{x}) = a(\bar{u} \cdot \bar{x});$$

or, since $\bar{u} \cdot \bar{x} = f(\bar{x})$,

$$f(\bar{x} + \bar{y}) = f(\bar{x}) + f(\bar{y}) \text{ and } f(a\bar{x}) = af(\bar{x}) \tag{4}$$

for all $\bar{x}$, $\bar{y} \in E^n$, $a \in E^1$.

It follows that $(\forall a, b \in E^1)$ $(\forall \bar{x}, \bar{y} \in E^n)$

$$f(a\bar{x} + b\bar{y}) = f(a\bar{x}) + f(b\bar{y}) = af(\bar{x}) + bf(\bar{y}).$$

By induction (which we leave to the reader), given any scalars $a_1, a_2, \dots, a_m \in E^1$ and vectors $\bar{x}_1, \dots, \bar{x}_m \in E^n$, we obtain

$$f\left(\sum_{i=1}^{m} a_i \bar{x}_i\right) = \sum_{i=1}^{m} a_i f(\bar{x}_i). \tag{5}$$

In other words, *the map f carries every linear combination of vectors $\bar{x}_1, \dots, \bar{x}_m$ in E^n into the corresponding linear combination of the function values $f(\bar{x}_i)$, $i = 1, 2, \dots, m$.* We express this by saying that *f preserves linear combinations*, or *preserves vector addition and multiplication by scalars*. Mappings with that property turn out to be of great importance for the theory of vector spaces in general (cf. §1, Note 3). They are called *linear maps* (because they preserve linear combinations).

In particular, for Euclidean spaces E^n and E^r, we have the following.

Definition 2.

A mapping $f\colon E^n \to E^r$ is said to be *linear* iff it preserves linear combinations, i.e., satisfies (4) and hence (5) (see above). Linear maps of E^n *into* E^1, $f\colon E^n \to E^1$ $(r = 1)$, are called *linear functionals*.

Theorem 2. *A mapping $f\colon E^n \to E^1$ is a linear functional iff there is a vector $\bar{u} \in E^n$ such that $(\forall \bar{x} \in E^n)$*

$$f(\bar{x}) = \bar{u} \cdot \bar{x} = \sum_{k=1}^{n} u_k x_k.^{1}$$

Proof. If such a vector $\bar{u}$ exists then, as was shown above, f satisfies (4) and hence is linear. Conversely, if f is linear, then f preserves linear combinations. Now, by Theorem 2 of §1, *every* $\bar{x} \in E^n$ is such a combination, namely,

$$\bar{x} = \sum_{k=1}^{n} x_k \bar{e}_k.$$

Thus, by (5),

$$f(\bar{x}) = f\left(\sum_{k=1}^{n} x_n \bar{e}_k\right) = \sum_{k=1}^{n} x_k f(\bar{e}_k), \quad \bar{x} \in E^n.$$

Here, since f is a map *into* E^1, the function values $f(\bar{e}_k)$ are in E^1, i.e., certain real numbers. Then let $f(\bar{e}_k) = u_k \in E^1$, $k = 1, 2, \dots, n$, and set $\bar{u} = (u_1, \dots, u_n)$. Then we have, for all $\bar{x} \in E^n$,

$$f(\bar{x}) = \sum_{k=1}^{n} x_k f(\bar{e}_k) = \sum_{k=1}^{n} x_k u_k = \bar{x} \cdot \bar{u} = \bar{u} \cdot \bar{x},$$

by the properties of dot products. Thus u is the desired vector, and all is proved. □

Note 1. *The vector $\bar{u}$ of Theorem 2 is unique.* Indeed, suppose there are *two* vectors, $\bar{u}$ and $\bar{v}$ such that

$$\bar{u} \cdot \bar{x} = f(\bar{x}) = \bar{v} \cdot \bar{x}$$

for all $\bar{x} \in E^n$. Then

$$(\bar{u} - \bar{v}) \cdot \bar{x} = \bar{u} \cdot \bar{x} - \bar{v} \cdot \bar{x} = 0$$

for *all* $\bar{x} \in E^n$. But, by Problem 9 of §4, this implies that $\bar{u} - \bar{v} = \bar{0}$, i.e., $\bar{u} = \bar{v}$ after all. Thus $\bar{u}$ is unique indeed.

We now establish a connection between hyperplanes and those linear functionals that are *not identically zero*.[2]

[1] In other words, *all* linear functionals on E^n are of the kind that we considered above, i.e., arise from dot products, as in equation (2).

[2] We say that a function $f\colon E^n \to E^1$ is *identically zero*, and write $f \equiv 0$, iff $f(\bar{x}) = 0$ for *all* $\bar{x} \in E^n$. Otherwise, we write $f \not\equiv 0$. The latter means that $f(\bar{x}) \neq 0$ for *at least one* $\bar{x} \in E^n$.

Our next theorem shows that hyperplanes are exactly all those sets in E^n whose equations are of the form $f(\bar{x}) = c$, where f is a linear functional not identically 0 and c is a real constant. More precisely, we have the following result.

Theorem 3. *A set $A \subseteq E^n$ is a hyperplane iff there is a linear functional $f\colon E^n \to E^1$, $f \not\equiv 0$, and some $c \in E^1$, such that*

$$A = \{\bar{x} \in E^n \mid f(\bar{x}) = c\},$$

i.e., A consists of exactly those points $\bar{x} \in E^n$ for which $f(\bar{x}) = c$.

Proof. If A is a hyperplane, its general equation (2) may also be written as $\bar{u} \cdot \bar{x} = c$ (since $\bar{u} \cdot \bar{x}$ is, *by definition*, the left-hand side of (2)). Thus $A = \{\bar{x} \in E^n \mid \bar{u} \cdot \bar{x} = c\}$. Setting $f(\bar{x}) = \bar{u} \cdot \bar{x}$, we obtain a *linear functional* $f\colon E^n \to E^1$, by Theorem 2. Then $A = \{\bar{x} \in E^n \mid f(\bar{x}) = c\}$. Moreover, as $\bar{u} \neq \bar{0}$ in (2), f is *not* $\equiv 0$ (Problem 9 of §4). Thus A is as stated in Theorem 3.

Conversely, if $A = \{\vec{x} \in E^n \mid f(\bar{x}) = c\}$, with f a linear functional $\not\equiv 0$, then again Theorem 2 yields a vector $\bar{u} \neq \bar{0}$ such that

$$f(\bar{x}) = \bar{u} \cdot \bar{x} = \sum_{k=1}^{n} u_k x_k$$

for all $\bar{x} \in E^n$. Then we obtain

$$A = \{\bar{x} \in E^n \mid f(\bar{x}) = c\} = \left\{\bar{x} \in E^n \mid \sum_{k=1}^{n} u_k x_k = c\right\},$$

and this means that A is exactly the set of points satisfying equation (2), i.e., a *hyperplane*. Thus all is proved. □

Note 2. This theorem could be accepted as an alternative *definition* of a hyperplane. It has the advantage that it replaces the notion of dot products by that of a linear functional, without any reference to "angles" or orthogonality (which are defined in *Euclidean* spaces only; cf. Note 4 in §3).

Examples.

(1″) Let $\bar{a} = (1, -2, 0, 3)$ and $\vec{u} = (1, 1, 1, 1)$ in E^4. Then the plane normal to $\vec{u}$ through $\bar{a}$ has the equation

$$(\bar{x} - \bar{a}) \cdot \vec{u} = \sum_{k=1}^{4} (x_k - a_k) u_k = 0,$$

or

$$(x_1 - 1) \cdot 1 + (x_2 + 2) \cdot 1 + (x_3 - 0) \cdot 1 + (x_4 - 3) \cdot 1 = 0,$$

or $x_1 + x_2 + x_3 + x_4 = 2$. The corresponding linear functional $f\colon E^4 \to E^1$ is defined by $f(\bar{x}) = x_1 + x_2 + x_3 + x_4$.

(2″) The two linear equations

$$x + 3y - 2 = 1 \text{ and } 2x + y - z = 0$$

(where x, y, z stand for x_1, x_2, x_3) represent two planes in E^3 with normal vectors

$$\vec{u} = (1, 3, -2) \text{ and } \vec{v} = (2, 1, -1),$$

respectively. (Note that, by formulas (1) and (2), the components u_k of the normal vector are exactly the coefficients of the variables x_k, here denoted by x, y, z; thus, in the first plane, $u_1 = 1$, $u_2 = 3$ and $u_3 = -2$, so that $\vec{u} = (1, 3, -2)$; similarly for $\vec{v}$.)

The corresponding linear functionals on E^3 (call them f and g, respectively) are given by

$$f(x, y, z) = x + 3y - 2z \text{ and } g(x, y, z) = 2x + y - z$$

(these are the left sides of the equations of the planes, without the constant terms). The second plane passes through $\bar{0}$ (why?), and so its vector equation is $(\bar{x} - \bar{0}) \cdot \vec{v} = 0$ or $\vec{x} \cdot \vec{v} = 0$, where $\vec{v} = (2, 1, -1)$. The equation of the first plane can be rewritten as

$$(x_1 - 1) + 3(x_2 - 0) - 2(x_3 - 0) = 0;$$

it passes through $\bar{a} = (1, 0, 0)$, and its vector equation is $(\bar{x} - \bar{a}) \cdot \vec{u} = 0$, with $\bar{a}$ and $\vec{u}$ as above. The angle between the planes is given by

$$\cos\langle \vec{u}, \vec{v} \rangle = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}|\,|\vec{v}|} = \frac{7}{\sqrt{14 \cdot 6}} = \frac{7}{\sqrt{84}} = \frac{7}{2\sqrt{21}}.$$

Their normalized equations are

$$\frac{x + 3y - 2z - 1}{\sqrt{14}} = 0 \text{ and } \frac{2x + y - z}{\sqrt{6}} = 0.$$

Problems on Hyperplanes in E^n (cf. also §6)

1. Given a hyperplane $3x_1 + 5x_2 - x_3 + 2x_4 = 9$ in E^4, find

(i) a few points that lie on it, and some that do not;

(ii) a unit vector normal to the plane (thus normalize the equation);

(iii) the angles between the plane and the basic unit vectors $\vec{e}_k$;

(iv) the equations of the planes parallel to the given plane and passing through

(a) the origin;

(b) $\bar{p} = (2, 1, 0, -1)$;

(v) the equations of the line through $\bar{0}$, perpendicular to the plane;

(vi) the *intercepts* of the plane, i.e., four numbers a, b, c, d such that the points $(a, 0, 0, 0)$, $(0, b, 0, 0)$, $(0, 0, c, 0)$, and $(0, 0, 0, d)$ lie on the plane (at these points the plane meets the four "axes");

(vii) the angle between the plane and the line

$$\frac{x_1 - 1}{3} = \frac{x_2}{4} = \frac{x_3}{5} = \frac{x_4 + 2}{-1};$$

(viii) the point of intersection of the plane and line given in (vii).

[Hint: Using parametric equations, express x_1, x_2, x_3, and x_4 in terms of t and substitute in the equation of the plane to evaluate t. Explain!]

2. Find the normal equation of the hyperplane in E^4 that

(i) is perpendicular to the line given in Problem 1(vii) and passes through the point

(a) $\bar{p} = (3, 1, -2, 0)$;

(b) $\bar{p} = (-1, 2, 1, 1)$;

(ii) is perpendicular to, and bisects, the line segment $L(\bar{a}, \bar{b})$, where $\bar{a} = (0, -1, 2, 2)$, $\bar{b} = (2, -3, 0, 4)$ (first find the midpoint of $L(\bar{a}, \bar{b})$);

(iii) contains the points

$$(2, 0, 0, -1),\ (-3, 0, 2, 3),\ (1, 1, 2, 0),\ \text{and } (0, 0, 0, 0).$$

[Hint for (iii): As the points lie on the plane, their coordinates satisfy its general equation, $ax_1 + bx_2 + cx_3 + dx_4 = e$. Substituting them, obtain four equations in the unknowns a, b, c, d, e. Solve them for the *ratios* b/a, c/a, d/a, e/a (assuming $a \neq 0$) and substitute into

$$x_1 + \frac{b}{a}\,x_2 + \frac{c}{a}\,x_3 + \frac{d}{a}\,x_4 = \frac{e}{a}.$$

This is the required equation.]

3. A reader acquainted with the theory of determinants will verify that the equation of a hyperplane in E^n through n given points $\bar{a}_1, \ldots, \bar{a}_n$ is

$$\begin{vmatrix} x_1 & x_2 & \ldots & x_n & 1 \\ a_{11} & a_{12} & \ldots & a_{1n} & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & 1 \end{vmatrix} = 0, \tag{6}$$

provided the determinant does not vanish *identically*, i.e., regardless of the choice of the point $\bar{x} = (x_1, x_2, \ldots, x_n)$.

[Hint: Each of the n points $\bar{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$ when substituted for $\bar{x} = (x_1, \ldots, x_n)$ in (6) makes the determinant *vanish* (for two rows become equal). Thus all $\bar{a}_i$ satisfy equation (6) and so lie in the plane represented by (6) (the equation being *linear* in $x_1, \ldots, x_n$, upon expansion by elements of the first row).]

Use this result for another solution of Problem 2(iii).

4. Show that the perpendicular distance from a point $\bar{p}$ to a hyperplane

$$\sum_{k=1}^{n} u_k x_k = c$$

(or $\bar{u} \cdot \bar{x} = c$, where $\bar{u}$ is a normal vector) in E^n is given by

$$\rho(\bar{p}, \bar{x}_0) = \frac{|\bar{u} \cdot \bar{p} - c|}{|\bar{u}|}.$$

(Here $\bar{x}_0$ is the *orthogonal projection* of $\bar{p}$, i.e., a point on the plane such that $\overrightarrow{px_0}$ is perpendicular to the plane.)

[Hint: Consider the line $\bar{x} = \bar{p} + t\vec{v}$, where $\vec{v} = -\vec{u}/|\vec{u}|$, and find the value of t for which $\bar{x} = \bar{p} + t\vec{v}$ lies on both the line and the plane. Then $|t| = \rho(\bar{p}, \bar{x}_0)$.]

Note. For a *directed* plane, this t is called the *directed* distance from $\bar{p}$ to the plane (it may be *negative*). Unless otherwise stated, the direction of the plane is so chosen that the constant c in $\bar{u} \cdot \bar{x} = c$ is *positive*. Thus the directed distance is defined *always*, except when $c = 0$.

5. Let $P = 0$ and $P' = 0$ be the equations of two intersecting planes in E^3. (Here P stands for $\sum_{k=1}^{3} u_k x_k - c$, and P' stands for $\sum_{k=1}^{3} v_k x_k - d$.) Show that, for any choice of k, $k' \in E^1$, the equation

$$kP + k'P' = 0$$

represents a plane passing through the intersection line of the planes $P = 0$ and $P' = 0$, and that *all* such planes in E^3 can be obtained by a suitable choice of k and k'. **Note:** $kP + k'P' = 0$ is called the equation of the *pencil of planes* passing through the intersection line of the two given planes; k, k' are called *parameters*.

[Hint: To show that *all* the required planes can be so obtained, take any point $\bar{p} \in E^3$ and prove that the parameters k, k' can always be so chosen that the plane $kP + k'P' = 0$ passes through $\bar{p}$.]

6. Find the direction cosines of the intersection line of two planes in E^3:

$$2x - 3y + z = 4 \text{ and } x + y - 2 = 1.$$

Also give a set of parametric equations for the line.

[Hint: The points of the line satisfy the equations of *both* planes, hence also all equations that follow from them by eliminating one of the variables x, y, z. Thus, obtain two equations: one in x and y, the other in x and z only. Choose x as the parameter t: $x = t$, and also express y and z in terms of t, thus obtaining the parametric equations.]

7. From Problem 4, find the distance between two parallel planes: $\bar{u}\cdot\bar{x}=c$ and $\bar{u}\cdot\bar{x}=d$ in E^n. (Answer: $|c-d|/|\vec{u}|$.) Give an example in E^3.

§6. Review Problems on Planes and Lines in E^3

1. Determine whether the plane $4x-y+3z+1=0$ contains the points $(-1,\ 6,\ 3)$, $(3,\ -2,\ -5)$, $(0,\ 4,\ 1)$, $(2,\ 0,\ 5)$, $(2,\ 7,\ 0)$, $(0,\ 1,\ 0)$.

2. A point M moves from $(5,\ -1,\ 2)$ in a direction parallel to OY. At what point will it meet the plane $x-2y-3z+2=0$?

3. What special properties have the planes

(a) $3x-5z+1=0$? (b) $9y-2=0$?

(c) $x+y-5=0$? (d) $2x+3y-7z=0$?

(e) $8y-3z=0$?

4. Find equations of the planes

(a) parallel to the XOY-plane and passing through $(2,\ -5,\ 3)$;

(b) containing OZ and the point $(-3,\ 1,\ -2)$;

(c) parallel to OX and passing through $(4,\ 0,\ -2)$ and $(5,\ 1,\ 7)$.

5. Find the x, y, z intercepts of the planes

(a) $2x-3y-z+12=0$; (b) $5x+y-3z-15=0$;

(c) $x-y+z-1=0$; (d) $x-4z+6=0$;

(e) $5x-2y+z=0$.

6. Draw the lines of intersection between the coordinate planes and the plane $5x+2y-3z-10=0$.

7. The plane $3x+y-2z=18$ and the coordinate planes form a tetrahedron $OABC$. Find the sides of the cube inscribed in that tetrahedron, with one vertex lying in the given plane, while three faces of the cube lie in the coordinate planes.

8. Find an equation of the plane passing through $(7,\ -5,\ 1)$ and marking off equal positive intercepts on the three coordinate axes.

9. A tetrahedron $OABC$ lying in the second octant has three of its faces in the coordinate planes. Find an equation of the fourth face, given that three of its edges equal $CA=5$, $BC=\sqrt{29}$, and $AB=6$.

10. Normalize the equations of the planes

(a) $2x-9y+6z=22$,

(b) $10x+2y-11z=0$, and

(c) $6x-yx-z=33$.

11. Find the distance from the origin $\bar{0}$ to the plane $15x-10y+6z=190$.

12. Find the plane whose distance from the origin equals 6, given the ratios between its intercepts: $a:b:c=1:3:2$.

13. Find the direction cosines of the line perpendicular to the plane $2x-y+2z=-9$.

14. Repeat Problem 13, assuming the line is perpendicular to the plane with intercepts are $a=11$, $b=55$, $c=10$.

15. Find the angle between the planes YOZ and $x-y+\sqrt{2}\,z=5$.

16. Find the point symmetric to $\bar{0}$ with respect to the plane $x-y+\sqrt{2}\,z=5$.

17. Find an equation of the plane given that the perpendicular dropped on it from the origin meets the plane at $(3,\ -6,\ 2)$.

18. Find the distance between the given point and the given plane:

(a) $(3,\ 1,\ -1)$, $22x+4y-20z=45$.

(b) $(4,\ 3,\ -2)$, $3x-y+5z+1=0$.

(c) $(2,\ 0,\ -1/2)$, $4x-4y+2z=17$.

19. Find the altitude $h_{\bar{a}}$ of the pyramid with vertices $(0,\ 6,\ 4)=\bar{a}$, $(1,\ -1,\ 4)$, $(-2,\ 11,\ -5)$, and $(3,\ 5,\ 3)$.

20. Find an equation of the plane through $(7,\ 4,\ 4)$ perpendicular to $\overline{ab}$ if $\bar{a}=(1,\ 3,\ -2)$, $\bar{b}=(1,\ -1,\ 0)$.

21. Find the point symmetric to $(1,\ 2,\ 3)$ with respect to the plane $-3x+y+z=1$.

22. The plane of a mirror is $2x-6y+3z=42$. Find the image of $(3,\ 7,\ 5)$.

23. Find the angle between the two given planes:

(a) $x-4y-z+9=0$ and $4x-5y+3z=1$;

(b) $3x-y+2z=-15$ and $5x+9y-3z=1$;

(c) $6x+2y-4z=17$ and $9x+3y-6z=4$.

24. Find the angle between two planes through $(-5,\ 16,\ 12)$ given that one of them contains the axis OX and the other contains OY.

25. Find equations of the planes

(a) through $(-2,\ 7,\ 3)$ and parallel to the plane $x-4y+5=1$;

(b) through the origin and perpendicular to the two planes $2x-y+5z=-3$ and $x+3y-z=7$;

(c) passing through (3, 0, 0) and (0, 0, 1) and forming an angle of 60° with the plane XOY.

26. Find an equation of the plane containing the OZ-axis and forming an angle of 60° with the plane $2x + y - \sqrt{5}\,z = 7$.

27. Verify that the planes

$$2x - 2y + z = 3,\ 3x - 6z + 1 = 0,\ \text{and}\ 4x + 5y + 2z = 0$$

are perpendicular to each other, and find the transformation formulas to a system of coordinates in which these planes would become, respectively, the XOY, YOZ, and ZOX planes.

In the following problems, the results of Problems 4–6 of §5, are used.

28. Given the points (6, 1, −1), (0, 5, 4), and (5, 2, 0), find the plane whose distances from these points are −1, 3, and 0, respectively.

29. Find the planes bisecting the angles between the planes

$$3x - y + 7z = 4 \text{ and } 5x + 3y - 5z + 2 = 0.$$

30. Find a point on the OZ-axis equidistant from the two planes

$$x + 4y - 3z = 2 \text{ and } 5x + z + 8 = 0.$$

31. Find the distance between the planes

$$11x - 2y - 10z = 45 \text{ and } 11x - 2y - 10z = -15.$$

(First check that they are parallel.)

32. Find the center of the sphere inscribed in the tetrahedron formed by the plane $2x + 3y - 6z = 4$ and the coordinate planes.

33. Find the planes parallel to the plane $14 + 3x - 6y - 2z = 0$ given that the distance between the latter and each of them is 3.

34. Find the plane passing through $\bar{0}$ and the points (1, 4, 0), (3, −2, 1).

35. Find the equations of the faces of the tetrahedron with vertices (0, 0, 2), (3, 0, 5), (1, 1, 0), (4, 1, 2).

36. Find the volume of the tetrahedron of Exercise 35.

37. Verify the coplanarity or noncoplanarity of the points

(a) (3, 1, 0), (0, 7, 2), (−1, 0, −5), (4, 1, 5);

(b) (4, 0, 3), (1, 3, 3), (0, 2, 4), (1, −1, 1).

38. Find the intersection point of the given three planes:

(a) $2x - 3y + 2z = 9$, $x + 2y + 3z = 1$, $5x + 8y - z = 7$;

(b) $-3x + 12y + 6z = 7$, $3x + y + z = 5$, $x - 4y - 2z + 3 = 0$;

(c) $3x - z + 5 = 0$, $5x + 2y - 13z = -23$, $2x - y + 5z = 4$.

39. Verify whether the four given planes meet at a single point:

(a) $5x - z = -3$, $2x - y + 5z = 4$, $3y + 2z = 1$, $3x + 4y + 5z = 3$;

(b) $5x + 2y = 6$, $x + y = 3$, $2x - 3y + z = -8$, $3x + 2z = 1$.

40. A plane passes through the line of intersection of the planes

$$x + 5y + 2 = z \text{ and } 4x + 3 - y = 1.$$

Find its equation if

(a) it passes through the origin;

(b) it passes through (1, 1, 1);

(c) it is parallel to OY;

(d) it is perpendicular to the plane $2x - y + 5z = 3$.

41. In the pencil of planes determined by the planes $3x + y + 3z = 2$ and $x - 2y + 5z = 1$, find planes perpendicular to these planes.

42. Find an equation of the plane perpendicular to the plane $5x - y + 3z = 2$ and intersecting with it along a line lying in the XOY plane.

43. Find an equation of the plane tangent to the sphere

$$x^2 + y^2 + z^2 = 1$$

and containing the intersection line of the planes

$$5x + 8y + 1 = z \text{ and } x + 28y + 17 = 2z.$$

(For the notion of "sphere", cf. Problem 7 of §4.)

44. In the pencil of planes

$$x + 3y - 5 + t(x - y - 2z + 4) = 0,$$

find a plane with equal intercepts a, b, c.

45. Which of the coordinate planes belongs to the pencil of planes

$$4x - y + 2z - 6 + t(6x + 5y + 3z - 9) = 0?$$

46. Find the plane passing through the intersection line of the planes $x + 5y + z = 0$ and $z = 4$ at an angle of 45° to the plane $x - 4y - 8z = -12$.

47. Find the three planes that are each parallel to a coordinate axis and pass through the line

$$\frac{x-3}{2} = \frac{y+1}{-1} = \frac{z+3}{4}.$$

48. Verify that the given two lines intersect and find the intersection point, as well as the equation of the plane passing through them:

(i) $\dfrac{x-2}{-3}=\dfrac{y}{2}=\dfrac{z+5}{5}$ and $\dfrac{x+15}{-7}=\dfrac{y+4}{-3}=\dfrac{z-8}{4}$;

(ii) $\dfrac{x+1}{0}=\dfrac{y+1}{5}=\dfrac{z-3}{3}$ and $\dfrac{x-8}{3}=\dfrac{y+2}{-2}=\dfrac{z-6}{0}$;

(iii) $x=4+3t$, $y=7+6t$, $z=-10-2t$ and $x=-3-t$, $y=5t$, $z=2+8t$.

49. In each case find the direction cosines and parametric equations of the intersection line of the two given planes:

(i) $x-2y+3z+4=0$, $2x+3y-z=0$;

(ii) $4x-y+5z=2$, $3x+3y-2z=7$.

50. In Problem 49, find the an equation of plane passing through line (i) and parallel to line (ii).

51. Find the perpendicular distance from the point $\bar{p}=(2,\,-1,\,2)$ to the line

(i) $\dfrac{x-1}{2}=\dfrac{y}{1}=\dfrac{z+2}{-3}$;

(ii) $\dfrac{x+5}{3}=\dfrac{y-1}{-1}=\dfrac{z+4}{5}$.

Also find the perpendicular distance between the two lines.

[Hint: Cf. Problems 11 and 13 of §4. Alternatively, project (orthogonally) the vector $\overrightarrow{(1,\,0,\,-2)(-5,\,1,\,-4)}$ on the unit vector perpendicular to both lines using cross products; cf. Problems 4 and 2 of §4.]

§7. Intervals in E^n

Consider the rectangle in E^2 shown in Figure 18. Its interior (without the perimeter) consists of all points $(x,\,y)\in E^2$ such that

$$a_1<x<b_1 \text{ and } a_2<y<b_2,$$

i.e.,

$$x\in(a_1,\,b_1) \text{ and } y\in(a_2,\,b_2).$$

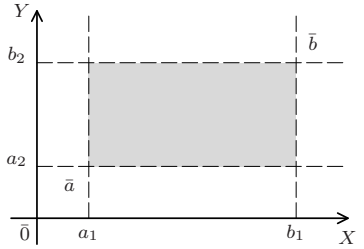


FIGURE 18

Thus it is *the cross product of two line intervals*, $(a_1,\,b_1)$, $(a_2,\,b_2)$. To include also all or some sides, we would have to replace open line intervals by closed, half-closed, or half-open ones. Similarly, cross products of three line intervals yield rectangular parallelepipeds in E^3. We may also consider cross products of n line intervals. This leads us to the following definition.

Definition 1.

By an *interval* in E^n, we mean the Cartesian product of any n intervals in E^1 (some may be open, some closed or half-open, etc.).

In particular, given $\bar{a}=(a_1,\,\dots,\,a_n)$ and $\bar{b}=(b_1,\,\dots,\,b_n)$, with $a_k\le b_k$, $k=1,\dots,n$, we define the *open interval* $(\bar{a},\,\bar{b})$, the *closed interval* $[\bar{a},\,\bar{b}]$, the *half-open interval* $(\bar{a},\,\bar{b}]$, and the *half-closed interval* $[\bar{a},\,\bar{b})$ as follows. First,

$$\begin{aligned}(\bar{a},\,\bar{b})&=(a_1,\,b_1)\times(a_2,\,b_2)\times\cdots\times(a_n,\,b_n)\\&=\{\bar{x}\in E^n\mid a_k<x_k<b_k,\ k=1,\,2,\,\dots,\,n\}.\end{aligned}$$

Thus $(\bar{a},\,\bar{b})$, the cross product of n *open* line intervals $(a_k,\,b_k)$, is the set of all those points $\bar{x}$ in E^n whose coordinates x_k all satisfy the inequalities $a_k<x_k<b_k$, $k=1,\dots,n$. Similarly,

$$\begin{aligned}[\bar{a},\,\bar{b}]&=[a_1,\,b_1]\times[a_2,\,b_2]\times\cdots\times[a_n,\,b_n]\\&=\{\bar{x}\in E^n\mid a_k\le x_k\le b_k,\ k=1,\,2,\,\dots,\,n\};\\(\bar{a},\,\bar{b}]&=(a_1,\,b_1]\times(a_2,\,b_2]\times\cdots\times(a_n,\,b_n]\\&=\{\bar{x}\in E^n\mid a_k<x_k\le b_k,\ k=1,\,2,\,\dots,\,n\};\\[\bar{a},\,\bar{b})&=[a_1,\,b_1)\times[a_2,\,b_2)\times\cdots\times[a_n,\,b_n)\\&=\{\bar{x}\in E^n\mid a_k\le x_k<b_k,\ k=1,\,2,\,\dots,\,n\}.\end{aligned}$$

While in E^1 there are only these four types of intervals, in E^n we can form many more kinds of them by cross-multiplying *different* (mixed) kinds of line intervals. In all cases, the points $\bar{a}$ and $\bar{b}$ are called the *endpoints* of the interval. If $a_k=b_k$ for some k, the interval is called *degenerate*. We often denote intervals by *single* capitals; e.g., $A=(\bar{a},\,\bar{b})$.

Note 1. A point $\bar{x}$ belongs to $(\bar{a},\,\bar{b})$ only if the inequalities $a_k<x_k<b_k$ hold *simultaneously* for $k=1,\,2,\,\dots,\,n$. This is impossible if $a_k=b_k$ for some k. Thus *a degenerate open interval is always empty*. Similarly for other *nonclosed* intervals. A closed interval contains at least its endpoints $\bar{a}$, $\bar{b}$.

Definition 2.

If $\bar{a}$ and $\bar{b}$ are the endpoints of an interval A in E^n, their distance $\rho(\bar{a},\bar{b})=|\bar{b}-\bar{a}|$ is called the *diagonal* dA of A; the n differences $b_k-a_k=\ell_k$ are called its n *edgelengths*; their product

$$\prod_{k=1}^{n}\ell_k=\prod_{k=1}^{n}(b_k-a_k)$$

is called the *volume* of A (in E^2 it is its *area*, in E^1 its *length*), denoted $\operatorname{vol} A$ or vA.

The point $\bar{c} = \frac{1}{2}(\bar{a} + \bar{b})$ is called the *center* of A.

The set difference $[\bar{a}, \bar{b}] - (\bar{a}, \bar{b})$ is called the *boundary* of any interval with endpoints $\bar{a}$ and $\bar{b}$; it consists of $2n$ "faces" defined in a natural manner. (How?)

If all edgelengths $\ell_k = a_k - b_k$ are *equal*, A is called a *cube* (in E^2, a *square*).

If one of the ℓ_k is 0, then A is degenerate and $\operatorname{vol} A$, being the product of *all* the ℓ_k, is 0.

In E^2, we can split an interval into two subintervals by drawing a *line* (in E^3, a *plane*) perpendicular to one of the axes (see Figure 19 below). To "imitate" this in E^n, we use *hyperplanes* (see §5). A hyperplane perpendicular to the k-th axis (i.e., to $\vec{e}_k$) can be defined as the set of all those points $\bar{x}$ in E^n whose k-th coordinate equals some fixed number c (the other coordinates may be arbitrary). Briefly, we call it "*the hyperplane* $x_k = c$". If $a_k < c < b_k$ ($\bar{a}$ and $\bar{b}$ being the endpoints of A), then A splits into two disjoint sets:

$$P = \{\bar{x} \in A \mid x_k < c\} \text{ and } Q = \{\bar{x} \in A \mid x_k \geq c\},$$

or

$$P = \{\bar{x} \in A \mid x_k \leq c\} \text{ and } Q = \{\bar{x} \in A \mid x_k > c\}.$$

We shall now show that P and Q are indeed *intervals*, with $vA = vP + vQ$.

Theorem 1. *If an interval $A \subset E^n$ with endpoints $\bar{a}$ and $\bar{b}$ is split by a hyperplane $x_k = c$ ($a_k < c < b_k$), then the partition sets P and Q (as above) are intervals, and one of them is closed if A is. In particular, if $c = \frac{1}{2}(a_k + b_k)$ (the plane bisects the k-th edge), then the k-th edgelength of P and Q equals $\frac{1}{2}\ell_k = \frac{1}{2}(b_k - a_k)$; the other edgelengths equal those of A.*

Moreover, the volume of A is the sum of vP and vQ: $vA = vP + vQ$.

Proof. To fix ideas, let A be half-open, i.e., $A = (\bar{a}, \bar{b}]$; let $a_1 < c < b_1$ (i.e., we cut the *first* edge), and let

$$P = \{\bar{x} \in A \mid x_1 \leq c\},$$
$$Q = \{\bar{x} \in A \mid x_1 > c\}$$

(i.e., we include the cross section $x_1 = c$ *in* P). Consider the points

$$\bar{p} = (c, a_2, a_3, \ldots, a_n) \text{ and}$$
$$\bar{q} = (c, b_2, b_3, \ldots, b_n)$$

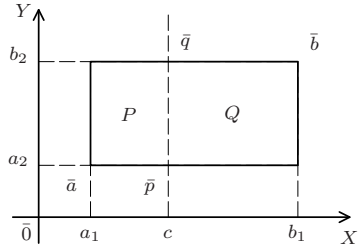


Figure 19

(see Figure 19), so that $p_1 = q_1 = c$, while $p_k = a_k$ and $q_k = b_k$ for $k \geq 2$. To prove that P is an interval, we show that $P = (\bar{a}, \bar{q}]$.

Indeed, if some $\bar{x}$ is in P, then, by definition, $\bar{x} \in A$ and $a_1 < x_1 \leq c = q_1$, and $a_k < x_k \leq b_k = q_k$, $k = 2, \ldots, n$. Thus $a_k < x_k \leq q_k$ for *all* k, i.e., $x \in (a, q]$. Reversing steps, we also see that $\bar{x} \in (a, q]$ implies $\bar{x} \in P$. Thus $P \subseteq (\bar{a}, \bar{q}] \subseteq P$, i.e., $P = (\bar{a}, \bar{q}]$. Quite similarly it is shown that $Q = (\bar{p}, \bar{b}]$. Thus P and Q are indeed intervals. It is clear that if A is *closed*, i.e., $A = [\bar{a}, \bar{b}]$, the same proof yields $P = [\bar{a}, \bar{q}]$ (so P is closed!). This proves the first part of the theorem.

Next, we compute the edgelengths of P and Q. For $k \geq 2$, we have $q_k = b_k$ and $p_k = a_k$. Thus the edgelengths of $P = (\bar{a}, \bar{q}]$ are $q_k - a_k = b_k - a_k$, i.e., *the same as those of* A (for $k \geq 2$); similarly for Q. On the other hand, the *first* edgelength of P is $q_1 - a_1 = c - a_1$ and that of Q is $b_1 - p_1 = b_1 - c$. If $c = \frac{1}{2}(a_1 + b_1)$, both expressions simplify to $\frac{1}{2}(b_1 - a_1)$. This proves the second part of the theorem.

Finally, the formula $vA = vP + vQ$ is proved by computing vA and vQ; we leave the details to the reader. Thus the theorem is proved. □

Note that, by including the cross section $x_1 = c$ in Q (instead of P), we could make Q closed (if A itself is). Thus *the choice is ours*; but we cannot make *both* P and Q closed. (Why?) Also note that, by what was shown above, a *half-open* interval $(a, b]$ can be split into two *half-open* intervals P and Q; similarly for half-closed intervals.

Next, we consider partitions into *more than two* subintervals. One important case is where we draw n hyperplanes, each *bisecting* one of the edges of an interval A and perpendicular to the corresponding axis. The first hyperplane bisects the first edge, leaving the others unchanged (as was shown in Theorem 1). The resulting two subintervals P and Q then are *both* cut (each into two parts) by the second hyperplane, which bisects the second edge in A, P, and Q. Thus, we get four disjoint intervals (see Figure 20 for E^2). The third hyperplane bisects the third edge in each of them. This yields eight subintervals. Thus each successive hyperplane *doubles* the number of the subintervals. After all n steps, we thus obtain 2^n intervals, with *all* edges bisected, so that every edgelength in each of the 2^n subintervals equals $\frac{1}{2}$ of the corresponding edgelength of A. Moreover, if A is closed then, as previously noted, we can make any one of them (but *only one*) closed, by properly manipulating the cross sections at each of the n steps. This argument yields the following result.

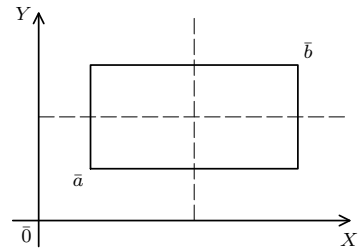


Figure 20

Theorem 2. *By drawing n hyperplanes bisecting the edges of an interval $A \subset E^n$, one can split A into 2^n disjoint subintervals whose edgelengths equal one half of the corresponding edgelengths of A and whose diagonals equal $\frac{1}{2}dA$. Any one (but only one) of the subintervals can be made closed if A is closed.*

Indeed, all this was proved except the statement about the diagonals. But if $\bar{a}$ and $\bar{b}$ are the endpoints of A, then clearly

$$dA = |\bar{b} - \bar{a}| = \sqrt{\sum_{k=1}^{n} (b_k - a_k)^2} = \sqrt{\sum_{k=1}^{n} \ell_k^2}.$$

Since the edgelengths of the subintervals are $\frac{1}{2}\ell_k$, their diagonals, by the same formula, equal

$$\sqrt{\sum_{k=1}^{n} \frac{1}{4}\ell_k^2} = \frac{1}{2}\sqrt{\sum_{k=1}^{n} \ell_k^2} = \frac{1}{2}dA,$$

as claimed.

Our next theorem states an important property of the volume, called its *additivity*. It generalizes the last clause of Theorem 1.

Theorem 3. *If an interval $A \subset E^n$ is split, in any manner, into m mutually disjoint subintervals $A_1, A_2, \ldots, A_m$, then*

$$vA = \sum_{i=1}^{m} vA_i.$$

Briefly, "*the volume of the whole equals the sum of the volumes of the parts*."

Proof. The case $m = 2$ was proved in Theorem 1.

Now, using induction, suppose additivity holds for any number of subintervals *less* than a certain m ($m > 1$). We must show that it also holds for m subintervals. To begin, let

$$A = \bigcup_{i=1}^{m} A_i \quad (A_i \text{ disjoint}).$$

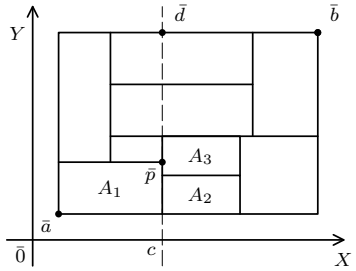


FIGURE 21

As $m > 1$, one of the A_i (say, $A_1 = [\bar{a}, \bar{p}]$) must have some edgelength *less* than the corresponding edgelength of A (say, ℓ_1). Now cut all of A into $P = [\bar{a}, \bar{d}]$ and $Q = A - P$ by the hyperplane $x_1 = c$ ($c = p_1$) (to fix ideas, we assume A and A_1 *closed*, but the proof works also in all other cases). Then (see Figure 21) $A_1 \subseteq P$ *while* $A_2 \subseteq Q$. For simplicity, we also assume that the hyperplane cuts *each* A_i into two subintervals A_i' and A_i'' (one of which may be empty); so

$$P = \bigcup_{i=1}^{m} A_i', \quad Q = \bigcup_{i=1}^{m} A_i''.$$

Actually, however, P and Q are split into *less* than m (nonvoid) intervals, since $A_1'' = \emptyset = A_2'$ by construction. Thus, by our inductive hypothesis,

$$vP = \sum_{i=1}^{m} A_i' \text{ and } vQ = \sum_{i=1}^{m} vA_i''$$

(where $vA_1'' = 0 = vA_2'$). Also, by Theorem 1, $vA = vP + vQ$ and $vA_i = vA_i' + vA_i''$. Thus

$$vA = vP + vQ = \sum_{i=1}^{m} vA_i' + \sum_{i=1}^{m} vA_i'' = \sum_{i=1}^{m} (vA_i' + vA_i'') = \sum_{i=1}^{m} vA_i,$$

and the inductive proof is complete. □

Note 2. The theorem and its proof remain valid also if some of the A_i contain *common* faces but it fails if the A_i overlap beyond that (i.e., have some *internal* points in common). As special cases, we obtain the additivity of *areas* of intervals in E^2 and *lengths* of intervals in E^1.

The proofs of the following corollaries are left to the reader.

Corollary 1. *The distance between any two points of an interval $A \subset E^n$ never exceeds the diagonal of A. Moreover, dA is the supremum of all such distances (provided $A \neq \emptyset$).*

(Hint for the second clause: If $\bar{a} \neq \bar{b}$ are the endpoints of A, consider the *line segment* $L(\bar{a}, \bar{b})$ whose length is $|\bar{b} - \bar{a}| = dA$. Show that $L(\bar{a}, \bar{b}) \subseteq A$. Given $0 < \epsilon < \frac{1}{2}dA$, show that $L(\bar{a}, \bar{b})$ contains two points $\bar{x}, \bar{y}$ such that $\rho(\bar{x}, \bar{y}) = |\bar{x} - \bar{y}| > dA - \epsilon$; e.g., take $x = \bar{a} + \frac{1}{2}\epsilon\vec{u}$ and $\bar{y} = \bar{b} - \frac{1}{2}\epsilon\vec{u}$, where

$$\vec{u} = \frac{\bar{b} - \bar{a}}{|\bar{b} - \bar{a}|}.$$

Then apply Corollary 1 and Note 4 of §9 in Chapter 2.)

Corollary 2. *Every interval $A \subset E^n$ contains all line segments $L[\bar{p}, \bar{q}]$ whose endpoints $\bar{p}$ and $\bar{q}$ lie in A.*

(This property is called *convexity*. Thus all intervals are *convex* sets. See also Problem 7 of §4.)

Corollary 3. *The volume, the edgelengths, and the diagonal of a subinterval never exceed those of the containing interval.*

Corollary 4. *Every nondegenerate interval in E^n contains rational points, i.e., points whose coordinates are rational.*

(Hint: Apply the density of rationals in E^1 for each coordinate separately.)

Problems on Intervals in E^n

1. Complete the missing details in the proof of Theorem 1. In particular, show that $Q = (\bar{p}, \bar{b}]$ and that $vA = vP + vQ$. Then, assuming that A is closed, modify the proof so as to make Q closed.

2. Prove Corollaries 1 through 4.

2′. Verify Note 2.

3. Give a suitable definition of a "face" of an interval $A \subset E^n$ and of its 2^n "vertices" (the endpoints are only two of them).

4. Compute the edgelengths, the diagonal, and the volume of $[\bar{a}, \bar{b}]$ in E^4, given that $\bar{a} = (1, -2, 4, 0)$ and $\bar{b} = (2, 0, 5, 3)$. Is it a cube? Find all its "vertices" (see Problem 3). Split it by the plane $x_4 = 1$ and verify Theorem 1 (last part) by actually computing the volumes involved.

5. Verify that the cross product of n line intervals (a_k, b_k), $k = 1, \dots, n$, coincides with the set $\{\bar{x} \in E^n \mid a_k < x_k < b_k\}$. (Thus justify the second part of Definition 1.) Show also that Definition 1 could be stated *inductively*: An interval in E^n is the cross product of an interval in E^{n-1} by a line interval. (Use the inductive definition of an n-tuple, given in §6 of Chapter 2.)

***6.** A nonempty family of (arbitrary) sets is called a *semi-ring* of sets iff

(i) it contains the intersection of any two (hence any finite number) of its members; that is, if A and B are members of the family, so is $A \cap B$; and

(ii) the difference $A - B$ of any two members can always be represented as a union of a finite number of disjoint members of the family; i.e., $A - B = \bigcup_{i=1}^{m} C_i$ for some disjoint sets C_i belonging to the family.

Given this definition, solve the following problems:

(a) Prove that all intervals in E^1 satisfy (i) and (ii) and hence constitute a semi-ring; show that so also do the *half-open* intervals in E^1 alone; similarly for the *half-closed* intervals. Disprove this for *open* intervals and for *closed* intervals.
[Hint: (ii) fails.]

(b) Do question (a) for intervals *in* E^n; in particular, show that all half-open intervals in E^n form a semi-ring.

[Hint: Use the inductive definition given at the end of Problem 5, and apply induction on the number n of dimensions; i.e., assuming all for E^{n-1}, prove it for E^n.]

***7.** A set in E^n is said to be *simple* iff it is the union of a finite number of disjoint intervals (in particular, all intervals are simple). Prove the following:

(a) *If A and B are simple, so is $A \cap B$.*
[Hint: Let

$$A = \bigcup_{i=1}^{m} A_i, \quad B = \bigcup_{k=1}^{r} B_k.$$

Then

$$A \cap B = \bigcup_{i=1}^{m} \bigcup_{k=1}^{r} (A_i \cap B_k). \quad \text{(Verify!)}$$

If A_i and B_k are intervals, so are all $A_i \cap B_k$ by Problem 6 (since the intervals form a *semi-ring*). The sets $A_i \cap B_k$ are disjoint if so are A_i or B_k. Thus $A \cap B$ is a finite union of disjoint intervals, i.e., $A \cap B$ is simple.]

Extend this, by induction, to intersections of any finite number of simple sets: *If $A_1, A_2, \dots, A_r$ are simple, so is $\bigcap_{k=1}^{r} A_k$.*

(b) *If A is simple and B is an interval, then $A - B$ is simple.*
[Hint: Let $A = \bigcup_{i=1}^{m} A_i$, where the A_i are disjoint intervals. Then

$$A - B = \bigcup_{i=1}^{m} (A_i - B). \quad \text{(Verify!)}$$

By Problem 6, $A_i - B$ is the union of some disjoint intervals $C_1, C_2, \dots, C_{n_i}$. Thus

$$A - B = \bigcup_{i=1}^{m} \bigcup_{k=1}^{n_i} C_k,$$

with *all* C_k disjoint. (Why?)]

(c) *If A and B are simple, so is $A - B$.*
[Hint: Let $B = \bigcup_{i=1}^{m} B_i$ for some disjoint intervals B_i. Then

$$A - B = A - \bigcup_{i=1}^{m} B_i = \bigcap_{i=1}^{m} (A - B_i),$$

by duality laws. By (b), each $A - B_i$ is simple, and so is

$$\bigcap_{i=1}^{m} (A - B_i)$$

by (a).]

(d) *If A and B are simple, so is $A \cup B$ (similarly for all finite unions, by induction).*
[Hint: $A \cup B = (B - A) \cup A$; A is a disjoint union of intervals (by assumption); so is $B - A$, by (c); hence, so is $A \cup B$.]

***8.** A nonempty family $\mathcal{M}$ of (arbitrary) sets is called a *ring of sets* iff

$$(\forall A, B \in \mathcal{M}) \quad A - B \in \mathcal{M} \text{ and } A \cup B \in \mathcal{M}.$$

(We then also say that $\mathcal{M}$ is *closed under finite unions and differences*.) Infer from Problem 7 that all simple sets in E^n form a *ring*. Moreover, show that if $\mathcal{C}$ is a semi-ring of sets (cf. Problem 6), then all finite unions of disjoint members of $\mathcal{C}$ form a ring.

[Hint: Proceed as in Problem 7.]

***9.** Prove the *subadditivity of the volume* for intervals A, B_1, B_2, ..., B_m (not necessarily disjoint): *If* $A = \bigcup_{i=1}^{m} B_i$, *then*

$$vA \le \sum_{i=1}^{m} vB_i.$$

[Hint: Let $C_1 = B_1$ and $C_k = B_k - \bigcup_{i=1}^{k-1} B_i$, $k = 2, 3, \ldots, m$. Verify that the sets C_k are *disjoint* and that $A = \bigcup_{k=1}^{m} C_k$, with $C_k \subseteq B_k$. From Problem 7(d)(c), infer that each C_k is *simple*, and so is each $B_k - C_k$. Thus C_k is the union of some disjoint intervals D_{kj}, $j = 1, \ldots, m_k$, while B_k contains some *additional* intervals (those in $B_k - C_k$). Now, use additivity (Theorem 3) to obtain

$$\sum_{j=1}^{m_k} vD_{kj} \le vB_k$$

and, from $A = \bigcup_{k=1}^{m} C_k$,

$$vA = \sum_{k=1}^{m} \sum_{j=1}^{m_k} vD_{kj} \le \sum_{k=1}^{m} vB_k,$$

as required.]

§8. Complex Numbers

As we have already noted, E^n *is not a field*, because of the lack of a vector multiplication that would satisfy the field axioms. Now we shall define such a multiplication, but only for E^2. Thus E^2 will become a field which we shall call the *complex field*, denoted C.

In this connection, it will be convenient to introduce some notational and terminological changes. Points of E^2, when regarded as elements of the field C, will be called *complex numbers* (each being an ordered pair of real numbers). We shall denote them by lower case letters (preferably z), *without* a bar or an arrow; e.g., $z = (x, y)$ denotes a complex number with coordinates x and y. We shall preferably write (x, y) instead of (x_1, x_2). The coordinates x and y of z are also called the *real* and *imaginary* parts of z, respectively.

If $z = (x, y)$, then $\overline{z}$ will denote the complex number $(x, -y)$, called the *conjugate* of z. Thus $\overline{z}$ has the same real part as z, but its imaginary part is the additive inverse of that of z. Geometrically, the point $\overline{z}$ is symmetric to z with respect to the x-axis (see Figure 22).

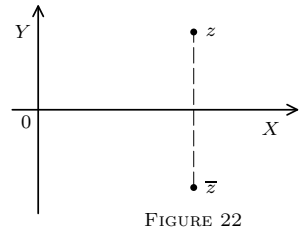


FIGURE 22

Complex numbers of the form $(x, 0)$, i.e., those with vanishing imaginary part, are called *real points* of C. For brevity, we shall simply write x for $(x, 0)$; e.g., $2 = (2, 0)$. In particular, we write 1 for $\bar{e}_1 = (1, 0)$ and call it the *real unit* in C. Points of the form $(0, y)$, with vanishing real part, are called (*purely*) *imaginary numbers*. In particular, the unit vector $\bar{e}_2$ is such a number since $\bar{e}_2 = (0, 1)$; we shall now denote it by i and call it the *imaginary unit* in C. Apart from these notational and terminological peculiarities, all our former definitions that were given for E^n remain valid in $E^2 = C$. In particular, this applies to the definition of the sum and difference,

$$(x, y) \pm (x', y') = (x \pm x', y \pm y'),$$

and that of the absolute value: If $z = (x, y)$, then $|z| = \sqrt{x^2 + y^2}$. Similarly, if $z = (x, y)$ and $z(x', y')$, then $\rho(z, z') = \sqrt{(x - x')^2 + (y - y')^2}$. Hence, also, all previous theorems remain valid.

We now define the new multiplication in C. The definition may seem strange at first sight, but it makes a field out of E^2, as will be seen.

Definition 1.

The *product* of two complex numbers $(x, y) = z$ and $(x', y') = z'$ is the complex number $(xx' - yy', xy' + yx')$, denoted $(x, y)(x', y')$ or zz'.

Theorem 1. *$E^2 = C$ is a field under addition and multiplication as defined above, with the zero element $0 = (0, 0)$ and unity $1 = (1, 0)$.*

Proof. We only must show that multiplication obeys the field axioms I–VI (as for addition, all is proved in Theorem 1 of §1).

Axiom I (closure law) is obvious from Definition 1: if z, z' are in C, so is zz'. To prove commutativity, we take any two complex numbers, $z = (x, y)$ and $z' = (x', y')$, and verify that $zz' = z'z$. Indeed, by definition,

$$zz' = (xx' - yy', xy' + yx'), \text{ while } z'z = (x'x - y'y, x'y + y'x);$$

but the bracketed expressions coincide, by the commutative laws for *real* numbers. Thus, indeed, $zz' = z'z$. Associativity and distributivity are proved in a similar manner, and we leave it to the reader.

Next, we show that $1 = (1, 0)$ is the "unity" element required in Axiom IV(b), i.e., that for any number $z = (x, y) \in C$, we have $1z = z$. In

fact, by Definition 1,

$$1z = (1,0)(x,y) = (1x - 0y, 1y + 0x) = (x - 0, y + 0) = (x,y) = z$$

(here we have used the corresponding laws for *reals*).

It remains to establish Axiom V(b), i.e., to show that every complex number $z = (x,y) \neq (0,0)$ has a multiplicative *inverse* z^{-1} such that $zz^{-1} = 1$. It turns out that this inverse is obtained by setting

$$z^{-1} = \Big(\frac{x}{|z|^2}, \frac{-y}{|z|^2}\Big),$$

where $|z|^2 = x^2 + y^2$. In fact, with z^{-1} so defined, we have

$$\begin{aligned} zz^{-1} &= (x,y)\Big(\frac{x}{|z|^2}, \frac{-y}{|z|^2}\Big) = \Big(\frac{x^2+y^2}{|z|^2}, \frac{-xy+yx}{|z|^2}\Big) \\ &= \Big(\frac{x^2+y^2}{|z|^2}, 0\Big) = (1,0) = 1. \end{aligned}$$

Thus, indeed, $zz^{-1} = 1$, as required, and all is proved. □

We now obtain some immediate corollaries.

Corollary 1. $i^2 = -1$. *In fact, by definition,*

$$i^2 = (0,1)(0,1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1,0) = -1.$$

Thus the complex field C *has an element* $i = (0,1)$ *whose square is* $-1 = (-1,0)$, whereas there is no such element in E^1, by Corollary 3 in §4 of Chapter 2. This is not a contradiction since that corollary was proved only for *ordered* fields (it is based on Axioms VII–IX). This only shows that C cannot be *ordered*, so as to satisfy Axioms VII–IX. Thus we shall define no inequalities ($<$) in C.

From our definitions one easily obtains the following equations for "real points" $(x,0)$ and $(x',0)$:

$$(x,0) + (x',0) = (x + x', 0) \text{ and } (x,0) \cdot (x',0) = (xx', 0).$$

(Verify!) Thus *two "real points" in* C *are added* (*multiplied*) *by simply adding* (*multiplying*) *their real parts*, x and x', while the imaginary part, i.e., 0, remains unchanged, as an "onlooker" only. Similarly for subtraction and division. In other words, when carrying out field operations on "real points" in C, we may safely forget about the distinction between the real number x ($x \in E^1$) and the real point $(x,0)$ in C. The real points in C behave exactly like real numbers. One easily verifies that they form a field (called the *real subfield of* C), and we may even order them exactly as we order their real parts, i.e., by setting

$$(x,0) < (x',0) \iff x < x'.$$

Then the real points in C become an *ordered field* that, mathematically, is an exact copy of E^1. Geometrically, it is the x-axis in the xy-plane representing C.

(*More precisely, one can describe this situation by using the notion of *isomorphism* defined in §14 of Chapter 2. The mapping $x \to (x,0)$ is an isomorphism of E^1 onto the real subfield of C, since it preserves addition, multiplication, and order. (Verify!))

Therefore it is customary not to distinguish between real numbers and real points in C, "identifying" x with $(x,0)$ in C, as was explained above. With this convention, E^1 *becomes simply a subset* (*and a subfield*) *of* C. Henceforth, we shall simply say that "x is real" or "$x \in E^1$," instead of saying that "$x = (x,0)$ is a *real point* in C." We then also obtain the following result.

Theorem 2. *Every complex number* z *has a unique representation as a sum:* $z = x + yi$, *where* x *and* y *are real and* $i = (0,1)$ *is the imaginary unit.*

Proof. By our convention, x and y stand for $(x,0)$ and $(y,0)$, respectively; thus $x + yi = (x,0) + (y,0) \cdot (0,1)$. Computing the right side expression from definitions, we obtain for any $x, y \in E^1$

$$x + yi = (x,0) + (y \cdot 0 - 0 \cdot 1, y \cdot 1 + 0 \cdot 0) = (x,0) + (0,y) = (x,y).$$

Thus $(x,y) = x + yi$ *for any* $x, y \in E^1$. If, in particular, we take the coordinates of z for x and y in that formula, we obtain $z = (x,y) = x + yi$, which is the required representation.

To prove its uniqueness, suppose that we also have $= x' + y'i$, where $x' = (x',0)$ and $y' = (y',0)$. But then, as was shown above,

$$z = (x',0) + (y',0) \cdot (0,1) = (x',y'),$$

and so $z = (x',y')$. Since also $z = (x,y)$, we have $(x,y) = (x',y')$, i.e., the pairs (x,y) and (x',y') are the same, and so $x = x'$, $y = y'$ after all. Thus the theorem is proved. □

We shall now consider the geometric representation of complex numbers as points of the Cartesian plane (see Figure 23). The x-axis comprises all the "real points"; the y-axis consists of all "imaginary" points". The rest of the plane represents all the other complex numbers. Instead of the Cartesian coordinates (x,y), we may also use *polar coordinates* (r,θ), where $r = \sqrt{x^2+y^2}$ is the absolute value $|z|$ of $z = (x,y)$ and θ is the (counterclockwise) rotation angle from the

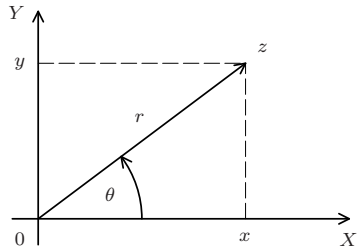


FIGURE 23

x-axis to $\overline{0z}$ (represented as the directed line segment $\overline{0z}$). Clearly, z is uniquely determined by r and θ, but θ is not uniquely determined by z; indeed, the same point of the plane results if θ is replaced by $\theta + 2n\pi$ ($n = 1, 2, \dots$); r and θ are called, respectively, the *modulus* and *argument* of $z = (x, y)$. By elementary trigonometry, we have $x = r\cos\theta$ and $y = r\sin\theta$. Substituting this in $z = x + yi$ (see Theorem 2), we obtain the following corollary.

Corollary 2. $z = r(\cos\theta + i\sin\theta)$ ("*trigonometric form of* z").

In conclusion, we note that since C is a field, all consequences of the field Axioms I–VI (but not VII–IX) apply to it. Quotients and differences are defined as in §3 of Chapter 2, and all propositions proved there for (unordered) fields apply to C.

Problems on Complex Numbers

1. Complete the proof of Theorem 1 (associativity, distributivity, etc.).

1′. Verify that the "real points" in C form an *ordered field*.

2. Prove that $z\overline{z} = |z|^2$. Infer that if $z \neq 0$, then $z^{-1} = \overline{z}/|z|^2$.

3. Show that the conjugate of the sum (product) of z and z' in C equals the sum (product) of their conjugates:

$$\overline{z + z'} = \overline{z} + \overline{z'}, \quad \overline{zz'} = \overline{z} \cdot \overline{z'}.$$

Show also (by induction) that

$$\overline{z^n} = (\overline{z})^n \quad (n = 1, 2, \dots)$$

and that

$$\overline{\sum_{k=1}^{n} a_k z^k} = \sum_{k=1}^{n} \overline{a}_k \overline{z}^k.$$

***4.** From Problem 3 infer that the map $z \to \overline{z}$ is an isomorphism of C *onto itself* (such isomorphisms are called *automorphisms*).

5. Compute

(a) $(1 + 2i)(3 - i)$;

(b) $(1 + 2i)/(3 - i)$;

(c) i^n, $n = 1, 2, \dots$;

(d) $(1 \pm i)^n$;

(e) $1/(1 + i)^n$;

(f) $(x + 1 + i)/(x + 1 - i)$, $x \in E^1$;

(g) $(z + 1 + i)(z + i - i)(z - 1 + i)(z - 1 - i)$.

Do (a), (b), and (f) in *two* ways:

1) Use definitions only, and use the notation (x, y) instead of $x + yi$.

2) Use all laws valid in a field. In fractions, multiply the numerator and the denominator by the conjugate of the denominator to get a *real* denominator.

6. Solve the equation $(2, -1)(x, y) = (3, 2)$ for x and y.

7. Use Corollary 2 to show that if

$$z' = r'(\cos\theta' + i\sin\theta') \text{ and } z'' = r''(\cos\theta'' + i\sin\theta''),$$

then the modulus r of the product $z = z'z''$ equals $r'r''$, i.e., $|z| = |z'|\,|z''|$, and the argument θ of z equals $\theta' + \theta''$. Hence, derive the geometric interpretation of the product: to multiply two complex numbers z' and z'' means to multiply the vector $\overrightarrow{0z'}$ by the scalar $|z''|$ and rotate it counterclockwise around $\overline{0}$ by the angle θ''. Consider the cases $z'' = i$ and $z'' = -1$.

[Hint: Expand

$$r'(\cos\theta' + i\sin\theta') \cdot r''(\cos\theta'' + i\sin\theta'')$$

and apply the laws of trigonometry.]

8. Use induction to extend the result of Problem 7 to products of n complex numbers. Also derive *de Moivre's formula*: If $= r(\cos\theta + i\sin\theta)$, then

$$z^n = r^n(\cos n\theta + i\sin n\theta).$$

Using it, solve again 5(c), (d), and (e).

9. From Problem 8 derive that, for every complex number $z \neq 0$, there are exactly n complex numbers w such that $w^n = z$ ($n = 1, 2, \dots$); they are called the *n-th roots* of z.

[Hint: If

$$z = r(\cos\theta + i\sin\theta) \text{ and } w = r'(\cos\theta' + i\sin\theta'),$$

the equation $w^n = z$ implies $(r')^n = r$ and $n\theta' = \theta$, and conversely, so that $r' = \sqrt[n]{r}$ and $\theta' = \theta/n$. While r' is thus determined *uniquely*, there are different choices of θ', since θ may be replaced by $\theta + 2k\pi$ without affecting z. Thus,

$$\theta' = \frac{\theta + 2k\pi}{n}, \quad k = 1, 2, \dots .$$

Distinct points w result only for $k = 0, 1, \dots, n - 1$ (after which they repeat cyclically).]

*§9. Vector Spaces. The Space C^n. Euclidean Spaces

I. We have occasionally mentioned that there are vector spaces other than E^n. Now we shall dwell on this matter in more detail.

Let V be an arbitrary set whose elements will be called "points" or "vectors" (even though they may have nothing in common with E^1 or E^n). Suppose that a certain binary operation (call it "addition") has somehow been defined in V *in such a manner that the first five axioms for real numbers hold for this* "*addition*". That is, we have the closure law, $(\forall x,\ y \in V)\ x+y \in V$, commutativity, and associativity; there is a (unique) zero-element, denoted $\vec{0}$, such that $(\forall x \in V)\ x+\vec{0}=x$; and each vector $x \in V$ has a (unique) additive inverse $-x$, such that $x+(-x)=\vec{0}$. A set V together with such an operation is called an *Abelian* or *commutative group*.

Note. If commutativity is not assumed, V is simply called a *group*. In this section, however, only commutative groups will be considered. Note that the operation $(+)$ need not be the ordinary addition, and sometimes other symbols are used instead of "+". For an example of a *noncommutative* group, see Problem 8 in §6 of Chapter 1.

Next, let F be any field (e.g., E^1 or C); its elements will be called *scalars*; its zero-element will be denoted by 0, and its unity by 1. Suppose that yet another operation (call it "multiplication of scalars by vectors") has been defined that assigns to every scalar $a \in F$ and every vector $x \in V$ a certain vector $ax \in V$, called the *a-multiple* of x, and suppose that it satisfies the following laws: $(\forall a, b \in F)\ (\forall x, y \in V)$

$$a(x+y)=ax+ay,\ \ (a+b)x=ax+bx,\ \ (ab)x=a(bx),\ \text{ and } 1x=x.$$

In other words, we assume that *all laws of Theorem 1 of §1 are valid*. In this case, V together with these two operations is called a *vector space*, or a *linear space*, *over the field* F; F is called its *field of scalars* or *scalar field*.

Examples.

(a) E^n is a vector space over E^1 (its scalar field), with operations as defined in §1. So also is R^n, the set of all points with rational coordinates, i.e., ordered n-tuples $(x_1, \ldots, x_n)$ of *rationals*; but its field of scalars is R, not E^1.

We also could choose R as the field of scalars *for all of* E^n. This would yield a *different* vector spaces: E^n *over* R, not over E^1. It contains R^n as a *subspace* (a smaller space over the same field).

(b) Let F be any field, and let F^n be the set of all n-tuples of elements of F $(x_1, x_2, \ldots, x_n)$, $x_k \in F$, with sums and scalar multiples defined exactly as for E^n (with F playing the role of E^1). Then F^n is a vector space *over* F. (The proof is exactly as in Theorem 1 of §1.)

(c) *Every field* F *is also a vector space* under the addition and multiplication defined in F, with F treated as its own field of scalars. (Verify!)

(d) Let V be a vector space over a field F, and let W be the set of all mappings $f: A \to V$ from some arbitrary set $A \neq \emptyset$ into V. Define the sum of two such maps f and g, denoted $f+g$, by setting

$$(f+g)(x)=f(x)+g(x) \text{ for all } x \in A.$$

(Here "$(f+g)$" is to be treated as *one* letter (function symbol). Thus, "$(f+g)(x)$" means "$h(x)$" where $h=f+g$.) Similarly, given $a \in F$ and $f \in W$, we define the map $(af): A \to V$ by

$$(af)(x)=af(x).$$

Then, under these operations, W *is a vector space over the same field* F. (Verify!) In particular, taking $V=E^1$ or $V=C$, we obtain *the vector space of all real-valued functions* $f: A \to E^1$ (with $F=E^1$) or that *of all complex-valued functions* $f: A \to C$ (with $F=C$ or $F=E^1$).

In every vector space V over a field F we can define *linear combinations* of vectors, i.e., sums of the form

$$\sum_{k=1}^{m} a_k x_k \quad (a_k \in F,\ x_k \in V),$$

hence also *linearly dependent* and *independent sets of vectors* (cf. §1, Problem 8). Moreover, given two vector spaces V and W over the same field F, we can consider *linear maps* $f: V \to W$, i.e., mappings which preserve linear combinations, so that

$$(\forall x, y \in V)\ (\forall a, b \in F) \quad f(ax+by)=af(x)+bf(y)$$

(cf. §5, Definition 2). Such a map is called a *linear functional* (on V) if the range space W is simply the scalar field F of V, so that $f: V \to F$. (Recall that a field F may be treated as a vector space.)

Vector spaces over E^1 (respectively, C) are called *real* (respectively, *complex*) vector spaces. Complex spaces can always be transformed into real ones by restricting their scalar field C to its real subfield (which we identify with E^1).

II. An important example of a complex linear space is C^n, i.e., the set of all n-tuples $x=(x_1, \ldots, x_n)$ of *complex* numbers x_k (now treated as *scalars*), with sums and scalar multiples defined as in E^n. In order to avoid confusion with conjugates of complex numbers, we shall not use the notation $\bar{x}$ for a vector in C^n, writing simply x for it. Dot products in C^n are defined by

$$x \cdot y=\sum_{k=1}^{n} x_k \overline{y}_k,$$

where $\overline{y}_k$ is the *conjugate* of the complex number y_k (cf. §8). Note that if $y_k \in E^1$, then $\overline{y}_k = y_k$. Thus, for points with real coordinates,

$$x \cdot y = \sum_{k=1}^{n} x_k y_k,$$

in agreement with our definition of $x \cdot y$ in E^n.

The reader will easily verify (exactly as for E^n) that for $x, y \in C^n$, we have the following:

(i) $x \cdot y \in C$; thus $x \cdot y$ is a *scalar*, not a vector.

(ii) $x \cdot x \in E^1$ *and* $x \cdot x \geq 0$; i.e., the dot product of a vector by *itself* is a *real* number ≥ 0. Moreover, $x \cdot x = 0$ *iff* $x = \vec{0}$.

(iii) $x \cdot y = \overline{y \cdot x}$ (= conjugate of $y \cdot x$). Thus commutativity *fails* in general.

(iv) $(\forall a, b \in C)\ (ax) \cdot (by) = (a\overline{b})\,(x \cdot y)$; hence

(iv′) $(ax) \cdot y = a(x \cdot y) = x \cdot (\overline{a}y)$.

(v) $(x + y) \cdot z = x \cdot z + y \cdot z$ and

(v′) $z \cdot (x + y) = z \cdot x + z \cdot y$ (distributive laws).

Observe that (v′) follows from (v) by using (iii). Verify!

III. Sometimes (but not always) dot products can also be defined on complex or real linear spaces other than C^n or E^n *in such a manner that they satisfy the laws* (i)–(v) *listed above* (with C replaced by E^1 if the space is real). If these laws hold, the space is called a (complex or real) *Euclidean space*.[1] In particular, C^n is a complex Euclidean space, and E^n is a real Euclidean space.

In every Euclidean space (real or complex), one can define *absolute values* of vectors by setting $|x| = \sqrt{x \cdot x}$ (this root exists in E^1 since $x \cdot x \geq 0$ by formula (ii) above). In particular, this definition applies to C^n and E^n (cf. §2, Note 3). Then, similarly as was done for E^n, one obtains the following laws, valid for all vectors x, y and any scalar a:

(a′) $|x| \geq 0$; and $|x| = 0$ iff $x = \vec{0}$;

(b′) $|ax| = |a|\,|x|$;

(c′) $|x + y| \leq |x| + |y|$ (triangle inequality);

(d′) $|x \cdot y| \leq |x|\,|y|$ (Cauchy–Schwarz inequality).

In particular, these laws are valid in C^n and E^n.

The proof is analogous to that of Theorem 2 of §2. Only the Cauchy–Schwarz inequality requires a somewhat different approach, as follows.

[1] Note that the scalar field in a Euclidean space is always C or E^1. The same applies to *normed* linear spaces, to be defined later.

If $|x \cdot y| = 0$, there is nothing to prove. Thus let $x \cdot y \neq 0$, and put

$$a = \frac{x \cdot y}{|x \cdot y|} \neq 0.$$

Let t be an *arbitrary real* number, $t \in E^1$, and consider the expression $(tx + ay) \cdot (tx + ay) \geq 0$ (see formula (ii) above). Removing brackets (by distributivity) and using (iii) and (iv), we obtain

$$\begin{aligned} 0 &\leq (tx + ay) \cdot (tx + ay) \\ &= tx \cdot tx + ay \cdot tx + tx \cdot ay + ay \cdot ay \\ &= t^2|x|^2 + (a\overline{t})\,(y \cdot x) + (t\overline{a})\,(x \cdot y) + |a|^2|y|^2 \quad (\text{for } a\overline{a} = |a|^2 \text{ in } C). \end{aligned}$$

As $t \in E^1$, we have $\overline{t} = t$. Also, as

$$a = \frac{x \cdot y}{|x \cdot y|},\ \text{we have } \overline{a} = \frac{\overline{x \cdot y}}{|x \cdot y|}.$$

Thus

$$(t\overline{a})\,(x \cdot y) = t\,\frac{\overline{x \cdot y}}{|x \cdot y|}\,(x \cdot y) = \frac{t}{|x \cdot y|}\,|x \cdot y|^2 = t\,|x \cdot y|.$$

Similarly,

$$(at)\,(y \cdot x) = t|x \cdot y|,\ \text{and } |a|^2 = a\overline{a} = \frac{|x \cdot y|^2}{|x \cdot y|^2} = 1.$$

Substituting, we get

$$0 \leq t^2|x|^2 + 2t|x \cdot y| + |y|^2$$

for an *arbitrary* $t \in E^1$.

Here $|x|^2$, $|x \cdot y|$, and $|y|^2$ are fixed *real* numbers (by the definition of absolute value). We treat them as coefficients and t as a variable. Thus we have a quadratic trinomial in t which remains nonnegative for *all* $t \in E^1$. By elementary algebra (which we assume known) its discriminant must be ≤ 0. Thus

$$4|x \cdot y|^2 - 4|x|^2|y|^2 \leq 0,\ \text{whence } |x \cdot y| \leq |x|\,|y|. \quad \square$$

Once absolute values have been defined and laws (a′)–(d′) have been established, we can also define *distances*, as in E^n, by setting $\rho(x, y) = |x - y|$ for any vectors x and y. We treat this matter in the next section in a more general setting, so we omit it here.

Finally, in any real or complex linear space V, we define *lines* and *line segments* exactly as in E^n. That is, given two fixed points a, $b \in V$, we define the line $\overleftrightarrow{ab}$ to be the set of all points $x \in V$ which are of the form

$$x = a + t(b - a) = (1 - t)a + tb,$$

where t varies *over* E^1 (not over all of C, even if the space is complex). Line segments are obtained by letting t vary over corresponding intervals in E^1 (cf. §4).

Problems on Linear Spaces

1. Prove that F^n in Example (b) is a vector space, i.e., satisfies all laws stated in Theorem 1 of §1. Similarly for W in Example (d).

2. Verify that inner products (dot products) in C^n obey laws (i)–(v). Which of the laws would fail if these products were defined by

$$x \cdot y = \sum_{k=1}^{n} x_k y_k \text{ instead of } \sum_{k=1}^{n} x_k \overline{y}_k?$$

How would this affect the definition of absolute values? Would such values satisfy laws (a′)–(d′)?

3. Complete the proof of properties (a′)–(c′) of absolute values in a Euclidean space V. What change in (a′) would result if property (ii) of dot products were weakened to say only that $x \cdot x \geq 0$ and $\vec{0} \cdot \vec{0} = 0$?

4. Define angles, directions, and orthogonality (perpendicularity) in a general Euclidean space, following the pattern of §3. Show that a vector v is orthogonal to all vectors of the space *iff* $v = \vec{0}$.

5. Define hyperplanes in C^n following the pattern of §5 (parts I and II), and prove Theorems 1, 2, and 3 of §5 for such hyperplanes.

6. Which (if any) of the problems following §5 remain valid for hyperplanes in C^n?

7. Prove the principle of *nested line segments*: *Every contracting sequence of closed line segments $L[a_m, b_m]$, $m = 1, 2, \ldots$, in a real or complex Euclidean space V has a nonempty intersection*,

$$\bigcap_{m=1}^{\infty} L[a_m, b_m] \neq \emptyset.$$

[Hint: All the line segments $L[a_m, b_m]$ lie on the line $x = a_1 + tu$, where $u = b_1 - a_1$. (Why?) In particular,

$$a_m = a_1 + t_m u \text{ and } b_m = a_1 + t'_m u \quad \text{for some } t_m, t'_m \in E^1.$$

Show that the intervals $[t_m, t'_m]$ in E^1 form a contracting sequence, i.e.,

$$[t_m, t'_m] \supseteq [t_{m+1}, t'_{m+1}], \quad m = 1, 2, \ldots .$$

Now, from Problem 11 in §9 of Chapter 2, infer that there is

$$t_0 \in \bigcap_{m=1}^{\infty} [t_m, t'_m] \quad \text{in } E^1,$$

and let $p = a_1 + t_0 u$. Then show that $p \in \bigcap_{m=1}^{\infty} L[a_m, b_m]$.]

8. Prove Note 3 at the end of §4 for lines *in any Euclidean space*.

9. Define the basic unit vectors e_k in C^n exactly as in E^n, and show that they are *linearly independent*, i.e.,

$$\sum_{k=1}^{n} a_k e_k = \vec{0} \quad (a_k \in C)$$

iff all a_k vanish.

10. Prove that if a set of vectors $B = \{v_1, \ldots, v_m\}$ in a vector space is linearly independent, then:

(a) B does not contain $\vec{0}$;

(b) every subset of B is linearly independent;

(c) if

$$\sum_{k=1}^{m} a_k v_k = \sum_{k=1}^{m} b_k v_k$$

for scalars $a_k, b_k \in F$, then necessarily $a_k = b_k$, $k = 1, 2, \ldots, m$.

*§10. Normed Linear Spaces

In §9 we saw how absolute values can be defined from inner products in Euclidean spaces. Sometimes, however, absolute values can be defined *directly*, even in non-Euclidean linear spaces (where there are no dot products), "bypassing" inner products altogether. All that is required is to assign, in some way or other, a real absolute value $|x|$ to every vector x *in such a manner that laws* (a′)–(c′) *specified in* §9 *are satisfied* (excluding (d′) since it has no sense if there are no dot products). A vector space equipped with such absolute values is called a *normed linear space*. Thus, we have the following definition.

Definition 1.

A *normed linear space* is a real or complex vector space V in which every vector v is associated with a real number $|v|$, called its *absolute value* (or *norm* or *magnitude*), such that, for any vectors $u, v \in V$ and any scalar a (in E^1 or C, as the case may be),

(i) $|v| \geq 0$;

(i′) $|v| = 0$ *iff* $v = \vec{0}$;

(ii) $|av| = |a|\,|v|$; and

(iii) $|u + v| \leq |u| + |v|$ (triangle inequality).

Sometimes we write $\|v\|$ for $|v|$ or use other similar symbols.

Mathematically, the existence of absolute values in V amounts to the existence of a mapping $v \to |v|$ on V, i.e., a mapping $\varphi: V \to E^1$, with function values $\varphi(v)$ written as $|v|$, satisfying the laws (i)–(iii). Any such mapping is called a *norm map* (briefly, "*norm*") on V. Thus, to define absolute values in V means to define a norm map $v \to |v|$ on V, satisfying (i)–(iii). Often this can be done in many different ways, thus giving rise to *different* norms on V, all satisfying (i)–(iii).

Note 1. There also are maps $v \to |v|$ that satisfy (i), (ii), and (iii) but only a weaker form of (i′), namely, $|\vec{0}| = 0$, so that $|v|$ may vanish if $v \neq \vec{0}$. Such maps are called *semi-norms*, and vector spaces equipped with such maps are called *semi-normed linear spaces*.

Examples.

(1) Every Euclidean space (in particular, E^n and C^n) is also a normed linear space, with the norm defined by

$$|v| = \sqrt{v \cdot v}.$$

Indeed, as was shown in §9, absolute values so defined satisfy (a′)–(c′), i.e., laws (i)–(iii) of Definition 1. In E^n and C^n, one can also define $|v|$ *directly* in terms of coordinates, setting

$$|v| = \sqrt{\sum_{k=1}^{n} |v_k|^2},$$

which is equivalent to $|v| = \sqrt{v \cdot v}$. This is the so-called *standard* norm on E^n (C^n).

(2) One can also define various "nonstandard" norms on E^n and C^n; e.g., fix some real number $p \geq 1$ and put

$$\|v\| = \sqrt[p]{\sum_{k=1}^{n} |v_k|^p}.$$

It can be shown that this yields another norm map $v \to \|v\|$. (See Problems 9–11 below.)

(3) A *semi-norm* on E^n and C^n is obtained by setting

$$|v| = |v_1| \text{ where } v = (v_1, v_2, \ldots, v_n);$$

e.g., if $v = (0, 1, 1, \ldots, 1)$, then $|v| = 0$ because $v_1 = 0$. Thus formula (i′) fails here, but the remaining laws (i)–(iii) do hold, as is easily verified. Therefore, we have a semi-norm here, not a norm.

(4) Let W be the set of all *bounded* real functions on a set $A \neq \emptyset$, i.e., maps $f: A \to E^1$ such that

$$(\forall x \in A) \quad |f(x)| < c$$

for some constant c (depending on f only). Due to boundedness, the set of all absolute values $|f(x)|$, for a given $f \in W$, has a l.u.b. *in* E^1; we denote it by $\|f\|$. Thus

$$\|f\| = \sup |f(x)|, \quad x \in A.$$

We also define operations in W as in Example (d) of §9, i.e., setting for any $a \in E^1$ and any $f, g \in W$,

$$(\forall x \in A) \quad (f+g)(x) = f(x) + g(x) \text{ and } (af)(x) = a \cdot f(x).$$

Thus the maps $f + g$ and af are defined on A.

It is easy to show that these definitions make W a normed linear space, with norm $\|f\| = \sup |f(x)|$ for $f \in W$. (Here each function $f \in W$ is to be treated as a "vector" or "point" in W.) Leaving other details to the reader, we verify the triangle inequality: $\|f + g\| \leq \|f\| + \|g\|$. By definition, we have, for $f, g \in W$,

$$|(f+g)(x)| = |f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\|. \tag{4′}$$

(The last inequality holds because $\|f\| = \sup |f(x)|$ and $\|g\| = \sup |g(x)|$.) By (4′), $\|f\| + \|g\|$ is an upper bound of all expressions $|(f+g)(x)|$, $x \in A$. Thus $\|f\| + \|g\|$ cannot be *less* than $\sup |(f + g)(x)|$, $x \in A$. But, by definition, $\sup |(f + g)(x)| = \|f + g\|$. Thus $\|f + g\| \leq \|f\| + \|g\|$, as required.

Formula (4′) also shows that the function $f + g$ is *bounded* on A and hence is a member of W. Thus we have the closure law

$$(\forall f, g \in W) \quad f + g \in W.$$

The reader will easily verify that also $af \in W$ when $a \in E^1$ and $f \in W$ (i.e., af is bounded if f is) and that W also has all other properties of a normed linear space over E^1.

Definition 2.

In every normed (or semi-normed) linear space V, we define the *distance* $\rho(u, v)$ between two points $u, v \in V$ by $\rho(u, v) = |u - v|$.

The resulting distances depend, of course, on the norm defined in V. In particular, using the standard norm in C^n or E^n (cf. Example 1), we have

$$\rho(u, v) = \sqrt{\sum_{k=1}^{n} |u_k - v_k|^2}.$$

If, instead, the "nonstandard" norm of Example (2) is used, we obtain

$$\rho(u,v) = \sqrt[p]{\sum_{k=1}^{n} |u_k - v_k|^p}.$$

Under the semi-norm of Example (3), we have $\rho(u,v) = |u_1 - v_1|$. In the space W described in Example (4), we have $\rho(f,g) = \|f-g\| = \sup |f(x) - g(x)|$, $x \in A$.

In all cases, distances are nonnegative real numbers (for so are all absolute values by definition). Moreover, proceeding exactly as in the proof of Theorem 3 of §2, we see that distances resulting from any norm on V ("norm-induced" distances) obey the laws stated there, i.e.,

(1) $\rho(u,v) \geq 0$;

(1′) $\rho(u,v) = 0$ *iff* $u = v$;

(2) $\rho(u,v) = \rho(v,u)$ (symmetry law); *and*

(3) $\rho(u,w) \leq \rho(u,v) + \rho(v,w)$ (triangle inequality).

The details are left to the reader.

Note 2. Distances resulting from a *semi-norm* ("seminorm-induced" distances) have the same properties, except that (1′) is replaced by the weaker law $\rho(u,u) = 0$; so distances may vanish even if $u \neq v$ (which is excluded under *norm*-induced distances).

Moreover, in normed and semi-normed spaces, distances are *translation invariant*; that is, the *distance $\rho(u,v)$ does not change if both u and v are increased by one and the same vector x*, so that we have the following:

(4) $\rho(u,v) = \rho(u+x, v+x)$ (translation invariance).

Indeed, by definition,

$$\rho(u+x, v+x) = |(u+x) - (v+x)| = |u - v| = \rho(u,v).$$

Problems on Normed Linear Spaces

1. Prove laws (1), (2), and (3) for distances in semi-normed spaces and (1′) for normed spaces. Show also that $|\rho(u,w) - \rho(v,w)| \leq \rho(u,v)$.

2. Complete the proof of the assertions made in Example (4) as to the space W.

3. Verify that Example (3) yields a semi-norm; i.e., verify properties (i), (ii), and (iii) of Definition 1. Give examples of points u, v such that $\rho(u,v) = 0$, though $u \neq v$, under distances induced by that semi-norm.

4. Verify that Note 3 at the end of §4 applies to normed linear spaces (not only to Euclidean spaces), with lines defined as in §9.

5. Prove the *principle of nested line segments* (Problem 7 of §9) for normed linear spaces in general.

6. Let M be the set of all infinite *bounded* sequences $\{x_m\}$ in E^1 (or in C), i.e., sequences such that

$$(\forall m) \quad |x_m| \leq c$$

for some fixed $c \in E^1$.[1] We briefly denote such a sequence by a *single* letter (e.g., x) and use the same letter, with subscripts, to denote the terms x_m; thus $x = (x_1, x_2, \dots, x_m, \dots)$. Addition of sequences is defined termwise, i.e.,

$$x + y = (x_1 + y_1, x_2 + y_2, \dots, x_m + y_m, \dots).$$

Similarly, for $a \in E^1$ ($a \in C$),

$$ax = (ax_1, ax_2, \dots, ax_m, \dots).$$

Show that this makes M a vector space (with each bounded sequence treated as a single "point" in M). Also solve a similar problem for the set S of *all* sequences in E^1 (or C).

7. Continuing Problem 6, define a norm on M by

$$\|x\| = \sup_m |x_m|, \quad m = 1, 2, \dots .$$

Verify properties (i)–(iii) of Definition 1 for that norm, and give a formula for distances in M.

[Hint: Proceed as in Example 4.]

8. Verify that Example 4 remains valid also if W is defined to be the set of all bounded functions from A into the *complex field* C, with all other definitions unchanged.

9. In differential calculus it is shown that

$$a^{1/p}\, b^{1/q} \leq \frac{a}{p} + \frac{b}{q}$$

if a, b, p, $q \in E^1$, $a \geq 0$, $b \geq 0$, $p > 0$, $q > 0$, and

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Assuming this result, prove *Hölder's inequality*: If $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any x_k, $y_k \in C$,

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q}.$$

[1] The constant c may be different for different sequences in M.

[Hint: Let

$$A = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \text{ and } B = \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q}.$$

If $A = 0$ or $B = 0$, then all x_k or all y_k vanish, and the inequality is trivial. Thus assume $A \neq 0$, $B \neq 0$. Then, setting

$$a = \frac{|x_k|^p}{A^p} \text{ and } b = \frac{|y_k|^q}{B^q}$$

in the "calculus" inequality stated above, obtain

$$\frac{|x_k y_k|}{AB} \leq \frac{|x_k|^p}{pA^p} + \frac{|y_k|^q}{qB^q}, \quad k = 1, 2, \ldots, n.$$

Now add up these inequalities, substitute the values of A, B, and simplify.]

10. Prove the *Minkowski inequality*:

$$\left(\sum_{k=1}^{n} |x_k + y_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |y_k|^p\right)^{1/p}$$

for any real $p \geq 1$ and x_k, $y_k \in C$.

[Hint: If $p = 1$, this follows by the triangle inequality in C. If $p > 1$, let

$$A = \sum_{k=1}^{n} |x_k + y_k|^p \neq 0 \quad \text{(if } A = 0\text{, all is trivial).}$$

Then verify (writing $\sum$ for $\sum_{k=1}^{n}$ for simplicity):

$$\begin{aligned} A &= \sum |x_k + y_k|\, |x_k + y_k|^{p-1} \\ &\leq \sum |x_k|\, |x_k + y_k|^{p-1} + \sum |y_k|\, |x_k + y_k|^{p-1}. \end{aligned}$$

Now apply Hölder's inequality (Problem 9) to each of the last two sums, with $q = p/(p-1)$, so that $(p-1)q = p$ and $1/p = 1 - 1/q$. Thus obtain

$$A \leq \left(\sum |x_k|^p\right)^{1/p} \left(\sum |x_k + y_k|^p\right)^{1/q} + \left(\sum |y_k|^p\right)^{1/p} \left(\sum |x_k + y_k|^p\right)^{1/q}.$$

Now divide by $A^{1/q} = (\sum |x_k + y_k|^p)^{1/q}$ and simplify.]

11. Verify that

$$\|v\| = \sqrt[p]{\sum_{k=1}^{n} |v_k|^p}$$

defines a norm for E^n and C^n, satisfying the norm properties (i)–(iii), if $p \geq 1$.

[Hint: For the triangle inequality, use Problem 10. The rest is easy.]

Notation

Index