TOPICS IN DISCRETE MATHEMATICS

A.F. Pixley
Harvey Mudd College

July 21, 2010

Contents

Preface

This text is intended as an introduction to a selection of topics in discrete mathematics. The choice of topics in most such introductory texts is usually governed by the supposed needs of students intending to emphasize computer science in their subsequent studies. Our intended audience is somewhat larger and is intended to include any student seriously interested in any of the mathematical sciences. For this reason the choice of each topic is to a large extent governed by its intrinsic mathematical importance. Also, for each topic introduced an attempt has been made to develop the topic in sufficient depth so that at least one reasonably nontrivial theorem can be proved, and so that the student can appreciate the existence of new and unexplored mathematical territory.

For reasons that are not entirely clear, at least to me, discrete mathematics seems to be not as amenable to the intuitive sort of development so much enjoyed in the study of beginning calculus. Perhaps one reason for this is the fortuitous notation used for derivatives and integrals which makes such topics as the chain rule for derivatives and the change of variable theorems for integrals so easy to understand. But, for example, in the discrete calculus, (presented in Chapter 3 of this book), despite many efforts, the notation is not quite so natural and suggestive. It may also just be the case that human intuition is, by nature, better adapted to the study of the continuous world than to the discrete one. In any case, even in beginning discrete mathematics, the role of proper mathematical reasoning and hence the role of careful proofs seems to be more essential than in beginning continuous mathematics. Hence we place a great deal of emphasis on careful mathematical reasoning throughout the text.

Because of this, the prerequisites I have had in my mind in writing the text, beyond rigorous courses in single variable and multivariable calculus, include linear algebra as well as elementary computer programming. While little specific information from these subjects is used, the expectation is that the reader has developed sufficient mathematical maturity to begin to engage in reasonably sophisticated mathematical reasoning. We do assume familiarity with the meanings of elementary set and logical notation. Concerning sets this means the membership ($\in$) and inclusion ($\subset$) relations, unions, intersections, complements, cartesian products, etc.. Concerning logic this means the propositional connectives ("or", "and", "negation", and "implication") and the meanings of the existential and universal quantifiers. We develop more of these topics as we need them.

Mathematical induction plays an important role in the topics studied and an appendix on this subject is included. In teaching from this text I like to begin the course with this appendix.

Chapters 1 (Combinatorics) and 2 (The Integers) are the longest and the most important in the text. With the exception of the principle of inclusion and exclusion (Section 1.8) which is used in Section 2.5 to obtain Legendre's formula for the Euler ϕ-function, and a little knowledge of the binomial coefficients, there is little dependence

of Chapter 2 on Chapter 1. The remaining chapters depend on the first two in varying amounts, but not at all on each other.

Chapter 1

Combinatorics

1.1 Introduction

Combinatorics is concerned with the possible arrangements or configurations of objects in a set. Three main kinds of combinatorial problems occur: existential, enumerative, and constructive. Existential combinatorics studies the existence or nonexistence of certain configurations. The celebrated "four color problem"— Is there a map of possible "countries" on the surface of a sphere which requires more than four colors to distinguish between countries?— is probably the most famous example of existential combinatorics. Its negative "solution" by Appel and Haken in 1976 required over 1000 hours of computer time and involved nearly 10 billion separate logical decisions.

Enumerative combinatorics is concerned with counting the number of configurations of a specific kind. Examples abound in everyday life: how many ways can a legislative committee of five members be chosen from among ten Democrats and six Republicans so that the Republicans are denied a majority? In how many ways can such a committee, once chosen, be seated around a circular table? These and many other simple counting problems come to mind.

Constructive combinatorics deals with methods for actually finding specific configurations, as opposed to simply demonstrating their existence. For example, Los Angeles County contains at least 10 million residents and by no means does any human being have anywhere near that many hairs on his or her head. Consequently we must conclude (by existential combinatorics) that at any instant at least two people in LA County have precisely the same number of hairs on their heads! This simple assertion of existence is, however, a far cry from actually prescribing a method of finding such a pair of people — which is not even a mathematical problem. Constructive combinatorics, on the other hand, is primarily concerned with devising algorithms — mechanical procedures — for actually constructing a desired configuration.

In the following discussion we will examine some basic combinatorial ideas with emphasis on the mathematical principles underlying them. We shall be primarily

concerned with enumerative combinatorics since this classical area has the most connections with other areas of mathematics. We shall not be much concerned at all with constructive combinatorics and only in the following discussion of the Pigeonhole principle and Ramsey's theorem will we be studying a primary area of existential combinatorics.

1.2 The Pigeonhole Principle

If we put into pigeonholes more pigeons than we have pigeonholes then at least one of the pigeonholes contains at least two pigeons. If n people are wearing $n+1$ hats, then someone is wearing two hats. The purely mathematical content of either of these assertions as well as of the "LA County hair assertion" above is the same:

Proposition 1.2.1 *(Pigeonhole Principle) If a set of at least $n+1$ objects is partitioned into n non-overlapping subsets, then one of the subsets contains at least two objects.*

The proof of the proposition is simply the observation that if each of the n non-overlapping subsets contained at most 1 object, then altogether we would only account for at most n of the at least $n+1$ objects.

In order to see how to apply the Pigeonhole Principle some discussion of partitions of finite sets is in order. A *partition* π of a set S is a subdivision of S into non-empty subsets which are disjoint and exhaustive, i.e.: each element of S must belong to one and only one of the subsets. Thus $\pi = \{A_1, ..., A_n\}$ is a partition of S if the following conditions are met: each $A_i \neq \emptyset$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and $S = A_1 \cup \cdots \cup A_n$. The A_i are called the *blocks* or *classes* of the partition π and the number of blocks n is called the *index* of π and denote it by $index(\pi)$. Thus the states and the District of Columbia form the blocks of a partition of index 51 of the set of all residents of the United States. A partition π of a set S of n elements always has index $\leq n$ and has index $= n$ iff each block contains precisely one element. At the other extreme is the partition of index 1 whose only block is S itself. With this terminology the Pigeonhole Principle asserts:

> *If π is a partition of S with $index(\pi) < |S|$, then some block contains at least two elements. ($|S|$ denotes the number of elements in S.)*

Partitions of S are determined by certain *binary relations* on S called *equivalence relations*. Formally, R denotes a binary relation on S if for each ordered pair (a, b) of elements of S either a stands in the relation R to b (written aRb) or a does not stand in the relation R to b. For example, $=, <, \leq, |$ (the latter denoting divisibility) are common binary relations on the set $\mathbb{Z}$ of all integers. There are also common non-mathematical relations such as "is the brother of" among the set of all men, or "lives in the same state as" among residents of the United States (taking DC as a

state). The binary relation R on S is an equivalence relation on S if it satisfies the laws:

reflexive aRa for all $a \in S$,

symmetric aRb implies bRa for all $a, b \in S$,

transitive aRb and bRc imply aRc for all $a, b, c \in S$.

The relation of equality is the commonest equivalence relation, and in fact equivalence relations are simply generalizations of equality. An important example illustrating this is *congruence modulo* m: if m is any non-negative integer we say a *is congruent to* b *modulo* m (and express this by $a \equiv b \pmod{m}$) provided $a-b$ is divisible by m, i.e.: provided $m|(a-b)$. Equivalently, $a \equiv b \pmod{m}$ means that a and b have the same remainder r in the interval $0 \leq r < m$ upon division by m. This remainder is often called the *residue of* $a \pmod{m}$. Since $0|b$ iff $b = 0$, $a = b$ iff $a \equiv b \pmod{0}$, i.e.: equality is the special case of congruence modulo 0. Congruence and divisibility are discussed in detail in Chapter 2.

Equivalence relations and partitions of a set uniquely determine one another: given a partition π, the corresponding equivalence relation R is defined by aRb iff a and b are in the same π-block; R so defined is clearly an equivalence relation on S. Conversely, if R is an equivalence relation on S, for each $a \in S$ let the *R-block of* a, $a/R = \{b \in S : aRb\}$. From the definition of an equivalence relation above it is easy to check that the R-blocks are non-empty, disjoint, and exhaustive and hence constitute the blocks of a partition. The *index* of R is the index of the corresponding partition. If m is a positive integer, congruence mod m is an equivalence relation on $\mathbb{Z}$ of index m. The blocks, often called *residue classes* are the m subsets of integers, each residue class consisting of all those integers having the same remainder r upon division by m. Thus there are m distinct classes each containing exactly one of $r = 0, 1, \ldots, m-1$.

If S is the set of residents of LA County and for $a, b \in S$, we define aRb iff a and b have the same number of hairs on their heads at the present instant, then the index of R is the size of the set of all integers k such that there is at least one person in LA County having precisely k hairs on his head. This is no larger than the maximum number of hairs occurring on any persons head, and this is surely less than $|S|$. Hence some R-block contains at least two persons. If S is a set of $n+1$ hats, define aRb iff a and b are on the same one of n possible heads. Then $index(R) \leq n < |S| = n+1$.

Often a partition of a set is induced by a function. If S and T are finite sets and $f : S \to T$ is a function with domain S and taking values in T, for each $t \in T$ let $f^{-1}(t) = \{s \in S : f(s) = t\}$ be the inverse image of t. The collection of nonempty subsets $\{f^{-1}(t) : t \in T\}$ forms a partition of S with index $\leq |T|$ and corresponding equivalence relation R defined by aRb iff $f(a) = f(b)$. If T is strictly smaller than S then we have the following version of the Pigeonhole Principle:

> *If S and T are finite sets and $|T| < |S|$ and $f : S \to T$, then one of the subsets $f^{-1}(t)$ of S contains at least two elements.*

Thus in the hat example, think of S consisting of $n+1$ hats, T a set of n people, and $f : S \to T$ as an assignment of hats to people.

Examples of the Pigeonhole Principle

1. *Let S be a set of three integers. Then some pair of them has an even sum.*

This is clear since among any three integers there is either and even pair or an odd pair, and in either case the sum of the pair is even. Explicitly, let $T = \{0, 1\}$ and define $f : S \to T$ by $f(s) =$ the residue of s mod 2, i.e.: $f(s) = 0$ or 1 according as s is even or odd. Then one of $f^{-1}(0)$ or $f^{-1}(1)$ has at least two elements.

2.(G. Polya) *At any party of n people ($n \geq 2$ to make it a party), at least two persons have exactly the same number of acquaintances.*

Let S be the set of party participants and $N(a)$ be the number of acquaintances of a for each $a \in S$. Define aRb iff $N(a) = N(b)$. Since $=$ is an equivalence relation, from our definition it follows immediately that R is also an equivalence relation on S. Next we need to clarify the relation of "acquainted with". It is reasonable to assume that it is a symmetric relation and we shall need this assumption to justify our claim, which is that $index(R) < n$. It is also reasonable to assume either that it is reflexive (everyone is acquainted with himself) or irreflexive (no one is acquainted with himself). The important thing is that we decide one way or the other; the outcome will be the same either way. Let us assume that "acquainted with" is reflexive. From this it follows that for each $a \in S$, $1 \leq N(a) \leq n$. Now if $index(R) = n$ this would mean that $N(a) \neq N(b)$ for all distinct $a, b \in S$ and hence $N(a) = 1$ for some a and $N(b) = n$ for some b. But $N(b) = n$ implies that b is acquainted with everyone and hence with a so, by symmetry, a is acquainted with b, which contradicts $N(a) = 1$. Hence $index(R) < n$ and we are done, by the Pigeonhole Principle.

On the other hand if we assume that "acquainted with" is irreflexive then for each $a \in S$, $0 \leq N(a) \leq n-1$ and $index(R) = n$ again is easily seen to lead to a contradiction. Either way we analyze "acquainted with", we have shown that the function N is defined on the set S of n elements and has range of size less than n, so the function version of the Pigeonhole Principle could also be applied, perhaps most easily. Finally, it is perhaps worth noticing that the "acquainted with" relation is not reasonably assumed to be transitive and hence it cannot be the equivalence relation to which we apply the Pigeonhole Principle.

If we assume that the index of a partition on S is *much* much less than the size of S then we can conclude more about the block sizes. For example suppose $|S| = n^2+1$ and $index(\pi) = n$. Then if each block contained at most n elements, this would account for at most $n \cdot n = n^2$ of the elements of S. Hence some block must contain

at least $n+1$ elements. This is a special case of the following slight generalization of the Pigeonhole Principle.

Proposition 1.2.2 *(Generalized Pigeonhole Principle) If a finite set S is partitioned into n blocks, then at least one of the blocks has at least $|S|/n$ elements, i.e.: some block must have size at least equal to the average of the block sizes.*

PROOF: Suppose that the blocks are $A_1, ..., A_n$ and that $|A_i| < |S|/n$ for each i. Then we have

$$|S| = \sum_{i=1}^{n} |A_i| < \sum_{i=1}^{n} |S|/n = |S|,$$

a contradiction.

Notice that if $|S|/n = r$ is not an integer, then the conclusion must be that some block contains k elements where k is the least integer greater than r. For example, if a set of 41 elements is partitioned into 10 blocks, some block must contain at least 5 elements. Thus in the special case above where $|S| = n^2 + 1$, $(n^2+1)/n = n + 1/n$ so that we conclude that some block contains $n+1$ elements.

The following application illustrates the fact that the Pigeonhole Principle often occurs embedded in the midst of some more technical mathematical reasoning. First we clarify some properties of sequences of real numbers. Consider a finite sequence of real numbers

$$a_1, ..., a_n.$$

The sequence is *monotone* if it is either *non-increasing*, i.e.: $a_1 \geq a_2 \geq \cdots \geq a_n$, or *non-decreasing*, i.e.: $a_1 \leq a_2 \leq \cdots \leq a_n$. The terms

$$a_{i_1}, ..., a_{i_m}$$

are a *subsequence* provided

$$i_1 < i_2 < \cdots < i_m.$$

3. (P. Erdös and A. Szekeres) *Every sequence of n^2+1 real numbers contains a monotone subsequence of length at least $n+1$.*

Let the sequence be $a_1, ..., a_{n^2+1}$ and assume there is no non-decreasing subsequence of length $n+1$. Now we make a clever definition which will enable us to exploit the Pigeonhole Principle: for each $k = 1, ..., n^2+1$, let m_k be the length of the longest non-decreasing subsequence which starts with a_k. Then our assumption is equivalent to the assertion that $m_k \leq n$ for all k. Since also $1 \leq m_k$ for all k, the numbers $m_1, ..., m_{n^2+1}$ are n^2+1 integers all between 1 and n. Hence they are partitioned by equality into at most n blocks. Thus, by the Generalized Pigeonhole Principle some $n+1$ of them are equal. Let

$$m_{k_1} = \cdots = m_{k_{n+1}}$$

where

$$1 \leq k_1 < \cdots < k_{n+1} \leq n^2 + 1.$$

Suppose some $a_{k_i} < a_{k_{i+1}}$. Then since $k_i < k_{i+1}$, we could take a longest non-decreasing subsequence beginning with $a_{k_{i+1}}$ and put a_{k_i} in front of it to form a longer non-decreasing subsequence. But this implies $m_{k_i} > m_{k_{i+1}}$, contradicting their equality. Hence we must have $a_{k_i} \geq a_{k_{i+1}}$ for all $i = 1, ..., n$. Therefore

$$a_{k_1} \geq a_{k_2} \geq \cdots \geq a_{k_{n+1}}$$

is a non-increasing subsequence of length $n + 1$.

Exercises Section 1.2

1. Let $k_1, ..., k_m$ be a sequence of $m > 1$ (not necessarily distinct) integers. Show that some subsequence of consecutive terms has sum divisible by m. (Hint: Consider $S = \{k_1, k_1 + k_2, ..., k_1 + \cdots + k_m\}$.)

2. Given a set S of 10 positive integers all less than 100, show that S has two distinct subsets with the same sum.

3.a) Show that if $n + 1$ integers are chosen from $T = \{1, 2, ..., 2n\}$, there will be two which differ by 1.
b) In a) show that there will also be two, one of which divides the other. Hint: Let S be the set of $n + 1$ integers chosen from T. For each $m \in S$, $m = 2^p \cdot q$ where q is an odd number uniquely determined by m. Then let $f : S \to \{1, 3, \ldots, 2n - 1\}$ be defined by $f(m) = q$.
c) Show that if $n + 1$ integers are chosen from $T = \{1, 2, ..., 3n\}$, then there will be two which differ by at most 2.

4.The entire college student body is lined up at the door to a classroom and enter at the rate of one per second. How long will it take to be certain that a dozen members of at least one of the four classes (freshman, sophomore, junior, senior) has entered?

5. Suppose 5 points are chosen inside or on the boundary of a square of side 2. Show that there are two of them whose distance apart is at most $\sqrt{2}$.

6.a) Show that if 5 points are chosen from within (or on the boundary of) an equilateral triangle of side 1, there is some pair whose distance apart is at most 1/2.
b) Show that if 10 points are chosen from an equilateral triangle of side 1, there is some pair whose distance apart is at most 1/3.
c) For each integer $n \geq 2$ find an integer $p(n)$ such that if $p(n)$ points are chosen from from an equilateral triangle of side 1, there is some pair whose distance

apart is at most $1/n$? Hint: First show that the sum of the first n odd integers $(2n-1)+(2n-3)+\cdots+5+3+1=n^2$. (Notice that you are *not* asked to find a $p(n)$ which is the least integer with the desired property. Do you think the $p(n)$ you found is the least such integer?)

7. Prove the following version of the Generalized Pigeonhole Principle: If $f : S \to T$ where S and T are finite sets such that for some integer m, $m|T| < |S|$, then at least one of the subsets $f^{-1}(t)$ of S contains at least $m+1$ elements.

1.3 Ramsey's Theorem

In this section we will discuss a far reaching, important, and profound generalization of the Pigeonhole Principle, due to Frank Ramsey (1903-1930). It will provide us with an important example of the distinction between existential and constructive combinatorics. We shall prove only a restricted version of the theorem. We shall be content with this since it is the most important version and contains all of the ideas of the general version.

A popular "colloquial" special case of Ramsey's theorem is stated in terms of the "acquainted with" relation:

> *In a party of 6 or more acquaintances either there are 3 each pair of whom shake hands or there are 3 each pair of whom do not shake hands.*

We can also (and more conveniently) state this equivalently as a statement about graphs. Here a finite *graph* is a system $G = (V, E)$ where V is a finite set of "vertices" and E is a set of "edges", i.e.: a set of unordered pairs of vertices. Hence E is a symmetric relation on V. We shall also assume that E is irreflexive: this means that our graphs have no "loops"; hence in the statement above we assume no one is acquainted with himself. (Notice, incidentally, that Example 2 of the preceding section can also be restated as a property of graphs: in any graph with $n \geq 2$ vertices, at least two vertices lie on exactly the same number of edges.)

Corresponding to our assumption above that each pair of party participants are acquainted, our restricted version of Ramsey's theorem is concerned with *complete* graphs of order n, by which we mean that E consists of all pairs of distinct vertices from the set of n vertices. We denote the complete graph of order n — or more correctly, any copy of it — by K_n. It is easy to verify that K_n has $n(n-1)/2$ edges. (Do so!) In particular K_3 is usually called a triangle. Corresponding to pairs of persons shaking hands or not (again a symmetric relation), we suppose that some edges of the corresponding K_n are colored red or blue. Then the special case of Ramsey's theorem stated above takes the equivalent form:

> *For any coloring of the edges of K_n, $n \geq 6$, using colors red and blue, there is always a red K_3 or a blue K_3, i.e.: there is always a monochromatic triangle.*

We abbreviate this statement by writing:

$$\text{for } n \geq 6, \quad K_n \to K_3, K_3$$

and refer to this as a *Ramsey mapping*.

As a first step in proving Ramsey's theorem we first observe that if

$$K_6 \to K_3, K_3$$

is true then

$$K_n \to K_3, K_3$$

is true for all $n \geq 6$. This is because if we are given any coloring of a K_n, we can choose any six vertices; since these vertices are part of the K_n they form a colored K_6 which must contain a monochromatic triangle which is in turn contained in the K_n.

Next let us prove the assertion $K_6 \to K_3, K_3$. We could do this by considering cases, which would be tedious and not instructive. The following proof is both easy and — more important — we shall be able to generalize it. Hence we present it in detail:

Choose any one of the vertices v of the colored K_6. v lies on 5 edges of K_6. Each of these 5 edges is red or blue. By the Pigeonhole Principle some 3 of these 5 edges is monochromatic, say red. Let these 3 red edges join v to the vertices a, b, c. Consider the triangle abc. If it is monochromatic we are done; otherwise triangle abc contains a red edge, say $\{a, b\}$. Hence the triangle vab is red, and we are done.

Finally let us observe that $K_5 \not\to K_3, K_3$, i.e.: it is possible to color K_5 in such a way that there is no monochromatic triangle. To see this let the vertices be those of a regular pentagon and color the 5 "outside" edges red and the 5 "inside" edges blue.

Now we can state our restricted version of Ramsey's theorem:

Theorem 1.3.1 *(Ramsey's theorem — restricted version) If $m \geq 2$ and $n \geq 2$ are integers then there is an integer r such that*

$$K_r \to K_m, K_n.$$

This means that if the edges of K_r are colored arbitrarily using red and blue, then it contains a red K_m or a blue K_n.

For the reason given above (for K_6), if $K_r \to K_m, K_n$ then for any $q \geq r$, $K_q \to K_m, K_n$. The *Ramsey number* $r = r(m, n)$ is defined to be the *least integer* such that $K_r \to K_m, K_n$. From Ramsey's theorem, it is obvious that for each $m, n \geq 2$, $r(m, n)$ exists.

We have just proved above that $r(3, 3) = 6$. This is because we proved both $K_6 \to K_3, K_3$, which shows that $r(3, 3) \leq 6$ and $K_5 \not\to K_3, K_3$, which shows that $r(3, 3) > 5$. It is important to see that both proofs were necessary. We shall prove Ramsey's theorem below by showing that for given m and n there is *some* positive integer q for which $K_q \to K_m, K_n$, and hence there is a *least* q and this is what we

call $r(m,n)$. Our proof will give little clue as to what $r(m,n)$ actually is and, in fact, there is no general method for finding Ramsey numbers. Their determination is still a big mathematical mystery! We can, however, make some simple observations about $r(m,n)$:

i) $r(m,n) = r(n,m)$ for all m, n.

This is immediate if we observe that each coloring of K_r corresponds to a unique complementary coloring (interchange red and blue).

ii) $r(2,m) = m$ for all $m \geq 2$.

First $r(2,m) \leq m$ since if the edges of K_m are colored red or blue then we either have a red edge or all of K_m is blue, i.e.: $K_m \to K_2, K_m$. Second, $r(2,m) > m-1$, for if we color all edges of K_{m-1} blue, then we have no red edge and no blue K_m, i.e.: $K_{m-1} \not\to K_2, K_m$.

The $r(2,m) = r(m,2) = m$ are called the trivial Ramsey numbers. The only others known (as of when this was written) are the following:

n	=	3	4	5	6	7	8	9
$r(3,n)$	=	6	9	14	18	23	28	36
$r(4,n)$	=	9	18	25				

Some of these numbers, for example $r(3,8)$ and $r(4,5)$ have been discovered only in the last few years. Some recent estimates of still unknown Ramsey numbers are $45 \leq r(5,5) \leq 49$ and $102 \leq r(6,6) \leq 165$ but it seems very hard, short of massively impractical computations, to improve even the first of these.

Now for the proof of Ramsey's theorem. Actually we prove the following recursive inequality for the Ramsey numbers: for all integers $m, n \geq 3$, $r(m,n)$ exists and

$$r(m,n) \leq r(m-1,n) + r(m,n-1).$$

This will actually give us a recursively computable upper bound on the sizes of the ramsey numbers — but not a very good one. From the definition of $r(m,n)$ to show that $r(m,n)$ exists and satisfies this inequality, it is equivalent to show that if $N = r(m-1,n) + r(m,n-1)$, then

$$K_N \to K_m, K_n.$$

To prove this statement we use induction on the integer $m+n$. For the base step of the induction we observe that if $m, n \geq 3$ then the least relevant value of $m+n$ is $m+n=6$ with $m=n=3$ and we have already shown that

$$r(3,3) = 6 \leq 3+3 = r(2,3) + r(3,2).$$

Now for the induction step. Let p be a positive integer greater than 6. The induction hypothesis is: $r(m,n)$ exists and the inequality is true for all m and n with

$m + n < p$. Now let $m + n = p$; suppose

$$N = r(m-1, n) + r(m, n-1)$$

and that the edges of K_N are arbitrarily colored red or blue. We must prove that $K_N \to K_m, K_n$.

Pick a vertex v. v is connected to every other vertex and by the choice of N there are at least $r(m-1, n) + r(m, n-1) - 1$ of them. Then by the Pigeonhole Principle either

a) there are at least $r(m-1, n)$ vertices with red edges joining them to v, or

b) there are at least $r(m, n-1)$ vertices with blue edges joining them to v.

Case a): Since $(m-1)+n < m+n = p$, the induction hypothesis applied to these $r(m-1, n)$ vertices asserts that for the $K_{r(m-1,n)}$ formed by them,

$$K_{r(m-1,n)} \to K_{m-1}, K_n$$

i.e.: either
i) $K_{r(m-1,n)}$ contains a K_{m-1} with all red edges, in which case the vertices of this K_{m-1} together with v constitute a red K_m, and we are done, or
ii) $K_{r(m-1,n)}$ contains a K_n with all blue edges, in which case we also done.

Case b): This is symmetric with Case a). Since $m + (n-1) < m + n = p$, the induction hypothesis asserts that

$$K_{r(m,n-1)} \to K_m, K_{n-1}$$

i.e.: either
i) $K_{r(m,n-1)}$ contains a K_m with all red edges, in which case we are done, or
ii) $K_{r(m,n-1)}$ contains a K_{n-1} with all blue edges, in which case the vertices of this K_{n-1} together with v constitute a blue K_n, and we are done.

A good exercise at this point is to go through the first step in this induction argument, i.e.: from $r(3,3) = 6$ and $r(4,2) = 4$ use the method in the induction step above to establish that $r(4,3) \leq 10 = 6 + 4 = r(3,3) + r(4,2)$.

If we estimate $r(m, n)$ by $r(m-1, n) + r(m, n-1)$ we obtain for the first few values

n	$=$	3	4	5	6	7	8	9
$r(3,n)$	$\leq$	6	10	15	21	28	36	45
$r(4,n)$	$\leq$	10	20	35				

Comparing these estimates with the earlier table of known values indicates that our inequality provides only a rather poor estimate for $r(m, n)$.

Now we discuss the generalizations of our version of Ramsey's theorem and the general significance of what is now called "Ramsey theory". We begin by returning to Example 3 of Section 2 (Pigeonhole Principle), (the theorem of Erdös and Szekeres). In that example we showed that any real sequence of length n^2+1 contains a monotone subsequence of length $n+1$. Another way of looking at this result, is to ask the question: if one wants to find a monotone subsequence of prescribed length in an arbitrary sequence, is this possible if the sequence is taken to be sufficiently long? The answer is yes and in fact length n^2+1 was shown to be long enough to guarantee a monotone subsequence of length $n+1$. Let us show how to obtain, at least qualitatively, this same result from Ramsey's theorem. Let $a_1, ..., a_r$ be a sequence of length $r = r(n+1, n+1)$. In the complete graph K_r on the r vertices $\{1, ..., r\}$ label the edge joining vertices i and j by $i-j$ if $i < j$. Then color the edge $i-j$ red if $a_i < a_j$ and blue if $a_i \geq a_j$. Then since Ramsey's theorem asserts that $K_r \rightarrow K_{n+1}, K_{n+1}$, we conclude that the sequence has either a strictly increasing subsequence of length $n+1$ or a non-increasing one of length $n+1$, i.e.: the sequence has a monotone subsequence of length $n+1$. Stated still more generally, we conclude that no matter how "disordered" a sequence of real numbers is, if it is sufficiently long, (depending on n) it will contain a totally ordered subsequence of pre-assigned length $n+1$.

This illustrates the general significance of Ramsey's theorem and its extensions to what is called Ramsey theory. The general endeaver is to show that by taking a sufficiently large system or structure one can guarentee the existence of prescribed subsystems or substructures of a specified sort. In short "Complete disorder is impossible" (a quote from the mathematician T. S. Motzkin) encapsulates the theme.

These observations suggest that we generalize Ramsey's theorem. This can be done in many ways and we examine a few. First, we can extend our restricted version to any number of colors. Thus if $n_1, ..., n_k$ are prescribed positive integers ≥ 2 the ramsey number $r = r(n_1, ..., n_k)$ has the following defining property: if we color the edges of K_N, $N \geq r$, arbitrarily, using k colors, denoted by $1, ..., k$, we will have

$$K_N \rightarrow K_{n_1}, ..., K_{n_k},$$

meaning that the colored K_N will contain either a K_{n_1} of color 1, or a K_{n_2} of color 2, ... or a K_{n_k} of color k. This is easily established by using induction on k to show that

$$r(n_1, ..., n_k) \leq r(r(n_1, ..., n_{k-1}), n_k).$$

As an application, observe that this shows that

$$r(3,3,3) \leq r(r(3,3),3) = r(6,3) = 18.$$

With a bit of work it can be shown that actually $r(3,3,3) = 17$. In fact this is the only non-trivial mulicolor Ramsey number known at present.

Even further, we can extend Ramsey's theorem beyond just edge colorings. To understand this, for $t \geq 1$ let K_r^t be the collection of all t element subsets of a set of r

elements (vertices). For $t = 1$, K_r^1 is just the set of r vertices (or more precisely, the set of all one element subsets of the set of r vertices). For $t = 2$ we have $K_r^2 = K_r$, the complete graph on r vertices, already considered. For $t = 3$ we have the set of all triangles, etc. Now we consider colorings of these t element subsets of a set of r points and obtain the full (finite) Ramsey's theorem.

Theorem 1.3.2 *(Ramsey 1930) If $t \geq 1$ and $n_1, ..., n_k \geq t$, are prescribed integers, then there is an integer r such that*

$$K_r^t \to K_{n_1}^t, ..., K_{n_k}^t.$$

This means that for any coloring of the t element subsets of a set of r elements using k colors, there is either an n_1 element subset all of whose t element subsets are color 1, or an n_2 element subset all of whose t element subsets are color 2, ... or there is an n_k element subset all of whose t element subsets are color k. The least such integer r having the property guaranteed by Ramsey's theorem is the *Ramsey number* $r^t(n_1, ..., r_k)$. The proof of Ramsey's theorem consists in proving (by induction again) another recursive inequality.

A final striking application of Ramsey's Theorem is another theorem of Erdös and Szekeres. We say that a set S of points in the plane is *convex* if each triangle whose vertices are members of S does not contain another point of S in its interior or boundary. Thus any three points of the plane are convex and if each four point subset of S is convex then obviously S is convex.Thus to check a set for convexity we need only check its four point subsets for convexity. If a set is not convex we shall call it *concave*.

Theorem 1.3.3 *(Erdös and Szekeres) For any integer n there is an integer $E(n)$ such that any set $E(n)$ points of the plane with no three on a line, will contain an n-point convex set.*

We have no idea how large the number $E(n)$ might be, probably very large; however it is easy to show that it is certainly no larger than the Ramsey number $r^4(n, 5)$ and this will constitute a simple proof of the theorem. From what we already know, probably $r^4(n, 5)$ is much larger than $E(n)$, but since we know so little about either this shouldn't bother us.

The proof depends on the following lemma

Lemma 1.3.1 *There is no five point set in the plane each of whose four element subsets is concave.*

We leave the proof of the lemma as exercise 9, below. Then to complete the proof of the theorem, for a set of $E(n)$ points, no three on the same line, we agree to color each four element subset red if it is convex and blue if it is concave. Thus if we choose any $r^4(n, 5)$ points, no three on a line, there must be either a) an n element subset

with each four elements red and hence convex, or b) a five element subset with each four elements blue and hence concave. By the lemma, b) cannot occur and hence a), which implies that there is an n element convex subset, must occur.

There are also infinite versions of Ramsey's theorem. In fact Ramsey proved his original theorem first in the case of infinite subsets of the set of all integers. Here is a version:

Theorem 1.3.4 *For any finite coloring of $\mathbb{N} \times \mathbb{N}$ (i.e.: of the edges of the complete graph with vertices the set $\mathbb{N}$ of non-negative integers), there is an infinite subset $A \subset \mathbb{N}$ with $A \times A$ monochromatic.*

For example consider the simplest case of two colors. If we choose a very simple coloring, say we color the edge ab red if $a+b$ is even and blue if $a+b$ is odd, it is easy to find an infinite monochromatic subset. But for other colorings it is not at all obvious how to find one; for example color ab red if $a+b$ has an even number of prime factors and blue if $a+b$ has an odd number of prime factors. Hence it is remarkable that we can *always* (though not constructively!) find an infinite monochromatic subset.

It is also interesting to mention some examples of what are often called "Ramsey-type theorems". A famous example is the following theorem which has inspired many subsequent results of a similar character:

Theorem 1.3.5 *(van der Waerden's theorem, 1927) If the positive integers are partitioned into two disjoint classes in any way, one of the classes contains arbitrarily long arithmetic progressions.*

Beyond these there are also many mathematical results that illustrate Motzkin's maxim "Complete disorder is impossible". For example an important very basic fact about the real numbers is the following:

Theorem 1.3.6 *(Bolzano-Weierstrass theorem) Any bounded infinite sequence of real numbers has a convergent subsequence.*

Though this is not generally thought of as a Ramsey-type theorem, it does illustrate that " Complete disorder is impossible" (at least with a minimal hypothesis—in this case boundedness).

Finally let us recapture the original Pigeonhole Principle as a special case of Theorem 3.2. Let $t = 1$ and hence color the points of an r element set using k colors. Then $r^1(n_1, ..., n_k)$ is the least r which guarantees that for some i a subset of n_i points will be colored i. But it is easy to see that

$$r^1(n_1, ..., n_k) = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) + 1 = n_1 + \cdots + n_k - k + 1.$$

In particular if we take $n_1 = \cdots = n_k = 2$, this yields $r^1(2, ..., 2) = k+1$ which means simply that if $k+1$ points are colored using k colors, some pair have the same color — this shows explicitly how Ramsey's theorem extends the elementary Pigeonhole Principle.

Exercises Section 1.3

1. Show that $r(3,4) \leq 10 = 6+4 = r(4-1,3) + r(4,3-1)$ by carrying out the first induction step in the proof of Ramsey's theorem.

2. Use the fact that $r(3,3) = 6$ to prove that $r(3,3,3) \leq 17$. (Hint: Use the same strategy as in the proof of Ramsey's theorem.) A harder exercise is to show from this that $r(3,3,3) = 17$; this requires showing that $K_{16} \not\to K_3, K_3, K_3$. You may want to try this as an optional exercise.

3. Establish the multi-color version of Ramsey's theorem by using induction to show that

$$r(n_1, \ldots, n_k) \leq r(r(n_1, \ldots, n_{k-1}), n_k).$$

4. Let n_3 and t be positive integers with $n_3 \geq t$. Show that $r^t(t, t, n_3) = n_3$.

5. Let $n_1, ..., n_k, t$ be positive integers such that $n_1, ..., n_k \geq t$ and let m be the largest of $n_1, ..., n_k$. Prove that if $r^t(m, ..., m)$ (k arguments) exists then so does $r^t(n_1, ..., n_k)$ and, in fact, $r^t(m, ..., m) \geq r^t(n_1, ..., n_k)$. (This shows that to prove Ramsey's theorem it is enough to prove the special case $n_1 = \cdots = n_k$.)

6. A binary relation $\leq$ on a set P is a *partial order* (or just an *order*) if $\leq$ satisfies the laws:

reflexive $a \leq a$ for all $a \in P$,

antisymmetric for $a, b \in P$, $a \leq b$ and $b \leq a$ implies $a = b$,

transitive for all $a, b, c \in P$, $a \leq b$ and $b \leq c$ implies $a \leq c$.

$(P, \leq)$ is called a *partially ordered set* (briefly, a *poset*), or just an *ordered set*. A subset $C \subset P$ is a *chain* if $a \leq b$ or $b \leq a$ for all $a, b \in C$, i.e.: if each pair of elements is comparable with respect to $\leq$. A subset $A \subset P$ is an *anti-chain* if $a \not\leq b$ for all distinct $a, b \in A$, i.e.: no two elements are comparable. [Example: $(\mathbb{Z}^+, |)$ is a poset where $|$ is divisibility; the set of primes is an anti-chain and the set $\{2^n : n \in \mathbb{N}\}$ is a chain.] An important theorem due to R. P. Dilworth is

> *If $(P, \leq)$ is a poset with $|P| \geq ab+1$ where $a, b \in \mathbb{Z}^+$, then $(P, \leq)$ contains either a chain of $a+1$ elements or an anti-chain of $b+1$ elements.*

Prove the variation of this Ramsey-type theorem obtained by replacing $ab + 1$ by $r(a+1, b+1)$.

7. Does Dilworth's theorem (Exercise 6, above) provide a relation between $ab+1$ and $r(a+1, b+1)$? Explain. Answer the analogous question for the theorem of Erdös

and Szekeres of Example 3 in Section1.2.

8. Prove that every infinite sequence of real numbers has an infinite monotone subsequence.
9. Prove Lemma 1.3.1. To do this suppose the contrary. Let the five points be v, w, x, y, z; no three of these can lie on a line. Consider any four of them, say w, x, y, z. One of these, say w, must then lie in the interior of the triangle xyz. Now show that it is impossible to locate v so that each four element subset is concave. Hint: draw the lines joining w to each of x, y, z and then consider the two cases: v located in a) the interior, and b) the exterior, of triangle xyz.

1.4 Counting Strategies

In enumerative combinatorics a principal task is to devise methods for counting various kinds of configurations without simply making long lists of their members. In fact devising ingenious methods for counting is part of the very essence of what mathematics is all about, while simply listing, even with the aid of a computer, whatever it may be, is usually not mathematics.

A good general strategy for counting — and for solving problems in general is the *divide and conquer* strategy: devise a rational scheme for dividing a big problem into a manageable number of smaller ones and solve each of these separately. Thus the most common method of counting without mindlessly listing is to devise some rational method of partitioning the set of configurations under consideration, so that each of the blocks of the partition can be easily counted, and then sum up the sizes of the blocks. For example if you want to count all of the residents of the United States, pass the job onto the state governors (and the mayor of DC) and sum up the numbers they report. Thus, in general, if $\pi = \{A_1, ..., A_n\}$ is a partition of the set A, then since the A_i are disjoint we have

$$|A| = |A_1| + \cdots + |A_n|.$$

In the special case where all blocks have the same size, $|A_1| = \cdots = |A_n| = b$, we have simply

$$|A| = index(\pi) \cdot b.$$

Example 1. A student wants to take either a math or a CS course at 10:00 am. If 5 math courses and 4 CS courses are offered at this time, then assuming that these two sets of courses are disjoint, i.e.: not double listed, there are $5 + 4 = 9$ ways for the student to select his 10 o'clock class.

On the other hand, suppose the 10 o'clock math and CS courses are not disjoint; suppose two of the courses are listed as both math *and* CS, then the student has only $5 + 4 - 2 = 7$ choices.

Generally, if $A = A_1 \cup A_2$ but $A_1 \cap A_2 \neq \emptyset$, then

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|,$$

the reason being that in $|A_1| + |A_2|$ we have counted the elements in $A_1 \cap A_2$ twice. The formula we have just obtained is a special case of the principle of *inclusion and exclusion* which is concerned with the general case of $n \geq 2$ sets, and to which we shall return later.

The partitioning principle, or its generalization to the inclusion-exclusion principle, enables us to count the number of elements in a single set, or to count the number of ways in which we can choose a single element from a single set. If the set has been partitioned into $A = A_1 \cup A_2 \cup \cdots$, the principle tells us that we choose an element from A by choosing an element from A_1 *or* from A_2,*or* from A_3, etc. The operative conjunction here is the word *or*. For this reason the partitioning principle is also called the *addition* principle.

For another divide and conquer strategy suppose that we wish to *independently* choose elements a_1 from A_1 *and* a_2 from A_2, ... *and* a_n from A_n. Then the number of choices is evidently

$$|A_1| \times \cdots \times |A_n|$$

which is the number of n-tuples $(a_1, ..., a_n)$ where $a_i \in A_i,\ i = 1, ..., n$. (This is also just the number of elements in the cartesian product $A_1 \times \cdots \times A_n$ of the sets.) This is called the *multiplication* principle. The operative conjunction here is the word *and*; we can think of the choices as being made either simultaneously or in sequence and it does not matter if the sets A_i overlap so long as the choices are independent.

Example 2. Our student now wants to take both one of 5 math courses, each offered at 10 o'clock and one of 4 CS courses, each offered at 11. Then the number possible ways to schedule his 10 and 11 o'clock hours is $5 \times 4 = 20$. For another example suppose the student wants to take 2 out of 5 math courses and all are offered at both 10 and 11 then he can choose any of the 5 at 10 and, independently, any of the remaining 4 at 11 for a total of $5 \times 4 = 20$ choices.

Finally lets consider a more complicated example, one which combines partitioning with the multiplication principle.

Example 3. How many integers n in the interval $1 \leq n \leq 10^4$ have exactly one digit equal to 1?

We partition the set A of all such integers into the disjoint subsets A_k of k digit numbers, for $k = 1, ..., 5$. Then $A_1 = \{1\}$ and $A_5 = \{10,000\}$, so

$$|A_1| = |A_5| = 1.$$

We further partition A_2 into two disjoint sets: those with tens digit 1 and those with units digit 1. Since the tens digit cannot be 0 we thus have

$$|A_2| = 9 + 8 = 17.$$

For A_3 we partition this into the disjoint sets where the 1 occurs as the hundreds, tens, or units digit respectively, and obtain (adding the counts for these cases in order),

$$|A_3| = 9 \times 9 + 8 \times 9 + 8 \times 9 = 225.$$

Likewise we have

$$|A_4| = 9 \times 9 \times 9 + 8 \times 9 \times 9 + 8 \times 9 \times 9 + 8 \times 9 \times 9 = 2673.$$

Therefore

$$|A| = 1 + 17 + 225 + 2673 + 1 = 2917.$$

The following theorem is an important application of the multiplication principle.

Theorem 1.4.1 *If S and T are finite sets then the number of functions $f : S \to T$ (i.e.: with domain S and range a subset of T) is $|T|^{|S|}$.*

PROOF. If we denote the elements of S and T by $S = \{s_1, ..., s_n\}$ and $T = \{t_1, ..., t_m\}$, then for each s_i, $f(s_i)$ must be exactly one of $t_1, ..., t_m$. Hence to select a function we independently make n choices from the same set T. The multiplication principle then asserts that this can be done in

$$|T| \times \cdots \times |T| = |T|^{|S|}$$

ways.

What happens when one or both of S and T are empty? For $S \neq \emptyset$ and $T = \emptyset$, the argument just given counts the number of functions $f : S \to \emptyset$ as $0^{|S|} = 0$ (since we can choose no value for $f(s_i)$), which is consistent with the fact that $0^n = 0$ for $n \neq 0$. What about the case $S = \emptyset$? Often we take $m^0 = 1$ for $m \geq 0$ as a definition. Let us see that in fact there is exactly one function with domain $\emptyset$ and taking values in an arbitrary set T. To see this recall that a function $f : A \to B$ is simply a subset $f \subset A \times B$ with the properties,

i) if $a \in A$ then for some $b \in B$, $(a, b) \in f$ (i.e.: $domain(f) = A$), and

ii) if $(a, b), (a, c) \in f$ then $b = c$ (i.e.: f is single-valued).

Now $\emptyset \times T = \{(s, t) : s \in \emptyset, t \in T\} = \emptyset$ since there are no elements $s \in \emptyset$. Hence $\emptyset \subset \emptyset \times T$ is the only subset of $\emptyset \times T$ and the question is whether it satisfies properties i) and ii). But this is true since there are no $s \in \emptyset$ so that both i) and ii) are *vacuously* satisfied: each is an "if...then" statement with false hypothesis and hence is true. Therefore $\emptyset$ is the unique function with domain $\emptyset$ and having values in any other set. (We can also verify $0^{|S|} = 0$ for $S \neq \emptyset$ this way by observing that $S \times \emptyset = \emptyset$, but that in this case $\emptyset$ is not a function with domain S.)

The proof of this theorem suggests that if we start with sets as the basic entities of mathematics (functions, and numbers, etc. are then just special kinds of sets), then

it is reasonable to define m^n to be just the number of functions from an n element set to an m element set. If we do this then, rather than a definition, we have just shown that we then have no choice but to take $0^0 = 1$. More generally, this discussion suggests that we generalize exponentiation from numbers to sets and therefore for any sets S and T we define T^S to be the set of all functions $f : S \to T$, i.e.: with domain S and taking values in T. Then we define $|S|^{|T|}$ to equal $|S^T|$.

If S is any set and $A \subset S$, the *characteristic function* of A, denoted by χ_A is defined on S by

$$\chi_A(x) = \begin{array}{ll} 1 & \text{if } x \in A, \\ 0 & \text{if } x \in S - A. \end{array}$$

Obviously two subsets of S are equal iff they have the same characteristic functions. Hence the subsets of S are in 1-1 correspondence with their characteristic functions, and these are just the functions with domain S and having their values in the two element set $\{0, 1\}$. In this correspondence the empty set and S have the constants functions 0 and 1 as their respective characteristic functions. The last theorem then has the following important corollary.

Corollary 1.4.1 *If S is a finite set the number of subsets of S is $2^{|S|}$.*

Another interesting way to understand this corollary is to observe that if we arrange the elements of S in any sequence and, for any subset A, label the elements in A with 1's and those not in A with 0's. The resulting finite sequence is a list of the values of χ_A and is also a binary number in the range from 0 to $2^{|S|} - 1$. For example, if $S = \{s_0, s_1, s_2\}$, the subset $\{s_0, s_2\}$ is labeled by the sequence 1 0 1, the binary equivalent of the decimal number 5. Thus in this labeling we have counted the subsets of S in binary and found that there are $2^{|S|}$ of them.

Exercises Section 1.4

1. How many distinct positive integer divisors do each of the following integers have?

   a) $2^4 \times 3^2 \times 5^6 \times 7$

   b) 340

   c) 10^{10}

2. If A, B, C are finite sets show that

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

3. Among 150 people, 45 swim, 40 bike, and 50 jog. Also 32 jog but don't bike, 27 jog and swim, and 10 people do all three.

a) How many people jog but neither swim nor bike?

b) If 21 people bike and swim, how many people do none of the three activities?

4. How many 5 digit numbers can be constructed using the digits 2,2,2,3,5, each exactly once?

5. Five identical coins and two dice are thrown all at once. One die is white and the other is black. How many distinguishable outcomes are there? How many if both dice are white?

6. How many integers can be constructed using the digits 1,2,3,4, allowing for no repeated digits?

7. How many ways are there of rearranging the letters in the word "problems" which have neither the letter p first nor the letter s last.
Suggestion: Let A be the set of arrangements with p first and B be the set with s last. We want the number of arrangements in the complement $\overline{A \cup B}$ of $A \cup B$. Find a formula for $|\overline{A \cup B}|$ and substitute in the appropriate numbers.

1.5 Permutations and combinations

A *permutation* of a finite set S is simply an arrangement of the elements in a definite order. For example, the letters a, b, and c can be arranged in six orders:

abc acb bac bca cab cba

By the multiplication principle, if S has n elements, then a permutation of S is determined by choosing any of the n elements for the first position of the permutation, then, independently, any of the remaining $n-1$ elements for the second position, etc., so that we obtain

$$n(n-1)\cdots 3\cdot 2\cdot 1 = n!$$

permutations altogether.

A mathematically more precise way of defining a permutation is as a function; specifically a permutation of S is a function $f : S \to S$ which is both 1-1 and onto. In the example above, the permutation bca of the letters a,b,c can be thought of as the function f which replaces the natural order abc by bca, i.e.:

$$f(a) = b, \quad f(b) = c, \quad f(c) = a.$$

Likewise the arrangement abc is just the identity function: the function i which sends each of a,b,c into itself:

$$i(a) = a, \quad i(b) = b, \quad i(c) = c.$$

Thinking of permutations of a set S as 1-1 functions of S onto S is important since it suggests the way to extend our present considerations to the case where S is infinite, and where it is not so clear, for example if S is the set of real numbers, what is meant by "arranging the elements of S in a definite order". Of course for S finite, if $f : S \to S$ is not 1-1 then the range of f is a proper subset of S, so f is not onto. Also if f is not onto then (by the Pigeonhole Principle!) it is not 1-1. Hence for finite S, f is 1-1 iff it is onto, while this is not so for infinite sets.

Thinking of permutations as functions also enables us to consider the case where S is the empty set. As in the case where S is infinite it is again not clear what an arrangement of the elements of the empty set might be. However in this case the only function from S to S is the "empty" function $\emptyset$ as we saw earlier in Section 4. Moreover the function $\emptyset$ is both 1-1 and onto. This is because the definitions of a function being 1-1 or onto are in the form of "if...then..." sentences where in the case of the empty function the "if" clause is false so the "if...then" sentences are true, i.e.: the condition of the definition is vacuously satisfied. Therefore we conclude that the number of permutations of the empty set is 1 and for this reason it is makes sense to define $0! = 1$ which is what we hereby do. Notice that we are following the same sort of reasoning as in Section 4 where we generalized exponentiation to arbitrary sets by thinking of S^T as the set of all functions $f : T \to S$. We can now consider a permutation of *any* set S as a 1-1 function of S onto S.

Having pointed out the virtues of considering permutations as functions we should emphasize that for our present purposes, where we are concerned with finite counting problems and hence finite and usually non-empty sets, we will usually think of the permutations of a set as simply arrangements of the set in a definite order (and hence we will only invoke the function definition if we need to be very precise.) An important generalization of this more primitive way of thinking of permutations is to consider all possible arrangements of r elements chosen from a set of n elements and where, of course, $r \leq n$. For example we might want to count the number of 3 letter "words" which can be formed using any three distinct letters of the 26 letters of the English alphabet. We refer to these arrangements as the *r-permutations of S*. Hence the permutations we discussed above are just the n-permutations. We can then also think of the r-permutations of a set S as the set of 1-1 functions with domain an r-set and which take their values in S. r-permutations are easily counted, again using the multiplication principle: as before there are n choices for the first position, $n-1$ for the second, etc., and finally $n-(r-1)$ for the r-th position. We denote by $P(n,r)$ the number of r-permutations of an n-element set. We have just established

Theorem 1.5.1 *For non-negative integers r and n with $r \leq n$,*

$$P(n,r) = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Notice that we obtain the theorem for the case $r = 0, n \geq 0$, $P(n,0) = 1$ by our

agreement that $0! = 1$. For $r > n$ we take $P(n, r) = 0$. Also the set of all ordinary permutations of S is counted by $P(n, n) = n!$ as before.

Frequently in enumerative combinatorics a set of n elements is, for conciseness, called an *n-set* and an r element subset is called an *r-subset*. With this terminology, $P(n, r)$ is the total number of arrangements of all of the r-subsets of an n-set and this is the same as the number of r-permutations of an n-set.

Example 1. The number of ways a president, vice president, and secretary can be chosen from among 10 candidates is $P(10, 3) = 10!/(10 - 3)! = 10 \cdot 9 \cdot 8 = 720$. If we were instead counting the number of ways of choosing 3 persons from among 10, without regard to order, we would want all 3! arrangements of any 3 people as a single choice, i.e.: we would want to partition the 720 arrangements into blocks of 3! each, and it is the index of this partition which we want. Hence the answer would be $720/3! = 120$.

Example 2. Suppose we wish to assign the integers 1,2,...,9 to all of the possible 9 positions obtained by leaving blank some 7 of the 16 positions of a 4×4 array. The number of ways in which such an assignment can be made is just the number $P(16, 9) = 16!/7!$ of 9-permutations of the 16 positions of the array.

The following example illustrates that ingenuity is often required is solving simple combinatorial problems.

Example 3. Find the number of ways of ordering the 26 letters of the alphabet so that no pair of the 5 vowels will be adjacent.

We think of this problem as consisting of performing two tasks independently and then apply the multiplication principle: first we observe that there are 21! ways of arranging the consonants; after having done this we observe that for each of these arrangements, to avoid adjacency, the vowels must be matched with exactly 5 of the 22 blank positions among the arranged consonants (22 since we are including the positions before the first and after the last). Hence there are $P(22, 5)$ ways of arranging the vowels among any arrangement of the consonants. By the multiplication principle there are then

$$21! \times P(22, 5) = \frac{21!22!}{(22 - 5)!} = \frac{21!22!}{17!}$$

possible arrangements altogether.

So far the permutations we have considered have been *linear* meaning that we think of the r elements chosen from S as being arranged along a line. But if they are arranged around a circle then the number will be smaller. For example if the elements a,b,c were chosen from S and placed on a circle, then the arrangements

bac, acb, cba, which are different in the linear case, would now all be considered the same since what matters is the positions occupied by a,b,c relative to each other, i.e.: there is no longer a first position. But once we have positioned one of the r elements on the circle, their relative positions are fixed as in the linear case. We call these arrangements *circular r-permutations*.

To count the number of circular r-permutations of an n-set, we simply observe that to pass from linear permutations to circular permutations we partition the linear permutations into equivalence classes, where two arrangements of an r-subset are placed in the same class provided they are indistinguishable when placed on a circle, that is if one can be obtained from another by a suitable number of "end around carries" Then the number of circular r-permutations is the index of this partition. But each class obviously consists of just the r linear permutations corresponding to the r choices of first positions of all of the circular r-permutations in the class. But now we have a set of size $P(n,r)$ partitioned into classes all of size r, so the index of the partition is just $P(n,r)/r$. This proves the following theorem.

Theorem 1.5.2 *The number of circular r-permutations of an n-set is*

$$\frac{P(n,r)}{r} = \frac{n!}{r(n-r)!}.$$

Example 4. How many bracelets can be made by using r beads per bracelet and choosing the beads from a set of n beads of different colors?

Since each bracelet can be turned over without change, the number of bracelets is just half the number of circular r-permutations of an n set, i.e.: $P(n,r)/2r$. Notice that turning a bracelet over amounts to reversing the relative positions of the beads, i.e.: two arrangements are considered the same if one is just the reverse of the other. For example using 3 beads, abc and cba are considered the same. This means that we first partition the linear permutations into classes of r permutations per class to obtain the circular permutations, then we further partition these classes into still larger classes, two classes of circular permutations per larger class for a total of $2r$ per larger class. The index of this partition is the number of bracelets.

A permutation is a set of objects arranged in a particular order. Now we shall consider sets of objects with no consideration of order. If S is an n-set, any r-subset, with no consideration of its order, is called an *r-combination* of S. In Example 1 above we first counted the number of ways of choosing a president, vice president, and secretary, from among 10 candidates. This was the number of 3-permutations of a 10-set, since order was important. We also counted the number of ways of simply choosing 3 persons; this was the number of 3-combinations of a 10-set. We denote the number of r-subsets of an n-set by the symbol $\binom{n}{r}$. This symbol is read as "n choose r". These symbols are also called the *binomial coefficients* because of their role in the binomial theorem, which we will discuss shortly. They were introduced in

the eighteenth century by Euler and for this reason are sometimes also called Euler symbols. The importance of counting the number of choices of r objects from n objects was understood in India in 300BC and the basic properties of the binomial coefficients were studied in China in the 12th century. From their definition we observe

$$\left(\begin{array}{c} n \\ r \end{array} \right) = 0 \quad \text{if } r > n, \text{ and hence } \left(\begin{array}{c} 0 \\ r \end{array} \right) = 0 \text{ if } r > 0$$

since in these cases there are no r-subsets of the n-set. On the other hand the following are easily seen to be true for each non-negative integer n:

$$\left(\begin{array}{c} n \\ 0 \end{array} \right) = 1, \left(\begin{array}{c} n \\ 1 \end{array} \right) = n, \left(\begin{array}{c} n \\ n \end{array} \right) = 1.$$

For the most important cases we have the following very basic theorem.

Theorem 1.5.3 *For integers* $0 \leq r \leq n$,

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{P(n,r)}{r!}$$

and therefore

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)}{r!}.$$

PROOF The r-subsets of an n-set S are obtained by partitioning the r-permutations of S into equivalence classes, each class consisting of those r-permutations of the same r elements of S. Since there are $r!$ such permutations, each class has $r!$ elements. Since $\left(\begin{array}{c} n \\ r \end{array} \right)$ is the index of this partition we have

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!}.$$

This proves the theorem. Another interesting way to prove the theorem is to apply the multiplication principle: first choose an r-subset from S in (by definition) $\left(\begin{array}{c} n \\ r \end{array} \right)$ ways. For each such choice, there are $r!$ ways to arrange the objects, so the multiplication principle yields $\left(\begin{array}{c} n \\ r \end{array} \right) \times r! = P(n,r)$ by the definition of $P(n,r)$.

Observe that the special cases obtained by taking $r = 0, 1$ or n which were computed above directly from the definition of $\left(\begin{array}{c} n \\ r \end{array} \right)$ can also be obtained, using the formula in this theorem, since we have already defined $0! = 1$ by an analysis using

sets.

Example 5. In studying Ramsey's theorem in Section 3 we were concerned with the sets K_r^t of all t-subsets of an n-set. We now know that K_r^t has $\binom{r}{t}$ elements. In particular $K_r = K_r^2$, the complete graph on r vertices has $\binom{r}{2} = r(r-1)/2$ edges.

Example 6. Suppose each of n houses is to be painted one of the 3 colors: white, yellow, or green, and also that w of the houses are to be white, y are to be yellow, and g are to be green. How many painting plans are there with this allocation of colors?

Note that $w + y + g = n$. From the n houses we first can choose the w white houses in $\binom{n}{w}$ ways. For each of these choices there are $n-w$ houses left and from these we can select those to be painted yellow in $\binom{n-w}{y}$. From the remaining $n-w-y = g$ houses we must select all g to be painted green, i.e.: we have no choice, so this can be done in only 1 way, (which, in fact, is the value of $\binom{g}{g}$.) Hence, by the multiplication principle we can paint the houses altogether in

$$\binom{n}{w} \cdot \binom{n-w}{y} \cdot 1 = \frac{n!}{w!(n-w)!} \cdot \frac{(n-w)!}{y!(n-w-y)!}$$

ways. But $n-w-y = g$ and the factors $(n-w)!$ cancel so we obtain the symmetric formula

$$\frac{n!}{w!y!g!}$$

for the total number of painting plans. Incidentally, since the original problem was symmetric in the 3 colors the answer had to be symmetric! More abstractly, we have computed the number of ways in which we can partition an n-set into blocks of sizes w, y, and g, and then label the block of sizes w, y, g white, yellow, and green respectively. For example, if $n = 4$ and the block sizes are 2,1, and 1, this can be done in $4!/2!1!1! = 12$ ways.

Now we make two simple observations about the symbols $\binom{n}{r}$. First, choosing an r-subset is precisely equivalent to choosing its complement, an $(n-r)$-subset. Second, since the total number of subsets of an n-set is 2^n, this is what we will get if we add up all of the sizes of the r-subsets for $r = 0, ..., n$. These two observations are the proof of the next theorem.

Theorem 1.5.4 *a) For integers* $0 \le r \le n$,

$$\binom{n}{r} = \binom{n}{n-r}.$$

b) For any non-negative integer n

$$\left(\begin{array}{c} n \\ 0 \end{array}\right) + \left(\begin{array}{c} n \\ 1 \end{array}\right) + \left(\begin{array}{c} n \\ 2 \end{array}\right) + \cdots + \left(\begin{array}{c} n \\ n \end{array}\right) = 2^n.$$

In proving this theorem we have used a very simple idea which is worth naming because it is so useful in combinatorics:

> *EQUALITY PRINCIPLE: count the elements of a set in two different ways and set the two results equal.*

This is exactly what we did to obtain both parts of Theorem 5.4, perhaps more strikingly in part *b)*. The equality principle will be useful many times in what follows and should be almost the first proof strategy to think of in trying to establish many combinatorial identities.

The theorem could also be checked by direct substitution into the formula in Theorem 5.3. This is trivial in the case of equation *a)* and is tedious in the case of equation *b)*.

Finally we establish the binomial theorem — since, as noted above the symbols $\left(\begin{array}{c} n \\ r \end{array}\right)$ are named for it.

Theorem 1.5.5 *(Binomial theorem) If n is a non-negative integer and x and y are any numbers, real or complex, then*

$$\begin{aligned} (x+y)^n &= x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 + \cdots + \\ &\quad \frac{n(n-1)\cdots(n-r+1)}{r!}x^{n-r}y^r + \cdots + y^n \\ &= x^n + \left(\begin{array}{c} n \\ 1 \end{array}\right)x^{n-1}y + \left(\begin{array}{c} n \\ 2 \end{array}\right)x^{n-2}y^2 + \cdots + \left(\begin{array}{c} n \\ r \end{array}\right)x^{n-r}y^r + \cdots + y^n. \end{aligned}$$

In summation notation,

$$(x+y)^n = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array}\right) x^{n-r}y^r.$$

PROOF Write $(x+y)^n = (x+y)(x+y)\cdots(x+y)$, a product of n factors, each equal to $(x+y)$. Use the distributive law of elementary algebra to completely multiply out this product and then group together like terms. In multiplying out, for each of the factors $(x+y)$ we can choose either an x or a y. Hence, by the multiplication principle we obtain 2^n terms altogether and each has the form $x^{n-r}y^r$ for some $r = 0, ..., n$. (Check this by expanding $(x+y)^n$ for $n = 3$ or 4.) That is we partition the set of all 2^n terms in the expansion of $(x+y)^n$ into $n+1$ equivalence classes, putting into each class all those terms which are equal to each other. How many terms are there in the

class all of whose terms equal $x^{n-r}y^r$? The answer is that we obtain each of these terms by choosing y in r of the n factors and (with no choice) by taking the remaining $n-r$ to be x. Hence the number of terms equal to $x^{n-r}y^r$ is just the number $\binom{n}{r}$ of r-subsets in an n-set. Therefore the sum of these terms is $\binom{n}{r}x^{n-r}y^r$. Adding over all of the equivalence classes we obtain the binomial theorem.

We will return to the study of the binomial coefficients in Section 7.

Exercises Section 1.5

1. In how many ways can 4 people sit at a round table, if two seating arrangements are distinguished iff a) at least one person occupies different chairs? b) at least one person does not have the same neighbor on the left and the same neighbor on the right for both arrangements?

2. In how many ways can six people each be assigned to one of 8 hotels if no two people are assigned the same hotel?

3. How many 4 digit numbers can be formed with the digits 2,5,8, and 9 if a) no repetitions are allowed? b) repetitions are allowed? c) repetitions are allowed, but the numbers must be divisible by 4?

4. A committee of 5 members is to be chosen from 10 Democrats and 6 Republicans and is not to contain a majority of Republicans. In how many ways can this be done?

5. Five persons are to be chosen from 10 persons. Exactly two of the persons are husband and wife. In how many ways can the selection be made a) if the wife must be chosen whenever her husband is chosen? b) if both husband and wife must be chosen whenever either is chosen?

6. Recall that for a real number x, $[x]$ is the greatest integer $\leq x$. Show that a) the total number of r-permutations of an n-set, $n \geq 1$, for all $r \leq n$ is

$$\left(1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!}\right) n!$$

and that b) the expression in a) is equal to $[n! \cdot e]$.

7. In how many ways can a set of 8 marbles be divided to form a) a set of 2 and a set of 6? b) a set of 2 and a set of 5? c) two sets of 4 each?

8. Earlier we defined a binary relation on a set A to be simply a subset of $A \times A$. Now let us extend this definition and say that R is a binary relation *between* sets A

and B if R is a subset of $A \times B$. The *domain* of R is the subset of elements $a \in A$ for which $(a, b) \in R$ for some $b \in B$. The *range* of R is the subset of all $b \in B$ for which $(a, b) \in R$ for some $a \in A$. Hence if we think of A as the x-axis, B as the y-axis, and R as a subset of the plane, then the domain and range of R are just the projections of R into the axes. [A function $f : A \to B$ is then just a relation between A and B with domain A and the property that if $(a, b), (a, c) \in f$ then $b = c$.]

a) How many relations are there with the set $\{1, 2, 7\}$ as domain and the set $\{0, 5\}$ as range? How many functions?

b) How many relations are there whose domain is a subset of $\{1, 2, 3, 4\}$ and whose range is a subset of $\{5, 6, 7\}$? How many functions?

c) How many functions are there whose domain is $\{1, 2, \ldots, n\}$ and whose range is a subset of $\{0, 1\}$? How many relations?

9. In how many ways can 6 ladies and 6 gentlemen be seated around a circular table if the ladies and gentlemen are to occupy alternate seats?

10. How many sets of 3 integers each can be formed from the integers $\{1, 2, 3, \ldots, 20\}$ if no two consecutive integers are to be in the same set?

11. There are 100 students at a school and three dormitories A, B, and C with capacities 25, 35, and 40, respectively. a) How many ways are there to fill up the dormitories? b) Suppose that of the 100 students 50 are men and 50 are women and that A is an all men dorm, B is an all women dorm and C is co-ed. how many ways are there to fill up the dormitories?

12. Suppose there are n applicants for a job and three interviewers are asked to independently rank each of the applicants. After interviewing all of the applicants the interviewers present their results in a set of reports, each report consisting of their respective rankings of one of the candidates. It is decided that an applicant will be hired if he is ranked first by at least two of the three interviewers. Show that the fraction of the total number of possible reports which lead to the hiring of some applicant is

$$\frac{3n - 2}{n^2}.$$

Hint: Find the fraction of reports which do *not* lead to acceptance and subtract this number from 1. To do this notice that a particular report leads to the rejection of an applicant iff all three interviewers put a *different* applicant in first place and this can happen in $n(n - 1)(n - 2)$ different ways.

13. Find how many 8 letter "words" can be formed using the 26 letters of the alphabet if each word contains either 4 or 5 vowels. Each letter can occur any number of times in a word.

14. A teacher anticipates teaching the same course for the next 35 years. In order not to become bored with his jokes, he decides to tell exactly three distinct jokes every year and in no two different years to tell exactly the same three jokes. a) What is the minimum number of jokes that will accomplish this? b) What is the minimum number if he decides to never tell the same joke twice?

15. For given non-negative integers k and n with $k \leq n$, how many subsets of an n-set have fewer than k elements?

16. Find n if n is an integer for which $\begin{pmatrix} n \\ 12 \end{pmatrix} = \begin{pmatrix} n \\ 8 \end{pmatrix}$.

17. Prove that

$$\begin{pmatrix} n \\ 0 \end{pmatrix}^2 + \begin{pmatrix} n \\ 1 \end{pmatrix}^2 + \begin{pmatrix} n \\ 2 \end{pmatrix}^2 + \cdots + \begin{pmatrix} n \\ n \end{pmatrix}^2 = \begin{pmatrix} 2n \\ n \end{pmatrix}.$$

Do this in two ways using the following suggestions. a) Find the coefficient for $x^n y^n$ for each member of the identity $(x+y)^n(x+y)^n = (x+y)^{2n}$. b) Use the definition of $\begin{pmatrix} 2n \\ n \end{pmatrix}$ together with Theorem 5.4 a).

18. What is the coefficient of x^3y^4z in the expansion of $(x+y+z)^8$?

19. Suppose k and n are integers with $0 \leq k \leq n$. Explain why $n(n-1)(n-2)\cdots(n-k+1)$ is divisible by $k!$.

1.6 Permutations and combinations with repetitions

Sometimes we wish to consider r-subsets of an n-set S in which repeated elements from S occur. The following example concerns permutations and illustrates the kind of situation in which this might arise.

Example 1. Suppose we wish to arrange a group of 13 students in a line and that 3 of the students are freshmen, 5 are sophomores, 4 are juniors, and 1 is a senior. If we wish to distinguish between the students, then the number of ways of arranging the students is just the number of permutations of 13 objects, which is 13!. Now however, suppose we do not wish to distinguish between individual students but only among their classes, that is suppose the 3 freshmen are considered interchangeable, as are the 5 sophomores, etc. Let $P(13; 3, 5, 4, 1)$ denote the number of different ways the students can be arranged in this way. Each such arrangement is determined by

designating which of the 13 positions in the line are occupied by the 3 freshmen, which are occupied by the 5 sophomores, etc. Then for each such arrangement, the 3 freshmen can be arranged in 3! ways if we decide to distinguish between them. Thus by the multiplication principle there are $P(13;3,5,4,1)\cdot 3!$ ways of arranging the students if we wish to distinguish between the individual freshmen but not between the members of the other classes. Continuing in the this way, if we wish to distinguish between the individual members of each class, by the multiplication principle we obtain $P(13;3,5,4,1)\cdot 3!\cdot 5!\cdot 4!\cdot 1!$ total ways of arranging the 13 students in line. By the equality principle this must equal 13! Hence we have

$$P(13;3,5,4,1)=\frac{13!}{3!5!4!1!}.$$

This example can be generalized. Doing so we obtain the following theorem.

Theorem 1.6.1 *Suppose S is an n-set and $\{S_1,\ldots,S_k\}$ is a partition of S. If $|S_i|=r_i$ for $i=1,\ldots,k$, so that $r_1+\cdots+r_k=n$, then the number of permutations of S in which the elements of each S_i, $i=1,\ldots,k$, are not distinguished, is equal to*

$$P(n;r_1,\ldots,r_k)=\frac{n!}{r_1!r_2!\cdots r_k!}.$$

The proof of the theorem is a natural and simple extension of the proof given in the example, and hence we can omit it. Notice that if each of the blocks S_i of the partition of S consists of just a single element, i.e.: the index of the partition is n, then the formula of the theorem becomes $P(n;1,1,\ldots,1)=n!$, just the number of permutations of S, which is what we expect. Also, if we extend the definition of a partition to include empty blocks, so that some of the $r_i=0$, the theorem is still obviously true.

Notice that Example 6 of Section 5 (painting each of n houses one of 3 colors) was essentially the same as Example 1 above, except for the specific details. Also the formula obtained in that example was derived using binomial coefficients. Theorem 6.1 can also be derived in an analogous way; we leave this as an exercise.

If we think of either of these examples and the general statement of the theorem we see that if the numbers $r_1,\ldots,r_k$ are all different, we have actually just counted the number of ways in which the n-set S can be partitioned with prescribed index k and block sizes $r_1,\ldots,r_k$. If some of the r_i are equal then $P(n;r_1,\ldots,r_k)$ is larger than the number of partitions. For example, if we paint 7 houses, 3 white, 2 yellow, and 2 green, then the formula of Example 6, Sect. 5, tells us there are

$$\frac{7!}{3!2!2!}$$

painting plans. In particular, for example, if we paint the first 3 houses white, houses 4 and 5 yellow, and houses 6 and 7 green, this counts as a different painting plan

from that which is the same except that houses 4 and 5 are painted green, and houses 6 and 7 are yellow. These two different painting plans do, however, yield the same partition of the 7 houses. Thus we need to divide the number by 2! (the number of ways of arranging the 2 blocks of equal size) to obtain the number of partitions. In general we need to divide $P(n; r_1, \ldots, r_k)$ by $j!$ for any set of j repeated block sizes, to obtain the number of partitions. For example if r, s, t are distinct integers and $3r + 4s + t = n$, then the number of partitions of an n-set of index 8 and block sizes r, r, r, s, s, s, s, t is

$$\frac{n!}{r!r!r!s!s!s!s!t!3!4!}.$$

For this reason if all of the r_i are positive, we could refer to the $P(n; r_1, \ldots, r_k)$ by some such name as the numbers of *ordered partitions* of an n-set with the prescribed index k and block sizes $r_1, \ldots, r_k$. Most often, however, this is not done and the $P(n; r_1, \ldots, r_k)$, with all $r_i \geq 0$, are called the *multinomial coefficients* for the reason to be explained shortly.

Example 2. $P(6; 2, 2, 2) = 6!/2!2!2! = 90$. This is what we suggested above calling the number of *ordered partitions* of index 3 with each block size 2. On the other hand the number of *partitions* of index 3 with each block size 2 is $P(6; 2, 2, 2)/3! = 15$. Another way to count the number of partitions in this simple case is to observe that if the set to be partitioned is $\{1, 2, 3, 4, 5, 6\}$, a partition is completely determined by which of the five remaining elements is to be paired with 1 and, for each of these five choices, by determining how the remaining four elements can be partitoned into 2 blocks of size 2. This is determined by deciding for some one of the set of four, which of the remaining three to pair with it. Again we have $5 \times 3 = 15$.

Still another way of thinking of Theorem 6.1 is to consider the original n-set S to be what in combinatorics is usually called an *n-multiset*, i.e.: a set consisting of only k *distinct* objects but object 1 occurs r_1 times, object 2 occurs r_2 times,..., and object k occurs r_k times, for a total of n elements. We then denote S by

$$S = \{r_1 \cdot 1, r_2 \cdot 2, \ldots, r_k \cdot k\}$$

and refer to the r_i as the *repetition numbers* of the n-multiset S. This notation is in rather flagrant violation of the usual practice of only listing the elements of a set once in using the braces notation $\{\ldots\}$ to describe the set. For example the set $\{a, b\}$ would never be described by $\{a, a, b\}$. But in enumerative combinatorics we would even describe it as $\{2 \cdot a, 1 \cdot b\}$. So long as we observe that we are actually only describing a partition of S where we identify elements which we wish to consider indistinguishable, we are completely consistent with ordinary set theory notation. Notice that the notation

$$S = \{r_1 \cdot 1, \ldots, r_k \cdot\}$$

for an n-multiset S requires implicitly that $r_1 + \cdots + r_k = n = |S|$.

In the terminology of multisets we can restate Theorem 6.1 still another way:

Corollary 1.6.1 *The number of n-permutations of an n-multiset*

$$S = \{r_1 \cdot 1, \ldots, r_k \cdot k\}$$

is

$$P(n; r_1, \ldots, r_k) = \frac{n!}{r_1! r_2! \cdots r_k!}.$$

Example 3. The number of permutations of the letters of the word "mississippi" is just the number of 11-permutations of the multiset $S = \{1 \cdot m, 4 \cdot i, 4 \cdot s, 2 \cdot p\}$ and hence is

$$\frac{11!}{1!4!4!2!}.$$

The definition and notation for multisets suggests that we generalize the present discussion in two significant ways. First we can consider r-permutations of an n-multiset for $r < n$ and second we can allow some of the repetition numbers be infinite. If we consider r-permutations for $r < n$, we can no longer find a simple formula for the number of r-permutations, though there are general strategies for solving such problems. We will discuss one of these below. On the other hand the special case where all of the repetition numbers are infinite, or are all at least as large as r, is particularly easy and, in fact we have already seen the answer before:

Theorem 1.6.2 *If the multiset*

$$S = \{r_1 \cdot 1, \ldots, r_k \cdot k\}$$

is such that the repetition numbers r_i are all at least as large as r, then the number of r-permutations of S is k^r.

Proof To construct an r-permutation of S we can choose the first element element of the r-set to be any one of the k types of elements of S and for the second element, since each $r_i \geq r$, we can again independently choose any one of the k types, etc., so by the multiplication principle we can construct an r-permutation in k^r ways. The independence of each of our choices is guaranteed since the condition $r_i \geq r$ assures us that we will never run out of any of the k possible choices of the r elements.

We should also recall that the number of functions from an r-set into a k-set is also k^r and, in fact, constructing such a function is precisely equivalent to constructing an r-permutation from a multiset consisting of unlimited numbers of k types of elements. At the other extreme, as we noticed earlier, if the repetition numbers are all 1's, so the multiset becomes an ordinary set, then we are simply counting the number of 1-1 functions of the r-set into a k-set and this is just $P(n, r)$. This still leaves the general case where some, but not all, of the repetition numbers are infinite or where they are all finite but with sum less than r. As noted above, in these cases there is no general formula, but there are general strategies. One approach is (as is almost always the

case), to attempt to divide and conquer. The following example illustrates a case where, in particular, the partitioning principle applies:

Example 4. Find the number of 8-permutations of the multiset $S = \{3\cdot a, 2\cdot b, 4\cdot c\}$.

We notice that $|S| = 9$ so we can conveniently partition the total number of 8-permutations into 3 blocks:

a) those missing a single a. In this case we are simply counting the 8-permutations of $\{2\cdot a, 2\cdot b, 4\cdot c\}$. Here the sum of the repetition numbers is 8 so Theorem 6.1 applies and we count the number in this block to be

$$\frac{8!}{2!2!4!} = 420.$$

b) those missing a single b. Just as in a), the number of these is

$$\frac{8!}{3!1!4!} = 280.$$

c) those missing a single c. These number

$$\frac{8!}{3!2!3!} = 560.$$

By the partitioning principle the total number of 8-permutations of S is therefore $420 + 280 + 560 = 1260$.

If we had wanted to find the number of 7-permutations of S by partitioning and then applying Theorem 6.1, though not impossible, the solution would have involved a partition of index 6. (Why?) To count the 6-permutations this way would become rather tedious. (Another approach called the method of generating functions is usually better. A discussion of this method can be found in most texts on combinatorics.)

Now let us return to he formula of Theorem 6.1, and in particular, let us think of it as counting what we tentatively called *ordered partitions* of prescribed index and block sizes. The formula is reminiscent of the formula

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!},$$

for the binomial coefficient and, in fact if the index of the partition of S is 2 and one of the blocks has size r, then the other must have size $n - r$ and the two formulas coincide. (Notice that in this case we obtain the number of ordinary (unlabeled) partitions with block sizes r and $n - r$ only if $n \neq 2r$; if $n = 2r$ we have twice the number of partitions.) This suggests we now denote the numbers $P(n; r_1, \ldots, r_k)$ by

$$\left(\begin{array}{c} n \\ r_1\ r_2\ \cdots\ r_k \end{array} \right)$$

and call them the *multinomial* coefficients. The reason for this terminology is that they appear as the coefficients in the complete expansion of the multinomial $(x_1 + \cdots + x_k)^n$ and this generalizes the binomial theorem. This is what we prove now.

Theorem 1.6.3 *(Multinomial theorem) If n is a non-negative integer and $x_1, \ldots, x_k$ are any numbers, real or complex, then*

$$(x_1 + \cdots + x_k)^n = \sum_{r_1 + \cdots + r_k = n} \left(\begin{array}{c} n \\ r_1\ r_2\ \cdots\ r_k \end{array} \right) x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k},$$

where the notation indicates that the sum is taken over all k-tuples of non-negative integers with sum n.

Proof We just generalize the proof of the binomial theorem. As before we write $(x_1 + \cdots + x_k)^n = (x_1 + \cdots + x_k)(x_1 + \cdots + x_k) \cdots (x_1 + \cdots + x_k)$, a product of n factors. Using the distributive law of elementary algebra we completely multiply out this product and then group together like terms. Hence by the multiplication principle we obtain k^n terms altogether and each has the form $x_1^{r_1} \cdots x_k^{r_k}$ where $r_1 + \cdots + r_k = n$. (Again, as in the proof of the binomial theorem, Theorem 5.5, check this by a test case.) That is, we partition the set of all k^n terms into disjoint classes, each class corresponding to an ordered k-tuple $r_1, \ldots, r_k$ with sum n, i.e.: one class for each $x_1^{r_1} \cdots x_k^{r_k}$. How many terms are there in this class? The answer is that we obtain each such term in the class by choosing x_1 in r_1 of the factors $(x_1 + \cdots + x_k)$, x_2 in r_2 of them, etc., and x_k in r_k of the factors, and this is just the number

$$\left(\begin{array}{c} n \\ r_1\ r_2\ \cdots\ r_k \end{array} \right)$$

of n-permutations of the n-multiset $\{r_1 \cdot 1, \ldots, r_k \cdot k\}$. For example, in the expansion of $(x + y + z + w)^4$, the term x^2yz occurs

$$\left(\begin{array}{c} 4 \\ 2\ 1\ 1\ 0 \end{array} \right) = \frac{4!}{2!1!1!0!} = 12$$

times. (Check this by observing that there are $\left(\begin{array}{c} 4 \\ 2 \end{array} \right) = 6$ ways of choosing 2 x's from the 4 factors and for each of these there are 2 ways to choose a y and a z from the remaining 2 factors, and hence $6 \times 2 = 12$ ways to obtain x^2yz.)

COMBINATIONS OF MULTISETS Now let us turn to the problem of counting the number of *unordered* selections of r elements from a set, allowing repetitions. This means we are counting the *r-combinations* (which is the same as the r-subsets) of a multiset S. For example, if $S = \{2 \cdot a, 1 \cdot b, 3 \cdot c\}$ then the 5-combinations of S are just the three,

$$\{1 \cdot a, 1 \cdot b, 3 \cdot c\},\ \{2 \cdot a, 3 \cdot c\},\ \{2 \cdot a, 1 \cdot b, 2 \cdot c\},$$

while there are, (as in Example 4, above),

$$\frac{5!}{1!1!3!} + \frac{5!}{2!3!} + \frac{5!}{2!1!2!} = 60$$

5-permutations.

The most important theorem about r-combinations of multisets is the following, which counts the number when the repetition numbers are all at least as large as r, and in particular if all are infinite.

Theorem 1.6.4 *Let $S = \{r_1 \cdot 1, \ldots, r_n \cdot n\}$ be a multiset consisting of n distinct objects, each having repetition number $r_i \geq r$. Then the number of r-combinations of S is*

$$\left(\begin{array}{c} n+r-1 \\ r \end{array} \right) = \left(\begin{array}{c} n+r-1 \\ n-1 \end{array} \right).$$

Another (and more common) way of stating the theorem is

> *The number of ways of choosing r objects from n objects with repetitions allowed is*
>
> $$\left(\begin{array}{c} n+r-1 \\ r \end{array} \right).$$

Notice that the equality of the two binomial coefficients in the theorem is from Theorem 5.4 *a)*.

Proof The theorem is so important we give two instructive proofs. The first proof is geometric. Suppose a rectangle is divided by vertical and horizontal lines so there are r rectangles in each column and $n-1$ rectangles in each row; thus there are $r+1$ horizontal lines and n vertical lines. (Draw a picture for both the general case and, say, $r = 4$ and $n = 6$.) A path from the lower left corner to the upper right corner consists of $n+r-1$ segments. Since there are r vertical segments, the number of paths is equal to the number of subsets of r objects that can be chosen from $n+r-1$ objects, i.e.: $\left(\begin{array}{c} n+r-1 \\ r \end{array} \right)$. However, there are n positions in which the r vertical segments can be placed: any of the n vertical lines. The choice of path is determined by the choice of r of these positions from among the n, where repetitions

are allowed. Therefore the total number of ways of choosing r objects from n objects, with repetitions allowed is equal to the total number of paths, and we just observed that this is $\begin{pmatrix} n+r-1 \\ r \end{pmatrix}$. Hence the equality principle gives the result.

Second proof. We first observe that an r-combination of S has the form of a multiset $\{x_1 \cdot 1, \ldots, x_n \cdot n\}$ with each x_i a non-negative integer and $x_1 + \cdots + x_n = r$. Conversely, every ordered n-tuple of non-negative integers $x_1, \ldots, x_n$ with $x_1 + \cdots + x_n = r$ determines an r-combination of S. Therefore the number of r-combinations of S equals the number of non-negative integer solutions of

$$x_1 + \cdots + x_n = r.$$

(Each solution, also called a solution set, is understood to be an n-tuple $(x_1, \ldots, x_n)$ of non-negative integers.)

We will complete the proof by showing that the number of these solutions is also given by the number of permutations of the multiset

$$T = \{r \cdot x, (n-1) \cdot |\}$$

which, by Theorem 6.1, is

$$\frac{(r+(n-1))!}{r!(n-1)!} = \begin{pmatrix} n+r-1 \\ r \end{pmatrix}.$$

To see this, observe that if we are given a permutation of T, the $(n-1)$ |'s, arranged in a row, partition the r x's into n blocks: the block to the left of the first |, the blocks between adjacent |'s, and the block to the right of the last |. Some of the blocks may be empty. Let there be x_1 x's to the left of the first |, x_2 x's between the first and second |, etc., and finally x_n x's to the right of the last |. Then $x_1, \ldots, x_n$ are non-negative integers solving $x_1 + \cdots + x_n = r$. Conversely, given a solution set of $x_1 + \cdots + x_n = r$, we can obviously reverse these steps and construct a permutation of T. Hence the solution sets of $x_1 + \cdots + x_n = r$ are in 1-1 correspondence with the permutations of T. For example if $n = 5$ and $r = 6$,

$$|xxx||xx|x$$

is the permutation of $T = \{6 \cdot x, (5-1) \cdot |\}$ corresponding to the solution set $x_1 = 0, x_2 = 3, x_3 = 0, x_4 = 2, x_5 = 1$ of $x_1 + \cdots + x_5 = 6$. This completes the proof.

Example 5. A store sells 6 different kinds of cookies and has on hand at least a dozen of each kind. In how many ways can a dozen cookies be purchased?

Since we assumed that the store has at least a dozen of each kind of cookie on hand, then an order for a dozen cookies is just a 12-combination (since the order of

the cookies is unimportant) of the multiset $\{r_1 \cdot 1, \ldots, r_6 \cdot 6\}$ and this is

$$\left(\begin{array}{c} 6+12-1 \\ 12 \end{array} \right) = \left(\begin{array}{c} 17 \\ 12 \end{array} \right) = 6188.$$

If a bakery has n cookie cutters and wishes to make a batch of r cookies, how many different kinds of batches can it make? Since the bakery can use any of the cutters to make any of the cookies and since it can use any cutter as often as as it wants, the answer is $\left(\begin{array}{c} n+r-1 \\ r \end{array} \right)$.

In the second proof of Theorem 6.4 we established a 1-1 correspondence between the non-negative integer solutions of $x_1 + \cdots + x_n = r$ and the number of r combinations of a multiset S consisting of n different objects with infinite repetition numbers. Isolating out this important fact we have

Corollary 1.6.2 *The number of n-tuples $x_1, \ldots, x_n$ which are non-negative integer solutions of*

$$x_1 + \cdots + x_n = r$$

is

$$\left(\begin{array}{c} n+r-1 \\ r \end{array} \right).$$

It is easy to extend this corollary to count the number of solutions of $x_1 + \cdots + x_n = r$ in which the values of the x_i are bounded below by integers other than 0. The following example is sufficiently generic to make the method clear.

Example 6. Find the number of integer solutions of

$$x_1 + x_2 + x_3 = 12$$

in which $x_1 \geq 1,\ x_2 \geq 0,\ x_3 \geq 3$.

To solve this we introduce new variables

$$y_1 = x_1 - 1,\ y_2 = x_2,\ y_3 = x_3 - 3$$

so that the set of conditions on the x's is equivalent to the condition that the y's all be non-negative. Substituting, the original equation becomes

$$y_1 + y_2 + y_3 = 8$$

and by the corollary the number of non-negative solutions of this equations is

$$\left(\begin{array}{c} 3+8-1 \\ 8 \end{array} \right) = 45$$

so this is the number of non-negative solutions of the original equation.

Therefore, for example, if a store sells 3 kinds of cookies, then an order for a dozen cookies in which at least 1 of the first type and at least 3 of the third type occur, can be made up in 45 different ways.

A more general problem is to find the number of solutions of

$$x_1 + \cdots + x_n = r$$

where the x_i are to be integers satisfying

$$a_1 \leq x_1 \leq b_1, \ldots, a_n \leq x_n \leq b_n.$$

Making a change of variable we can replace all of the a_i by 0. Then we are left with the problem of counting r combinations of a multiset with repetition numbers which cannot be taken to be infinite. We will show later (Section 8) how the principle of inclusion and exclusion can be applied to solve this problem.

Exercises Section 1.6

1. Prove Theorem 6.1 by, using binomial coefficients, as in Example 6, Sect. 5.

2. By partitioning, find the number of 7-permutations of the multiset

$$S = \{3 \cdot a, 2 \cdot b, 4 \cdot c\}.$$

3. Let the multiset $S = \{r_1 \cdot 1, \ldots, r_k \cdot k\}$ where $r_1 = 1$. Let $n = r_2 + \cdots + r_k$. Show that the number of *circular* permutations of S is

$$\frac{n!}{r_2! \cdots r_k!}.$$

4. Show that the total number of combinations of all possible sizes of the multiset

$$S = \{r_1 \cdot 1, \ldots, r_k \cdot k\}$$

is $(r_1 + 1)(r_2 + 1) \cdots (r_k + 1)$.

5. How many integer solutions of

$$x_1 + x_2 + x_3 + x_4 = 30$$

satisfy $x_1 \geq 2, x_2 \geq 0, x_3 \geq -5, x_4 \geq 8$?

6. Find the number of r-combinations of the multiset $\{1 \cdot 1, \infty \cdot 2, \ldots, \infty \cdot n\}$.

7. How many different sums can one make from pennies, nickels, quarters, and five dollar bills, if exactly 4 pieces of money must be used?

8. How many different baskets of fruit, a dozen pieces of fruit per basket, can be made using oranges, apples, and bananas?

9. One throws r coins and s dice all at once, the coins being indistinguishable and the dice being indistinguishable. How many outcomes are distinguishable?

10. Show that the number of ways of choosing r or fewer objects from n with repetitions allowed is equal to $\left(\begin{array}{c} n+r \\ r \end{array} \right)$. Hint: show that the number of ways of choosing exactly r objects from $n+1$ objects with repetitions allowed is equal to the number of ways of choosing r or fewer objects from n objects with repetitions allowed.

1.7 The binomial coefficients

We have already seen that the numbers $\left(\begin{array}{c} n \\ r \end{array} \right)$, which are defined to be the number of r-element subsets of an n-element set, occur in several combinatorial contexts, in particular as the coefficients in the binomial theorem — from which they take their most common name. Because they occur so often in such a wide variety of surprisingly different mathematical situations, we devote this section to studying a few more of their elementary properties.

We begin by recalling that in addition to the formula

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!}$$

for their direct computation (Theorem 5.3), we also know, from their definition, that for all non-negative integers n, r,

$$\left(\begin{array}{c} n \\ r \end{array} \right) = 0 \text{ if } r > n, \text{ and } \left(\begin{array}{c} n \\ 0 \end{array} \right) = 1, \left(\begin{array}{c} n \\ n \end{array} \right) = 1.$$

We shall refer to the second and third of these values as the *boundary values* of the binomial coefficients. The computational formula applies only within these values, that is for $r = 0, \ldots, n$. We also know that for $0 \leq r \leq n$,

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \left(\begin{array}{c} n \\ n-r \end{array} \right)$$

since choosing an r-subset is equivalent to choosing its complement. We shall refer to this formula as the *symmetry* property. We use this terminology since if n is even and if we arrange the binomial coefficients in a row from $r = 0$ to $r = n$, then the formula says that they are symmetric about the "middle" coefficient $\begin{pmatrix} n \\ \frac{n}{2} \end{pmatrix}$. If n is odd then there is no middle coefficient, but

$$\begin{pmatrix} n \\ \left[\frac{n}{2}\right] \end{pmatrix} = \begin{pmatrix} n \\ \left[\frac{n}{2}\right] + 1 \end{pmatrix},$$

and they are symmetric about this common value. The next theorem is important since it provides an important *recursion formula* satisfied by the binomial coefficients.

Theorem 1.7.1 *For all integers r and n with $1 \leq r \leq n$,*

$$\begin{pmatrix} n \\ r \end{pmatrix} = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} + \begin{pmatrix} n-1 \\ r \end{pmatrix}.$$

Proof We could prove this identity by simply substituting in the formula of Theorem 5.3, but it is more instructive to prove it directly from the definition of $\begin{pmatrix} n \\ r \end{pmatrix}$. To do this let S be an n-set and fix an arbitrary element x of S. This is possible since we are assuming $n \geq 1$. Next we partition the set of all r-subsets of S into two disjoint blocks: a) those which contain x, and b) those which do not. r-subsets of type a) are chosen by choosing the remaining $r-1$ of their elements from $S - \{x\}$ and this can be done in $\begin{pmatrix} n-1 \\ r-1 \end{pmatrix}$ ways. r-subsets of type b) are chosen by choosing all r of their elements from $S - \{x\}$ and this can be done in $\begin{pmatrix} n-1 \\ r \end{pmatrix}$ ways. By the partitioning principle we conclude that $\begin{pmatrix} n \\ r \end{pmatrix}$ is just the sum of these two numbers. Notice that in the case $n = 1$ where $S = \{x\}$, we must have $r = 1$ and 1-subsets of type a) are obtained by choosing the empty set from the empty set which can be done in one way. 1-subsets of type b) are the 1-subsets of the empty set so there are none of these.

The significance of the recursion formula is that together with the boundary values, it enables us to obtain all of the binomial coefficients without using the direct computation of Theorem 5.3. The scheme, called *Pascal's triangle* is pictured below. The 1's bordering the triangle are the boundary values and, if we number the rows by $n = 0, 1, 2 \ldots$, starting from top, then the recursion formula of Theorem 7.1 says that we obtain the non-boundary entry in the r-th position of the n-th row by adding the two adjacent entries in the preceeding row. In fact, since the binomial coefficient

$\binom{n}{r} = 0$ for $r > n$ and if we *define* it to be 0 for $r < 0$, then we can generate even the boundary 1's (except for $\binom{0}{0} = 1$) by the recursion formula.

The Pascal triangle

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & & & & . & & . & & . & & \\
\end{array}
$$

Many of the relations satisfied by the binomial coefficients can be pictured in Pascal's triangle. For example Theorem 5.4 *b)*,

$$\binom{n}{0} + \cdots + \binom{n}{n} = 2^n$$

simply says that the sum of the entries in the n-th row is always 2^n.

If we look at the triangle we see that not only are the entries in each row symmetric, but also they increase from the left up to the "middle" and then decrease to the right. This is called the *unimodality* property. We can prove this fact in general (not just for the rows pictured above) by considering the ratio of two successive entries,

$$\frac{\binom{n}{r+1}}{\binom{n}{r}} = \frac{n!}{(r+1)!(n-r-1)!} \cdot \frac{r!(n-r)!}{n!} = \frac{n-r}{r+1}.$$

The numbers increase as long as the ratio is greater than 1, i.e.: as long as $n-r > r+1$ which means $r < \frac{1}{2}(n-1)$. For n even this means they increase up to the middle coefficient $\binom{n}{\frac{n}{2}}$ and decrease thereafter, as pictured. For n odd they increase up to the equal middle pair,

$$\binom{n}{\left[\frac{n}{2}\right]} = \binom{n}{\left[\frac{n}{2}\right]+1}$$

and decrease thereafter.

Here is a somewhat surprising connection between the binomial coefficients and the Ramsey numbers. Recall that in proving the 2-color version of Ramsey's theorem for graphs we actually established the upper estimate

$$r(n,r) \leq r(n-1,r) + r(n,r-1)$$

for the Ramsey number $r(n,r)$. This recursive upper bound leads to the following table of Ramsey number estimates.

r	$=$	2	3	4	5	6	7
$r(2,r)$	$=$	2	3	4	5	6	7
$r(3,r)$	$\leq$	3	6	10	15	21	28
$r(4,r)$	$\leq$	4	10	20	35		
$r(5,r)$	$\leq$	5	15	35			

(We are replacing m and n used in Section 3 to n and r respectively. Also we have included the trivial Ramsey numbers $r(2,r)$ and filled out a little more of the table.) Now the recursive upper bound for the Ramsey numbers looks a little like the recursive formula for the binomial coefficients. In fact, if we rotate the table of Ramsey number estimates clockwise through 45 degrees so that the diagonals of the table becomes rows, we see that we can obtain any entry in what have now become the rows by adding the two adjacent entries of the row above — just like the Pascal triangle! For example from $r(3,5) \leq 15$ and $r(4,4) \leq 20$ we obtain $r(4,5) \leq 35$. In fact we can easily show that we have the following upper bound for the Ramsey numbers:

$$r(n,r) \leq \left(\begin{array}{c} n+r-2 \\ r-1 \end{array} \right).$$

We leave the detailed proof of this as an exercise. Of course, as we noted in Section 3, this is not a very small upper bound for $r(n,r)$. The determination of smaller upper bounds than this is a major subject of research in Ramsey theory.

There are some other interesting combinatorial functions which satisfy recursive formulas similar to that for the binomial coefficients. We consider one such function. For non-negative integers n, r, define $S(n,r)$ to be the number of partitions of an n-set having index r. Just as in the proof of Theorem 7.1, for $n > 1$ fix an arbitrary element x of the n-set S. Then the set of all partitions of S of index r are of two disjoint types: a) those in which $\{x\}$ is one of the blocks, and b) those in which the block containing x properly contains $\{x\}$. Partitions of type a) consist of all of the partitions of $S - \{x\}$ of index $r-1$ together with $\{x\}$ as the r-th block; there are $S(n-1,r-1)$ of this type. For partitions of type b), these can all be obtained by taking one of the $S(n-1,r)$ partitions of $S - \{x\}$ of index r and placing x in one of the r blocks. Therefore by the multiplication principle there are $r \cdot S(n-1,r)$ partitions of type b). Hence by the addition principle we have

Theorem 1.7.2 *For non-negative integers n and r with $1 \leq r \leq n$, $S(n,r)$, the the number of partitions of an n-set of index r satisfies*

$$S(n,r) = S(n-1,r-1) + r \cdot S(n-1,r).$$

The $S(n,r)$ are called the *Stirling numbers of the second kind*. This must mean that there are Stirling numbers of the first kind and indeed there are; we shall encounter them later. Just as in the case of the binomial coefficients, Theorem 7.2 enables us to compute all of the Stirling number from the appropriate boundary conditions. For the Stirling numbers these are:

$$S(n,n) = 1 \text{ for } n = 0,1,2,\ldots, \text{ and } \; S(n,0) = 0, \; S(n,1) = 1 \text{ for } n > 0.$$

These values are immediate from the definition. We also have that $S(n,r) = 0$ for $n < r$ directly from the definition. We have taken $S(0,0) = 1$ as a convention to make the recursion formula correct when this case arises. (In fact, by slightly reformulating our original definition of the index of a partition to allow index 1 if the sole block is empty, one can also *prove* $S(0,0) = 1$. Just show that the empty relation is an equivalence relation of index 1 on the empty set.)

For the remainder of this section we will establish some additional identities satisfied by the binomial coefficients. An easy one is

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n}{r} \left(\begin{array}{c} n-1 \\ r-1 \end{array} \right), \quad n,r \geq 1.$$

It is immediate from the formula

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n(n-1)\cdots(n-r+1)}{r(r-1)\cdots 2 \cdot 1}.$$

Besides combinatorial arguments, another way to obtain interesting identities for the binomial coefficients is to appeal to the binomial theorem. For example if we start with the special case

$$(1+x)^n = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) x^r,$$

and set $x = 1$ we obtain Theorem 5.4 *b)* again. On the other hand, setting $x = -1$ yields

$$\left(\begin{array}{c} n \\ 0 \end{array} \right) - \left(\begin{array}{c} n \\ 1 \end{array} \right) + \left(\begin{array}{c} n \\ 2 \end{array} \right) - \cdots + (-1)^n \left(\begin{array}{c} n \\ n \end{array} \right) = 0, \quad n \geq 1.$$

(In case $n = 0$ setting $x = -1$ yields the not very interesting (but true!) fact that both

$$(1-1)^0 \text{ and } \left(\begin{array}{c} 0 \\ 0 \end{array} \right)$$

equal 1.) Such an identity should have a combinatorial interpretation. To discover it we transpose the negative terms (assuming $n \geq 1$) to obtain

$$\binom{n}{0} + \binom{n}{2} + \cdots = \binom{n}{1} + \binom{n}{3} + \cdots$$

which simply asserts that the number of even element subsets equals the number of odd element subset. Since the total number of subsets is 2^n, each side of the above equality equals 2^{n-1}.

Another trick is to differentiate both sides of

$$(1+x)^n = \sum_{r=0}^{n} \binom{n}{r} x^r$$

one or more times and set x equal to appropriate values. For example, differentiating once we have

$$n(1+x)^{n-1} = \sum_{r=0}^{n} \binom{n}{r} r x^{r-1},$$

and setting $x = 1$, we have

$$n2^{n-1} = 1\binom{n}{1} + 2\binom{n}{2} + \cdots + n\binom{n}{n}.$$

The combinatorial interpretation of this formula, which it presumably must have, is not so readily apparent. (Can you find it?)

These results suggest that there are also interesting identities satisfied by the multinomial coefficients. This is true, but since their combinatorial meaning (the number of permutations of a multiset or the number of what we called labeled partitions of a set) is more complex than the combinatorial meaning of the binomial coefficients, we can expect multinomial identities to be less appealing than binomial identities.

Exercises Section 1.7

1. Prove the recursion identity (Theorem 7.1) by substituting in the formula of Theorem 5.3.

2. Prove the bound

$$r(n,r) \leq \binom{n+r-2}{r-1}$$

for the Ramsey numbers.

3. Give a combinatorial proof that for all non-negative integers n and r,

$$\binom{n}{0} + \binom{n+1}{1} + \cdots + \binom{n+r}{r} = \binom{n+r+1}{r}.$$

Hint: Use Exercise 10 of Section 6.

4. Without using binomial coefficients prove that the number of subsets of an n-set, $n \geq 1$, having an even number of elements equals the number having an odd number of elements.

5. In your calculus text it was (probably) proved that the Taylor series for the function $f(x) = (1+x)^\alpha$ converges for all real numbers α and $|x| < 1$. From this fact (which you may wish to review in your calculus text), give a reasonable definition of

$$\binom{\alpha}{r}$$

where α is any real number and r is any nonnegative integer.

6. Can you find a single binomial coefficient equal to

$$\binom{n}{r} + 3\binom{n}{r-1} + 3\binom{n}{r-2} + \binom{n}{r-3}?$$

7. Let S be an n-set with $n \geq 3$ and let a, b, c be three distinct elements of S. By counting the number of r-subsets of S which contain either a, b, or c, in two different ways prove that

$$\binom{n}{r} - \binom{n-3}{r} = \binom{n-1}{r-1} + \binom{n-2}{r-1} + \binom{n-3}{r-1}$$

for all $1 \leq r \leq n$.

8. Prove that for every integer $n \geq 2$

$$\binom{n}{1} - 2\binom{n}{2} + 3\binom{n}{3} + \cdots + (-1)^{n-1} n \binom{n}{n} = 0.$$

9. Prove that for every non-negative integer n

$$\binom{n}{0} + \frac{1}{2}\binom{n}{1} + \frac{1}{3}\binom{n}{2} + \cdots + \frac{1}{n+1}\binom{n}{n} = \frac{2^{n+1}-1}{n+1}.$$

10. For all positive integers m, n, and r, use a counting argument to prove

$$\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} = \binom{m+n}{r}.$$

This is called the *Vandermonde convolution formula* for the binomial coefficients.

1.8 The principle of inclusion and exclusion

We have already considered some special cases of the *principle of inclusion and exclusion* — also known as the *principle of cross-classification* — in Section 4 where we first observed that for sets A_1 and A_2,

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|.$$

This formula generalized the partitioning (or addition) principle for disjoint sets. In Exercise 2 of Section 4, this formula was extended to three sets:

$$|A_1 \cup A_2 \cup A_3| =$$

$$|A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|.$$

This formula is easily pictured in a Venn diagram.

In Exercise 8 of Section 4 we found it useful to compute the size of the complement of a union and this is the form in which we now want the general principle of inclusion and exclusion. Also, if all of the sets under consideration are subsets of a set S it is useful to determine the subset $A_i \subset S$ by some defining property p_i. For example if S is a set of colored objects, then A_1, A_2, A_3 might be the subsets of *red, blue,* and *green* objects, respectively. Then $A_1 \cap A_2$ is the subset of objects colored both *red* and *blue* while $A_1 \cup A_2$ is the subset of objects colored *red* or *blue* or both. Of course if we start with given subsets then we may still think of properties: any subset $A \subset S$ is determined by the property "$x \in A$", which every element x in S may or may not have. Finally we recall *DeMorgan's law* for sets: the complement of a union is the intersection of the complements. This means that

$$S - (A_1 \cup \cdots \cup A_m) = (\overline{A_1 \cup \cdots \cup A_m}) = \overline{A}_1 \cap \cdots \cap \overline{A}_m,$$

depending on whether we use set difference or set complement.

Theorem 1.8.1 *(Principle of inclusion and exclusion) Let S be a finite set and let $p_1, p_2, \ldots, p_n$ be a set of properties which each element of S may or may not have. For each $i = 1, \ldots, n$, let A_i be the subset of S consisting of all objects having property p_i. Then*

$$|S - (A_1 \cup \cdots \cup A_n)| =$$

$$|S| - \sum_i |A_i| + \sum_{i<j} |A_i \cap A_j| - \sum_{i_1<i_2<i_3} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| + \cdots + (-1)^n |A_1 \cap \cdots \cap A_n|.$$

Proof We could prove this by induction, as is certainly suggested by Exercise 2 of Section 4. Let us instead give an interesting counting argument. Let x be any element of the set S. Then, in general, for some k, $0 \leq k \leq n$, x will be in *precisely* k of the A_i; to be specific suppose

$$x \in A_{i_1} \cap \cdots \cap A_{i_k}.$$

With this assumption, let us count how much x contributes to each side of the equation we are trying to establish. If our count for this arbitrary $x \in S$ is the same for both sides we will have established the equality. Now for the right side x contributes to each term the following:

1 to $|S|$,

1 to $k = \begin{pmatrix} k \\ 1 \end{pmatrix}$ of $|A_1|, \ldots, |A_n|$,

1 to $\begin{pmatrix} k \\ 2 \end{pmatrix}$ of $|A_1 \cap A_2|,\ |A_1 \cap A_3|, \ldots,\ |A_{n-1} \cap A_n|$,

1 to $\begin{pmatrix} k \\ 3 \end{pmatrix}$ of $|A_1 \cap A_2 \cap A_3|, \ldots, |A_{n-2} \cap A_{n-1} \cap A_n|$,

. . .

1 to $\begin{pmatrix} k \\ k \end{pmatrix}$ of $|A_1 \cap \cdots \cap A_k|, \ldots, |A_{n-k+1} \cap \cdots \cap A_n|$, and

0 to the remaining terms.

Hence x contributes to the right side of the equation the amount

$$\begin{pmatrix} k \\ 0 \end{pmatrix} - \begin{pmatrix} k \\ 1 \end{pmatrix} + \ldots + (-1)^k \begin{pmatrix} k \\ k \end{pmatrix} = (1-1)^k$$

and this

$$= 0 \text{ if } k > 0,$$
$$= 1 \text{ if } k = 0.$$

But x contributes 1 to the left hand side of the equation if x has none of the properties, i.e.: if $k = 0$, and 0 if x has at least one of the properties, i.e.: if $k > 0$. This proves the theorem.

Example 1. How many integers between 1 and 100 are not divisible by any of the integers 2,3,5, and 7?

We take the set $S = \{1, 2, \ldots, 100\}$ and for each $k \in \{2, 3, 5, 7\}$ let A_k be the subset of integers in S which *are* divisible by k. Then if N is the set of integers between 1 and 100 which are not divisible by any of 2,3,5, or 7, we have

$$N = S - (A_2 \cup A_3 \cup A_5 \cup A_7).$$

The point of using the principle of inclusion and exclusion is that it will enable us to compute the size of this set easily since the sizes of the A_k and their intersections are easy to compute. To see this first notice that

$$|A_k| = \left[\frac{100}{k}\right],$$

(where $[x]$ denotes the greatest integer $\leq x$). Hence

$$|A_2| = 50, \;\; |A_3| = 33, \;\; |A_5| = 20, \;\; |A_7| = 14.$$

Next, observe that an integer n is divisible by both of two integers a and b iff n is divisible by the least common multiple of a and b. But if a and b are both prime, their lcm is their product. From this it follows that

$$|A_2 \cap A_3| = \left[\frac{100}{2 \cdot 3}\right] = 16,$$

and for the same reason

$$|A_2 \cap A_5| = 10, \;\; |A_2 \cap A_7| = 7, \;\; |A_3 \cap A_5| = 6, \;\; |A_3 \cap A_7| = 4, \;\; |A_5 \cap A_7| = 2,$$

Continuing, an integer is divisible by each of three or more primes iff it is divisible by their product. From this we have

$$|A_2 \cap A_3 \cap A_5| = \left[\frac{100}{2 \cdot 3 \cdot 5}\right] = 3;$$

likewise

$$|A_2 \cap A_3 \cap A_7| = 2, \;\; |A_2 \cap A_5 \cap A_7| = 1, \;\; |A_3 \cap A_5 \cap A_7| = 0,$$

and finally

$$|A_2 \cap A_3 \cap A_5 \cap A_7| = 0.$$

The principle of inclusion and exclusion then asserts that the number of integers between 1 and 100 which are not divisible by any of 2,3,5, or 7, is

$$|N| = |S - (A_2 \cup A_3 \cup A_5 \cup A_7)| =$$

$$100 - 50 - 33 - 20 - 14 + 16 + 10 + 7 + 6 + 4 + 2 - 3 - 2 - 1 - 0 + 0 = 22.$$

It is worth observing why this computation is interesting. We note that N does not contain any of 2,3,5,7, but does contain 1. What else does N contain? Recall that a positive integer n is said to be *composite* if $n = r \cdot s$ where neither r nor s is 1. Hence every positive integer is either 1, or is a prime, or is composite. We show that N can contain no composite integers. If it did, say $n \in N$ is composite, then if we take p to be the least factor of n different from 1, p is a prime. Hence $n = pq$ where q is some integer different from 1, and since p is the least such factor of n, $p \leq q$ and hence $p^2 \leq n \leq 100$. Therefore $p \leq \sqrt{n} \leq 10$, which means that the divisor p of n is one 2,3,5,7. Hence $n \notin N$. Hence N consists of 22 integers which except for 1 are all prime. Since N does not include 2,3,5,7, we conclude that the number of primes between 1 and 100 is $22 - 1 + 4 = 25$. Later, when our focus is on properties of the

integers, we will generalize this simple computation.

COMBINATIONS OF MULTISETS WITH LIMITED REPETITION

Let us return to the problem, initially considered in Section 6, of determining the number of r-combinations of a multiset. In Section 6 we showed (Theorem 6.4) that if the multiset $M = \{r_1 \cdot 1, \ldots, r_n \cdot n\}$ has repetition numbers $r_i \geq r$, then the number of r-combinations of M is

$$\left(\begin{array}{c} n+r-1 \\ r \end{array} \right).$$

Also we observed that this is equal to the number of non-negative integer solutions of

$$x_1 + \cdots + x_n = r.$$

We used this fact to solve the problem of determining the number of r-combinations of M where the number of choices of each object was bounded below by some integer other than 0. Now we want to consider the more general problem, formulated at the end of Section 6, of determining the number of r-combinations of M where the repetition number r_i may also be less than r, i.e.: where we have limited choice. Equivalently, we want to find the number of integer solutions of

$$x_1 + \cdots + x_n = r$$

where the x_i are to satisfy $a_i \leq x_i \leq r_i$ for all $i = 1, \ldots, n$. As in Section 6, by the change of variable $y_i = x_i - a_i$ we can transform the problem into one where we replace all a_i by 0. Hence we want to determine the number of non-negative integer solutions of

$$x_1 + \cdots + x_n = r$$

where $x_i \leq r_i$ for $i = 0, \ldots, n$. (The r_i have no particular relation to r.) To do this let us consider, along with the multiset $M = \{r_1 \cdot 1, \ldots, r_n \cdot n\}$, the multiset $M' = \{\infty \cdot 1, \ldots, \infty \cdot n\}$. Let S be the set of *all* non-negative integer solutions $(x_1, \ldots, x_n)$ of

$$x_1 + \cdots + x_n = r.$$

Then we already know that $|S|$ is the number of r-combinations of M' and hence

$$|S| = \left(\begin{array}{c} n+r-1 \\ r \end{array} \right).$$

For each i let p_i be the property of a solution $(x_1, \ldots, x_n)$: x_i is *not* $\leq r_i$, i.e.: p_i is the property that $r_i + 1 \leq x_i$. Then the corresponding subset of S is

$$A_i = \{(x_1, \ldots, x_n) \in S : r_i + 1 \leq x_i\}$$

so

$$\overline{A_i} = \{(x_1, \ldots, x_n) \in S : 0 \leq x_i \leq r_i\}.$$

Consequently

$$\overline{A_1} \cap \cdots \cap \overline{A_n} = \{(x_1, \ldots, x_n) \in S : 0 \leq x_i \leq r_i \text{ for all } i = 1, \ldots, n\},$$

and it is the size of this set which we want. We can compute its size, using the inclusion-exclusion formula of Theorem 8.1, from the sizes of the various intersections of the A_i.

To compute these we introduce the new variables $z_i = x_i - (r_i + 1)$ so that the condition $r_i + 1 \leq x_i$ is equivalent to $0 \leq z_i$. Substituting, we have

$$A_i = \textit{ the set of all non-negative solutions of}$$

$$x_1 + \cdots + z_i + \cdots + x_n = r - (r_i + 1).$$

Therefore

$$|A_i| = \left(\begin{array}{c} n + r - (r_i + 1) - 1 \\ r - (r_i + 1) \end{array} \right) = \left(\begin{array}{c} n + r - (r_i + 1) - 1 \\ n - 1 \end{array} \right).$$

Of course if $r - (r_i + 1) < 0$ then $|A_i| = 0$ since in this case there are no non-negative solutions. Likewise

$$A_i \cap A_j = \textit{ the set of all non-negative solutions of}$$

$$x_1 + \cdots + z_i + \cdots + z_j + \cdots + x_n = r - (r_i + 1) - (r_j + 1),$$

and therefore

$$|A_i \cap A_j| = \left(\begin{array}{c} n + r - (r_i + 1) - (r_j + 1) - 1 \\ n - 1 \end{array} \right)$$

or 0 (if $r - (r_i + 1) - (r_j + 1) < 0$.)

Similar formulas are easily obtained for intersections of three or more of the A_i.

Thus we have solved the problem of computing the number of r-combinations of a multiset with both upper and lower constraints on the choices. Specific applications of this method appear in the exercises.

DERANGEMENTS

A final interesting application of the principle of inclusion and exclusion is to the study of a special class of permutations called *derangements*. Given the (ordinary) set $S = \{1, 2, \ldots, n\}$, a derangement of S is a permutation $k_1, \ldots, k_n$ of S in which no $k_i = i$, i.e.: no element of S is in its natural position. Thus derangements are the arrangements of S in which the maximum amount of disturbance has occurred. For an n-set S, D_n denotes the number of derangements of S. For example if $S = \{1, 2, 3\}$, the derangements of S are 231 and 312 so $D_3 = 2$. Obviously $D_1 = 0$, $D_2 = 1$, and its not hard to discover, by direct counting, that $D_4 = 9$. Also we define $D_0 = 1$. (Why does this make sense?) The numbers D_n are called the *derangement numbers.*

The derangement numbers solve the celebrated *hat check problem:* Suppose n people have checked their hats; all lose their claim tickets so their hats are returned randomly. Then D_n is the number of ways in which their hats can be returned so that no one gets his own hat. Consequently $n! - D_n$ is the number of ways their hats can be returned so that at least one person gets his own hat.

The interesting thing about derangements is that we have a nice formula for the derangement numbers D_n. To obtain it we first notice a corollary of Theorem 8.1 which results if for each k each of the intersections $A_{i_1} \cap \cdots \cap A_{i_k}$ has the same size.

Corollary 1.8.1 *In Theorem 8.1 if*

$$|A_1| = \cdots = |A_n| = s_1,$$

$$|A_1 \cap A_2| = \cdots = |A_{n-1} \cap A_n| = s_2,$$

$$\cdot \quad \cdot \quad \cdot$$

$$|A_1 \cap \cdots \cap A_n| = s_n,$$

then

$$|S - (A_1 \cup \cdots \cup A_n)| = |S| - s_1 \binom{n}{1} + s_2 \binom{n}{2} - \cdots + (-1)^n s_n \binom{n}{n}.$$

PROOF Observe that for each k there are $\binom{n}{k}$ intersections $A_{i_1} \cap \cdots \cap A_{i_k}$, each with size s_k.

Theorem 1.8.2 *For all non-negative integers n,*

$$D_n = n! \left(\frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} \right).$$

PROOF We apply the corollary to the set S of all $n!$ permutations of $\{1, 2, \ldots, n\}$. Let A_i be the set of permutations $k_1, \ldots, k_n$ in S in which only $k_i = i$. Then each $|A_i| = (n-1)!$. The number of permutations in S having precisely two integers remain in their natural position is $(n-2)!$, so $|A_{i_1} \cap A_{i_2}| = (n-2)!$, and in general

$$s_k = |A_{i_1} \cap \cdots \cap A_{i_k}| = (n-k)!$$

for each $k = 0, 1, \ldots, n$. Therefore

$$s_k \binom{n}{k} = (n-k)! \frac{n!}{k!(n-k)!} = \frac{n!}{k!}.$$

Substituting in the corollary, we obtain the formula in the theorem.

The nice thing about this formula for D_n is that from the Taylor series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

$x = -1$ yields

$$e^{-1} = \frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots,$$

so that

$$e^{-1} = \frac{D_n}{n!} + (-1)^{n+1}\frac{1}{(n+1)!} + (-1)^{n+2}\frac{1}{(n+2)!} + \cdots.$$

Hence by the alternating series test for infinite series we know that e^{-1} and $D_n/n!$ differ by less than $1/(n+1)!$, which means, since $1/(n+1)!$ goes to 0 rapidly, that $n!e^{-1}$ is a good approximation to D_n. For example even taking n as small as 4 gives $4!e^{-1} = 8.83\cdots$, whereas $D_4 = 9$, so the error is less than 2%. For $n = 7$ one can approximate D_7 to three decimals in this way.

Finally, $D_n/n!$, which is approximately e^{-1}, is the ratio of derangements to the total number of permutations and hence is the probability that a randomly chosen permutation is a derangement. In the hat check problem above, it is the probability that no one gets his own hat back, i.e.: for reasonably large n the probability of no one getting his own hat back is a little bigger than 1/3. $1 - D_n/n!$ which is approximated by

$$1 - e^{-1} = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots$$

is the probability that a random permutation is not a derangement. Thus the probability that at least one person gets his own hat back becomes close to 2/3.

Like so much in mathematics, the principle of inclusion and exclusion, as we have stated it, is capable of generalization. A first step is to define for a set S, $|S|(r)$ to be the number of elements of S having exactly r of the properties $p_1, \ldots, p_n$. Then Theorem 8.1 gives us a formula for $|S|(0)$. Without too much more effort we could develop a more general formula for $|S|(r)$ for $r = 0, 1, 2, \ldots$. But we will resist the temptation.

Exercises Section 1.8

1. Let $S = \{1, 2, \ldots, 1000\}$. Find the number of integers in S which are divisible by none of 5,6, and 8.

2. Let $S = \{1000, 1001, \ldots, 9999\}$. Find the number of integers in S having all three of the digits 0,1,2 occur at least once. Hint: for $k = 0, 1, 2$ let A_k be the subset of S consisting of those integers in which the digit k does not occur.

3. Let $S = \{g, i, r, l, a, n, d, b, o, y\}$ How many permutations of S are there in which none of the words "girl","and","boy" occur as consecutive letters? Hint: Let A_g be the set of all permutations of S in which the word "girl" does occur. Notice that $|A_g|$ is just the number of permutations of the set of 7 symbols $\{girl, a, n, d, b, o, y\}$. Define A_a and A_b likewise.

4. Given 3 apples, 4 bananas, and 5 cherries, how many different baskets of 10 pieces of fruit can be chosen? (Note: This can be solved either by the general method described in the text or, more simply, by a direct counting argument.)

5. Find the number of integer solutions of the equation

$$x_1 + x_2 + x_3 + x_4 = 18$$

which satisfy

$$1 \leq x_1 \leq 5, \quad -2 \leq x_2 \leq 4, \quad 0 \leq x_3 \leq 5, \quad 3 \leq x_4 \leq 9.$$

6. If S is an n-set and T is an m-set and $m \leq n$, find the number of functions having domain S and range T.

7. Find the number of permutations of $\{1, 2, 3, 4, 5, 6, 7, 8\}$ in which no even integer is in its natural position.

8. Find the number of permutations of $\{1, 2, 3, 4, 5, 6, 7, 8\}$ in which exactly 4 integers are in their natural position.

8. Find a simple formula for the number of permutations of the set $\{1, 2, \ldots, n\}$ in which exactly k are in their natural position.

10. Use the equality principle to prove the identity

$$n! = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) D_{n-k}$$

where n is any non-negative integer.

11. a) Verify that the derangement numbers D_n satisfy the recursion formula

$$D_n = nD_{n-1} + (-1)^n$$

for all $n \geq 1$.
b) Show that D_n is an even integer according as n is an odd integer.

Chapter 2

The Integers

The integers are the most familiar mathematical entities. In this chapter we shall study some of their most fundamental properties. Perhaps we should emphasize at the outset that we shall not attempt to explain what the integers actually *are*, but rather we shall start with the rather informal assumption of some of their simple properties and from these we shall derive others. Elementary texts on set theory may be consulted for a discussion of how the integers can actually be constructed starting with some basic assumptions about sets. Aside from the basic facts about addition, subtraction, multiplication, and inequalities, familiar from elementary arithmetic, our basic assumption about the integers is the well ordering principle:

> **Well ordering principle.** *Every non-empty subset of* $\mathbb{N}$ *contains a least element.*

This principle, in the presence of our other assumptions, is logically equivalent to the Principle of Mathematical Induction. (See Appendix A.) Often it is useful to invoke a slightly more general version of the well ordering principle which results by taking m to be any integer (positive, negative, or zero) and replacing $\mathbb{N}$ by the set M of all integers at least as large as m.

First some notation. The set of all (positive and negative) integers, including 0 is denoted by $\mathbb{Z}$. The set of positive integers $\{1, 2, 3, \ldots\}$ is denoted by $\mathbb{Z}^+$, and the set of non-negative integers $\{0, 1, 2, 3, \ldots\}$ is denoted by $\mathbb{N}$, which stands for *natural numbers*. (The letter $\mathbb{Z}$ is used since it stands for the German word for numbers, *zahlen*.) Finally, the rational numbers, since they are quotients of members of $\mathbb{Z}$, are denoted by $\mathbb{Q}$, and for the record, the reals are of course denoted by $\mathbb{R}$, and the complex numbers by $\mathbb{C}$.

2.1 Divisibility and Primes

For integers m and n in $\mathbb{Z}$ we say that m *divides* n just in case there is an integer $k \in \mathbb{Z}$ such that $mk = n$. If m divides n we express this by writing $m|n$. It is important not

to confuse the *relation* of divisibility, denoted by the symbol |, with the *operation* of division, usually denoted by the symbol /. For example $2|6$ is the true sentence which expresses the fact that 2 divides 6, while $2/6$ denotes a certain (rational) number. If m does not divide n we express this by $m \not| n$. Thus $2|3$ is a false sentence so $2 \not| 3$ is true, while $2/3$ is again a rational number. Notice that $1|n$ for all $n \in \mathbb{Z}$ and $m|0$ for all $m \in \mathbb{Z}$. In particular notice that $0|0$, (and this is not at odds with the fact that $0/0$ is undefined.)

The following lemma summarizes some simple facts about divisibility.

Lemma 2.1.1 *(Properties of divisibility)*
a) For $m, n \in \mathbb{Z}^+$ $m|n$ implies $m \leq n$.
b) For $k, m, n \in \mathbb{Z}$ $k|m$ and $m|n$ implies $k|n$.
c) For $k, m, n \in \mathbb{Z}$ $m|n$ implies $mk|nk$.
d) For $a, b, k, m, n \in \mathbb{Z}$ $k|m$ and $k|n$ implies that $k|(am + bn)$.

Proof. Exercise.

Definition. A positive integer p is *prime* if
a) $p > 1$ and
b) for all $m \in \mathbb{Z}^+$, $m|p$ implies $m = 1$ or $m = p$.

The prime numbers are the multiplicative building blocks of the integers. The precise statement of this fact is the following theorem which is as important as its name suggests.

Theorem 2.1.1 *(Fundamental theorem of arithmetic) If n is a positive integer greater than 1, then*
a) n is a product of primes, and
b) apart from permutation of the factors, n can be expressed as a product of primes in only one way.

Proof. Part a): We use strong induction on n. If $n = 1$ there is nothing to prove. Assume that $n > 1$ and that for each $k < n$, $k = 1$ or k is a product of primes. If n is a prime we are done, otherwise n has a divisor $m \in \mathbb{Z}^+$ which is neither 1 nor n. By the Lemma above $n = mk$ where both $1 < m < n$ and $1 < k < n$. By the induction assumption both m and k are products of primes and hence so is their product n. This proves Part a).

The primes obtained in the factorization just obtained are not necessarily distinct, nor are they arranged in any particular order. If we arrange them in increasing order and group together sets of equal primes we obtain

$$n = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where $p_1 = 2,\ p_2 = 3,\ \cdots$, and each $n_i \in \mathbb{N}$. We often call this the *standard representation* of n. Notice that $n = 1$ iff all $n_i = 0$.

PROOF. Part b): We prove this by contradiction. Suppose that assertion b) is false; then by the Well Ordering Principle, there is a *least* positive integer $n > 1$ for which the assertion fails. Then we may suppose both

$$n = p_1 p_2 \cdots p_t, \text{ and}$$

$$n = q_1 q_2 \cdots q_s,$$

where the p_i and q_j are prime and no p_i is a q_j (for otherwise we could reduce n). We may further suppose $p_1 > q_1$. Define

$$m = (p_1 - q_1) p_2 \cdots p_t.$$

Then $0 < m < n$.

Now the factor $p_1 - q_1$ is not divisible by q_1, for $p_1 - q_1 = kq_1$ implies $p_1 = (k+1)q_1$, contradicting the primality of p_1. Also, by Part a) of the theorem, $p_1 - q_1$ is either 1 or $r_1 \cdots r_k$, r_i prime, so that

$$\begin{aligned} m &= p_2 \cdots p_t, \text{ or} \\ & \quad r_1 \cdots r_k p_2 \cdots p_t \end{aligned}$$

and in either case we have a prime factorization of m not involving q_1. But on the other hand,

$$\begin{aligned} m &= p_1 p_2 \cdots p_t - q_1 p_2 \cdots p_t \\ &= q_1 q_2 \cdots q_s - q_1 p_2 \cdots p_t \\ &= q_1 (q_2 \cdots q_s - p_2 \cdots p_t) \\ &= q_1 s_1 \cdots s_k, \end{aligned}$$

where the s_i are prime by Part a). Hence we have two distinct factorizations for $m < n$, a contradiction. This completes the proof of Part b).

The unique representation

$$n = p_1^{n_1} p_2^{n_2} \cdots = 2^{n_1} 3^{n_2} 5^{n_3} \cdots$$

for each $n \in \mathbb{Z}^+$ gives us a uniquely determined 1-1 correspondence

$$n \leftrightarrow (n_1, n_2, \ldots)$$

between the $n \in \mathbb{Z}^+$ and the set of all sequences $(n_1, n_2, \ldots)$ such that $n_i \in \mathbb{N}$ and all but finitely many n_i are 0. In fact it is often useful to take advantage of this correspondence in discussing multiplicative properties of $\mathbb{Z}$. Thus we can write integers in "exponential form", that is we can replace the integer n by its exponents

$$(n_1, n_2, \ldots)$$

where the n_i are uniquely determined by the Fundamental Theorem. In exponential form, if

$$n \leftrightarrow (n_1, n_2, \ldots) \text{ and } m \leftrightarrow (m_1, m_2, \ldots)$$

then

$$m = n \Leftrightarrow m_i = n_i, \;\; i = 1, 2, \ldots,$$

$$mn \leftrightarrow (m_1 + n_1, m_2 + n_2, \ldots), \text{ and } m|n \Leftrightarrow m_i \leq n_i, \;\; i = 1, 2, \ldots.$$

Using the last of the above equivalences properties of the divisibility relation on $\mathbb{Z}^+$ can be proved by appealing to corresponding properties about the $\leq$ relation on $\mathbb{N}$. For example the property

$$a|b \text{ and } b|a \Rightarrow a = b$$

becomes

$$\text{For all } i,\ a_i \leq b_i \text{ and } b_i \leq a_i \Longrightarrow \text{ for all } i,\ a_i = b_i.$$

A more important example is given below. If $a, b \in \mathbb{Z}^+$ and 1 is the only positive integer which divides both, then a and b are *relatively prime* (or *coprime*). In exponential form a and b are relatively prime iff $a_i b_i = 0$ for all i. (Why?)

Theorem 2.1.2 *(Euclid's Divisibility Theorem) If a and b are relatively prime positive integers and $a|bc$, then $a|c$.*

Proof. The hypothesis that $a|bc$ is equivalent to:
i) $a_i \leq b_i + c_i$ for all i,
while the assertion that a and b are coprime means that
ii) $a_i b_i = 0$ for all i.

We want to show that $a_i \leq c_i$ for all i. For any i for which $a_i = 0$ this is obviously true; otherwise for $a_i > 0$, multiply 1) by a_i, apply ii) and obtain $a_i^2 \leq a_i c_i$ which implies $a_i \leq c_i$.

Exercises Section 2.1

1. Use exponential notation to show that if $p, a, b \in \mathbb{Z}^+$, p a prime, then

$$p|ab \Rightarrow p|a \text{ or } p|b.$$

2. Show that if $1 < n \leq N$ and n is not prime, then n must be divisible by a prime $\leq \sqrt{N}$.

3. Using Exercise 2 it is easy to construct a table of primes up to a moderate limit N by a procedure called the "sieve of Eratosthenes":
In the list of integers

$$2, 3, 4, 5, 6, \ldots, N$$

strike out every 2nd integer beginning with $2^2 = 4$, leaving

$$2, 3, 5, 7, 9, 11, 13, 15, \ldots.$$

Next strike out every 3rd integer in the list beginning with $3^2 = 9$. Continue the process until the next remaining integer n, following the one all of whose multiples were last canceled, is $> \sqrt{N}$. Then the remaining integers are all primes between 1 and N. Use this method to construct a table of all primes less than 100.

4. Exercise 2 suggests that there are infinitely many primes. This fact was first published by Euclid. Prove it by contradiction as follows:

Suppose there are only finitely many primes $2, 3, 5, \ldots, p$, i.e.: p is the largest prime. Let

$$N = (2 \cdot 3 \cdot 5 \cdots p) + 1.$$

Show that N is not a prime but that none of the primes $2, 3, \ldots, p$ divides N.

5. Exercise 4 (falsely) suggests that for each n, $N_n = (p_1 p_2 \cdots p_n) + 1$ might be prime. What is the first n for which this fails?

6. a) Show that every odd prime is of the form $4n + 1$ or $4n + 3$ for some $n \in \mathbb{N}$.

b) Show that the product of primes of the form $4n + 1$ is again of this form.

7. Show that here are infinitely many primes of the form $4n + 3$. (Hint: suppose the contrary and that p is the greatest prime of this form. Let $N = (2^2 \cdot 3 \cdot 5 \cdots p) - 1$, and apply Exercise 6.) This exercise proves a special case of a celebrated theorem of Dirichlet which asserts that if a and b are relatively prime, then there are infinitely many primes of the form $an + b$.

8. Show that if $n > 1$ and $a^n - 1$ is prime, then $a = 2$ and n is prime. (Primes of the form $2^p - 1$, p prime, are called Mersenne primes.)

9. Given any $m > 1$, show that there are m consecutive composite integers. (Hint: consider $m!$.)

10. Prove that for every $n > 1$, the number $s_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is never an integer. Without a hint this is a hard problem. The solution is simplified if one uses a theorem known as *Bertrand's postulate:* for each prime p there is another prime between p and $2p$. (This was actually a conjecture of Bertrand's which was proven by Tchebychef in 1850. The proof is not elementary.) With or without the use of Bertrand's postulate, give a proof by contradiction, i.e.: suppose that s_n is an integer for some n. Then $n! s_n$ is also an integer. For an easier proof assuming Bertrand's postulate, then let p be the largest prime $\leq n$ and proceed from here to obtain a

contradiction. The proof is a little more complicated without assuming Bertrand's postulate; to proceed this way let p^α be the largest *prime power* $\leq n$ and proceed from here to obtain a contradiction.

2.2 GCD and LCM

We have already verified in the last section that the relation of divisibility on the positive integers has the following properties:

$a|a$ for all $a \in \mathbb{N}$. (reflexivity)
$a|b$ and $b|a$ implies $a = b$. (anti-symmetry)
$a|b$ and $b|a$ implies $a|c$. (transitivity)

Recall that binary relations having these three properties are called *partial orders* or often just *orders*. Other examples of partial orders include $\leq$ on any of $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{Q}$ etc., but not on $\mathbb{C}$. Another familiar example is the inclusion relation $\subset$ on sets. If R is a binary relation on the set S and which is reflexive, anti-symmetric, and transitive, then the system (S, R) is called a *partially ordered set* (abbreviated *poset*) or just an *ordered set*. If in addition to the three defining properties R satisfies the following additional condition, which is usually called *connectedness*,

for every a and b in S, either aRb or bRa.

then R is called a *total order* and the system (S, R) is called a *chain* or a *totally ordered set*. The relation $\leq$ on any of $\mathbb{N}$, $\mathbb{Z}$ etc., is the commonest example of a total order.

The concepts of *upper* (and *lower*) bounds of sets of real numbers and of the important concept of a *least upper bound* (and *greatest lower bound*) denoted respectively by lub and glb apply equally to any poset. Let us review their definitions:

For the poset (S, R) if T is a subset of S, an element $u \in S$ is an *upper bound* for T if

$$(\forall x)(x \in T \Rightarrow xRu).$$

The element $u \in S$ is the least upper bound of T (written $u = \text{lub}(T)$) if
u is an upper bound and
for every upper bound u' of T, uRu'.

An element $v \in S$ is a *lower bound* for T if

$$(\forall x)(x \in T \Rightarrow vRx).$$

The element $v \in S$ is the greatest lower bound of T (written $v = \text{glb}(T)$) if
v is a lower bound and

for every lower bound v' of T, $v'Rv$.

Notice that from the definition any two greatest lower bounds are equal and likewise for least upper bounds. Hence we are entitled to speak of *the* least upper bound (or greatest lower bound) provided one exists. Notice also that for $T \subset S$ if $\text{lub}(T)$ or $\text{glb}(T)$ exists it need not be in be in T, but only in S.

Recall from Calculus that an important property of the real numbers $\mathbb{R}$, as ordered by $\leq$, is that every non-empty set which is bounded above (respectively, below) has a lub (respectively glb) in $\mathbb{R}$. For example if T is the subset of $\mathbb{R}$ consisting of the set of all rational numbers with squares no larger than 2, then certainly T is bounded above and hence $\text{lub}(T)$ exists and in fact is $\sqrt{2}$, which, since it is irrational, is not in T. (See exercise 3, below.) Indeed the real numbers can be "constructed" from the rationals by a "completion" process which endows sets of rationals which are bounded above (below) with lub's (glb's). This process is described in many texts on advanced calculus or elementary real analysis.

A mathematical concept which unifies the concepts discussed above is that of a lattice. A *lattice* is a poset in which each pair of elements has both a least upper bound and a greatest lower bound. It is trivial that the poset $(\mathbb{R}, \leq)$ of the real numbers ordered by $\leq$ is a lattice as is any chain, since in this case for any pair of elements a, b, $\text{lub}(a, b) =$ the larger of a and b, and $\text{glb}(a, b) =$ the lesser.

Returning to the poset $(\mathbb{N}, |)$, if a and b are in $\mathbb{N}$ then $\text{glb}(a, b)$ should be an integer d which divides both a and b, i.e.: is a *common divisor*, and which is divisible by every common divisor, and hence is *greatest* relative to the order relation $|$. This integer is called the *greatest common divisor* of a and b. (By Lemma 2.1.1, a), it is also a common divisor which is *greatest* relative to the order relation $\leq$, but this is not the significant part of the definition.) Likewise $\text{lub}(a, b)$ should be an integer m which is divisible by both a and b, i.e.: it is a *common multiple*, and which divides every common multiple, and hence is *least* relative to $|$. This integer is called the *least common multiple* of a and b. We denote the greatest common divisor of a and b either by $\gcd(a, b)$ or (a, b) and the least common multiple by either $\text{lcm}(a, b)$ or $[a, b]$. (The notation (a, b) and $[a, b]$ is the most common.) Notice carefully that according to this definition we have for $a \neq 0$, $\gcd(a, 0) = a$ and $\text{lcm}(a, 0) = 0$. Also, from $0|0$ we have $\gcd(0, 0) = \text{lcm}(0, 0) = 0$. In general we have the following theorem:

Theorem 2.2.1 *The poset $(\mathbb{N}, |)$ is a lattice in which $[a, b]$ is the least upper bound of a and b and (a, b) is their greatest lower bound. For nonzero a, b these are given in exponential form by*

$$(a, b)_i = \min(a_i, b_i) \text{ and } [a, b]_i = \max(a_i, b_i).$$

The integers 1 and 0 are the unique smallest and (respectively) largest elements of the lattice.

For example we have, using exponential notation,
$600 \leftrightarrow (3, 1, 2, 0, \cdots)$ since $600 = 2^3 \cdot 3 \cdot 5^2$, and
$480 \leftrightarrow (5, 1, 1, 0, \cdots)$ since $480 = 2^5 \cdot 3 \cdot 5$, so
$(600, 480) \leftrightarrow (3, 1, 1, 0, \cdots)$, i.e.: $(600, 480) = 2^3 \cdot 3 \cdot 5 = 120$, and
$[600, 480] \leftrightarrow (5, 1, 2, 0 \cdots)$, i.e.: $[600, 480] = 2^5 \cdot 3 \cdot 5^2 = 2400$.

PROOF of Theorem 2.2.1: Let $d \leftrightarrow (\ldots, \min(a_i, b_i), \ldots)$. Then $d|a$ and $d|b$ since $\min(a_i, b_i)$ is $\leq$ both a_i and b_i. Hence d is a common divisor (i.e.: a lower bound) of both a and b. Suppose d' is *any* common divisor of both a and b. Then $d'_i \leq a_i$ and $d'_i \leq b_i$ for all i. Hence $d'_i \leq \min(a_i, b_i)$ so $d'|d$. Thus d is, in the sense of the relation $|$, greater than or equal to any lower bound for a and b, i.e.: $d = \gcd(a, b)$.

The proof of the other half of the theorem is obtained by letting

$$m \leftrightarrow (\ldots, \max(a_i, b_i), \ldots)$$

and following the pattern above. The integers 1 and 0 are the unique least and greatest elements of the lattice since for all $a \in \mathbb{N}$, $1|a$ and $a|0$.

Corollary 2.2.1 *For all $a, b \in \mathbb{N}$, $[a, b](a, b) = ab$.*

PROOF. For all i we have $\min(a_i, b_i) + \max(a_i, b_i) = a_i + b_i$.

Notice that a and b are relatively prime iff $(a, b) = 1$ iff $[a, b] = ab$ iff $a_i b_i = 0$ for all i.

In any lattice it is useful to take advantage of the fact that for any pair of elements, a, b, both $\text{glb}(a, b)$ and $\text{lub}(a, b)$ are uniquely determined elements of the lattice, and hence we can think of glb and lub as determining a pair of binary operations on the elements of the lattice, just as $+$ and $\cdot$ are binary operations on the integers. Then, just as $+$ and $\cdot$ obey certain *laws* or *identities*, it is natural to look for laws satisfied by glb and lub. For example, corresponding to the commutative laws $x + y = y + x$ and $x \cdot y = y \cdot x$ of arithmetic we easily see that $\text{glb}(x, y) = \text{glb}(y, x)$ and $\text{lub}(x, y) = \text{lub}(y, x)$ in any lattice, by direct appeal to the definition. It is easy to verify several additional laws. For example both operations are associative:

$$\text{glb}(x, \text{glb}(y, z)) = \text{glb}(\text{glb}(x, y), z), \qquad \text{lub}(x, \text{lub}(y, z)) = \text{lub}(\text{lub}(x, y), z).$$

In the lattice $(\mathbb{N}, |)$ the commutative and associative laws become

$$\begin{aligned} (x, y) &= (y, x), & [x, y] &= [y, x], \\ (x, (y, z)) &= ((x, y), z), & [x, [y, z]] &= [[x, y], z]. \end{aligned}$$

The verification of these is left as an exercise.

An important special property of the lattice $(\mathbb{N}, |)$ is that it satisfies the *distributive* identity: the operation of taking the lcm distributes through the operation of taking

the gcd, just as in ordinary arithmetic the operation of multiplication distributes through the addition operation. Thus we have

$$[x,(y,z)] = ([x,y],[x,z]).$$

But unlike ordinary arithmetic we also have distributivity of the gcd through the lcm:

$$(x,[y,z]) = [(x,y),(x,z)].$$

Because either of these laws can be obtained from the other by formally interchanging lub = lcm and glb = gcd, they are said to be *dual* laws. The proofs are left as exercises. In Section 2.7 we will see one reason for the significance of the distributive laws for the lattice $(\mathbb{N}, |)$.

Exercises Section 2.2

1. Suppose $a|bc$ and $(a,b) = d$. Show that the integer $\frac{a}{d}$ divides c.

2. If $n = a^k,\ n,a,k \in \mathbb{Z}^+$, n is called a (perfect) k-th power. In exponential representation this means that $n_i = ka_i$ for all i, i.e.: all exponents are divisible by k. Prove: if mn is a k-th power and $(m,n) = 1$, then m and n are each k-th powers. Use exponential representation in your proof.

3. For any $k,n \in \mathbb{Z}^+$ show that $n^{1/k}$ is either a positive integer or is irrational.(Hint: Suppose $n^{1/k}$ is not irrational; then $n^{1/k} = p/q$ where we may take $(p,q) = 1$. Show that this forces $q = 1$ so that $n^{1/k}$ is an integer.) From this show that $\sqrt{2}$ and $\sqrt[3]{2}$ are both irrational.

4. Prove: if $a|c$ and $b|c$ and $(a,b) = 1$, then $ab|c$.

5. Prove: for all $n \in \mathbb{Z}^+$, $n > 1$ is prime iff for every $a \in \mathbb{Z}^+$, either $(a,n) = 1$ or $n|a$.

6. Verify the associative and distributive laws for gcd and lcm described in the last paragraph of this section.

7. Show that not all lattices are distributive. Do this by first showing that the subspaces of a vector space form a lattice. (For this exercise it is enough to show this just for the vector space $\mathbb{R}^3$, 3-dimensional Euclidean space.) Show that the distributive law fails in this lattice.

8. Show that in any lattice

$$\text{lub}(x,\text{glb}(x,y)) = x) \quad \text{and} \quad \text{glb}(x,\text{lub}(x,y)) = x.$$

9. Prove that in any lattice if either of the two distributive laws holds then so does the other. Hint: Use Exercise 8.

2.3 The Division Algorithm and the Euclidean Algorithm

Although the process of division is familiar from childhood it is important to formulate it precisely; this is done in the following theorem.

Theorem 2.3.1 *(Division algorithm) If $a, b \in \mathbb{Z}$, $b > 0$, then there are unique integers q and r, with $0 \leq r < b$, such that $a = bq + r$.*

Proof. We recall that the quotient q is the number of times we can subtract the divisor b from the dividend a and leave a nonnegative remainder. With this in mind we let

$$M = \{a - bn : n \in \mathbb{Z}\}.$$

First observe that M contains nonnegative integers, for we may take n to be negative of large magnitude. Hence, by the well-ordering principle, there is a *least* nonnegative integer in M, call it r. Then $r = a - bq$ for suitable $n = q$ in $\mathbb{Z}$. Then certainly $r \geq 0$ so we only need to see that $r < b$. If not, then $b \leq r = a - bq$ which implies that $0 \leq a - b(q + 1) < a - bq = r$ which contradicts the fact that r was the least nonnegative integer in M. Hence $0 \leq r < b$.

For the uniqueness of r and q, suppose that we had both

$$a = bq_1 + r_1, \text{ with } 0 \leq r_1 < b \text{ and } a = bq_2 + r_2, \text{ with } 0 \leq r_2 < b.$$

We may suppose $r_1 \leq r_2$. Subtracting we then have

$$0 = b(q_2 - q_1) + r_2 - r_1.$$

From this we infer that b divides $r_2 - r_1$. But from the restrictions on r_1 and r_2 we have $0 \leq r_2 - r_1 < b$ which means that $r_2 - r_1 = 0$. Then we immediately have that $q_1 = q_2$, completing the proof.

Elementary as it is, it should be appreciated that the division algorithm provides a convenient way to solve a large class of problems by a divide and conquer strategy, and now we mean *divide* in both the sense of arithmetic as well as in the sense of partitioning. Since division of any integer by a positive divisor b leaves a unique remainder r in the set $\{0, 1, \ldots, b - 1\}$, we can frequently break a problem into b cases depending on the value of a particular remainder. For example, Exercise 6 of Section 2.1 asked for a proof that every odd prime is of the form $4n + 1$ or $4n + 3$ for some $n \in \mathbb{N}$. If the prime is p then the Division Algorithm asserts that $p = 4n + r$ where $r \in \{0, 1, 2, 3\}$. The values $r = 0, 2$ are excluded since $r = 0$ would mean that p was composite while $r = 2$ would mean that p is even. Hence $p = 4n+1$ or $p = 4n+3$.

If the integers a and b are factored into the product of primes then we have seen that it is easy to compute both (a, b) and $[a, b]$. This is however not usually practical since, as we shall discuss later in Section 2.8 factoring large numbers seems to be

an extremely time consuming process. Therefore it is preferable to compute (a, b) directly without detouring through the prime factorization of a and b. ($[a, b]$ can then be computed by dividing the product ab by (a, b)). The Euclidean algorithm—which is as old as its name implies—is the basic method for computing the gcd. The process consists of dividing one of the numbers by the other and then dividing the remainder into the preceding divisor. We repeat this process until the remainder is zero. The last nonzero remainder is the gcd. The following theorem makes this explicit.

Theorem 2.3.2 *(Euclidean algorithm) Let $n_0, n_1 \in \mathbb{N}$. Apply the division algorithm successively to obtain the sequence*

$$\begin{aligned} n_0 &= n_1 q_1 + n_2, \ \ 0 \le n_2 < n_1, \\ n_1 &= n_2 q_2 + n_3, \ \ 0 \le n_3 < n_2, \\ n_2 &= n_3 q_3 + n_4, \ \ 0 \le n_4 < n_3, \\ &\cdots \\ n_{t-3} &= n_{t-2} q_{t-2} + n_{t-1}, \ \ 0 \le n_{t-1} < n_{t-2}, \\ n_{t-2} &= n_{t-1} q_{t-1} + n_t, \ \ 0 \le n_t < n_{t-1}. \end{aligned}$$

Since $n_1 > n_2 > \cdots n_t \ge 0$, the process must eventually yield $n_{t+1} = 0$ so that the sequence of divisions can be terminated with the step

$$n_{t-1} = n_t q_t + 0, \ \ 0 = n_{t+1} < n_t.$$

Then $n_t = \gcd(n_0, n_1)$.

PROOF. From the last equation of the sequence we see that n_t divides n_{t-1}; then from the next to last equation we see that n_t divides n_{t-2} as well and from the next equation up the sequence that n_t divides n_{t-3}, and so on up the list. Finally from the top two equations in the sequence, we have that n_t divides both n_1 and n_0. Hence going up the sequence shows that n_t is a common divisor of n_1 and n_2.

Now let d be any common divisor of n_0 and n_1. Then the first equation of the sequence shows that $d|n_2$ as well, the second equation shows $d|n_3$ and so on, and the last equation shows that $d|n_t$. Hence going down the sequence shows that n_t is divisible by any common divisor. We conclude that n_t is the greatest common divisor.

Since each of the division steps in the Euclidean algorithm can be replaced by a sequence of subtractions (as is described in the proof of the division algorithm) we can express the Euclidean algorithm by means of subtractions only. In fact we can do this by means of the following logically simpler algorithm:

read (a, b);
while $a \neq b$ **do**
  begin
    if $a < b$ **then** interchange a and b;

replace a by $a - b$
end;
write(a).

This says that to compute $\gcd(a, b)$ we repeatedly replace the larger of the two by their difference until they become the same. The algorithm terminates with $a = b = \gcd$ of the original a and b. We leave the verification of this fact as an exercise. Of course, simple as it is to describe, this process involves no less actual computing than the original version. Later (Section 2.8) we will discuss in more detail how much computing is actually involved.

Either way we compute the greatest common divisor of two integers there is an additional important property of the Euclidean algorithm, namely that it expresses the gcd (a, b) in the form

$$(a, b) = ax + by,$$

for suitable integers x and y (not necessarily nonnegative), that is, as an integer linear combination of a and b. To see that this is so notice that the next to the last division in the Euclidean algorithm expresses

$$(n_0, n_1) = n_t = n_{t-2} - n_{t-1}q_{t-1},$$

that is, as an integer linear combination of n_{t-1} and n_{t-2}. Since the preceding equation expresses n_{t-1} as an integer linear combination of n_{t-2} and n_{t-3},

$$n_{t-1} = n_{t-3} - n_{t-2}q_{t-2},$$

substituting, we obtain

$$\begin{aligned}(n_0, n_1) &= n_{t-2} - (n_{t-3} - n_{t-2}q_{t-2})q_{t-1} \\ &= (q_{t-1}q_{t-2} + 1)n_{t-2} - q_{t-1}n_{t-3},\end{aligned}$$

that is, as an integer linear combination of n_{t-2} and n_{t-3}. Continuing up the sequence of divisions we finally obtain (n_0, n_1) as an integer linear combination of n_0 and n_1. This aspect of the Euclidean algorithm is important both theoretically and computationally. But while it is easy to see by guessing that

$$(7, 3) = 1 = 7x + 3y$$

with $x = -2, y = 5$ (or $x = 4, y = -9$), the computation is not generally so easy. For example, computing $(23134, 15257)$ by the Euclidean algorithm entails

$$\begin{aligned}23134 &= 15257 \cdot 1 + 7877 \\ 15257 &= 7877 \cdot 1 + 7380 \\ 7877 &= 7380 \cdot 1 + 497\end{aligned}$$

$$\begin{aligned} 7380 &= 497 \cdot 14 + 422 \\ 497 &= 422 \cdot 1 + 75 \\ 422 &= 75 \cdot 5 + 47 \\ 75 &= 47 \cdot 1 + 28 \\ 47 &= 28 \cdot 1 + 19 \\ 28 &= 19 \cdot 1 + 9 \\ 19 &= 9 \cdot 2 + 1 \\ 9 &= 1 \cdot 9 + 0. \end{aligned}$$

So $(23134, 15257) = 1$ and it is really quite tedious to substitute backwards to obtain

$$23134 \cdot (-1627) + 15257 \cdot (2467) = 1.$$

A more convenient way of computing (a, b) in the form $ax + by$ is as follows. Form the matrix

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

Now if there is a matrix $\begin{pmatrix} x & y \\ r & s \end{pmatrix}$, $x, y, r, s \in \mathbb{Z}$ such that

$$\begin{pmatrix} x & y \\ r & s \end{pmatrix} \begin{pmatrix} a & 1 & 0 \\ b & 0 & 1 \end{pmatrix} = \begin{pmatrix} d & x & y \\ 0 & r & s \end{pmatrix}$$

with $d \in \mathbb{N}$, then $d = ax + by$ and hence d is divisible by any common divisor of a and b. Further, if $\begin{pmatrix} x & y \\ r & s \end{pmatrix}$ is nonsingular and has inverse $\begin{pmatrix} X & Y \\ R & S \end{pmatrix}$, with entries X, Y, R, S in $\mathbb{Z}$, then

$$\begin{pmatrix} a & 1 & 0 \\ b & 0 & 1 \end{pmatrix} = \begin{pmatrix} X & Y \\ R & S \end{pmatrix} \begin{pmatrix} d & x & y \\ 0 & r & s \end{pmatrix},$$

and from this we see that $a = Xd$ and $b = Rd$, which means that d is a common divisor of a and b; hence $d = (a, b)$.

To implement this matrix process we do not explicitly determine the matrix $\begin{pmatrix} x & y \\ r & s \end{pmatrix}$, but simply recall from linear algebra that such a matrix having entries in $\mathbb{Z}$ exists iff the matrix

$$A = \begin{pmatrix} a & 1 & 0 \\ b & 0 & 1 \end{pmatrix}$$

can be *row reduced* to the form

$$B = \begin{pmatrix} d & x & y \\ 0 & r & s \end{pmatrix},$$

where the permissible row operations are the following:

i) interchange of two rows,

ii) addition of any integer multiple of one row to another row,

iii) changing the sign of either row.

(If we were working with matrices with entries from $\mathbb{Q}, \mathbb{R}$, or $\mathbb{C}$, we could extend type iii) operations to include multiplication by any nonzero multiplier, since the result of this action is reversible. In the present case we allow only multiplication by ± 1 since these are the only multiplications which are reversible in $\mathbb{Z}$. This guarantees that the inverse matrix $\begin{pmatrix} X & Y \\ R & S \end{pmatrix}$ does indeed have entries in $\mathbb{Z}$.)

We show that in fact the matrix A with $a, b \in \mathbb{N}$ can always be row reduced to the form B by operations of types i), and ii) only. Suppose $a \geq b > 0$. Then by the division algorithm we can use a type ii) operation to replace a by $r = -qb + a$ where $q \in \mathbb{Z}$ and $0 \leq r < b$. Hence repeated application of type ii) operations produces a sequence of matrices where at each step $\max(a, b)$ is strictly less than at the preceding step unless one of a or b is zero, which must eventually occur. If necessary we then interchange the two rows to obtain the matrix B.

Example. To compute $(714, 2177)$ we can use row operations to obtain

$$\begin{pmatrix} 714 & 1 & 0 \\ 2177 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 714 & 1 & 0 \\ 35 & -3 & 1 \end{pmatrix} \to \begin{pmatrix} 14 & 61 & -20 \\ 35 & -3 & 1 \end{pmatrix}$$

$$\to \begin{pmatrix} 14 & 61 & -20 \\ 7 & -125 & 41 \end{pmatrix} \to \begin{pmatrix} 0 & 311 & -102 \\ 7 & -125 & 41 \end{pmatrix}$$

$$\to \begin{pmatrix} 7 & -125 & 41 \\ 0 & 311 & -102 \end{pmatrix}$$

and conclude that

$$7 = (714, 2177) = 714(-125) + 2177(41).$$

In terms of computational complexity, that is in terms of the number of actual operations performed, all of these ways of implementing the Euclidean algorithm are essentially the same. For machine or hand calculation probably the matrix method is the easiest and least error prone.

The concepts of GCD and LCM can be extended from $\mathbb{N}$ to all of $\mathbb{Z}$ by defining $(a, b) = \pm(|a|, |b|)$. Of course we loose the uniqueness of (a, b) which we have if $a, b \in \mathbb{N}$, and thus $\mathbb{Z}$ is not a lattice, in fact it is not even a poset, since for example, $-1| + 1$ and $+1| - 1$, but $+1 \neq -1$. The concept is still important however.

Exercises Section 2.3

1. Show that if $n \in \mathbb{Z}$ then $3n^2 - 1$ is not a perfect square.

2. Use the matrix method to obtain x and y such that $23134x + 15257y = 1$.

3. a) If $a_1, \ldots, a_n$ are nonnegative integers how should $\gcd(a_1, \ldots, a_n)$ be defined?
b) Show that $(a, b, c) = ((a, b), c)$.
c) Show that if $a_1, \ldots, a_n \in \mathbb{N}$, then there are $x_1, \ldots, x_n \in \mathbb{Z}$ such that

$$(a_1, \ldots, a_n) = a_1x_1 + \cdots + a_nx_n.$$

d) Devise a matrix method for computing $(a_1, \ldots, a_n)$ and $x_1, \ldots, x_n$ satisfying c) above.

4. Show that if $(a^2, b^2) = 1$ then $(a, b) = 1$ and, more generally, for $n \in \mathbb{N}$,

$$(a^n, b^n) = (a, b)^n.$$

5. Diophantine equations are equations for which integer solutions are sought. The linear diophantine equations in two unknowns is

$$ax + by = c$$

where $a, b, c \in \mathbb{Z}$. An example is $2x + y = 1$. It is easily verified that this equation has solutions $x = -1, y = 3$, and more generally $x = -1 + t, y = 3 - 2t$ are solutions for any $t \in \mathbb{Z}$. On the other hand the equation $4x + 6y = 5$ has no integer solutions. (Why?)

a) Prove that the equation $ax + by = c$ has a solution iff $(a, b)|c$.
b) Assume that the equation $ax + by = c$, has a solution, i.e.: $(a, b)|c$. Let

$$a_1 = \frac{a}{(a, b)}, \quad b_1 = \frac{b}{(a, b)}, \quad k = \frac{c}{(a, b)}.$$

Then $(a_1, b_1) = 1$ so there are $x_0, y_0 \in \mathbb{Z}$ such that $a_1x_0 + b_1y_0 = 1$. Show that $x = x_0k + b_1t, y = y_0k - a_1t$ is a solution of the equation for all $t \in \mathbb{Z}$.
c) Show that every solution of $ax+by = c$ has the form $x = x_0k+b_1t, y = y_0k-a_1t$ for some $t \in \mathbb{Z}$ and where a_1, b_1, x_0, y_0 are chosen as in b) above.

2.4 Modules

Modules provide another important way of examining the integers and the relation of divisibility. A *module of integers* — which we will simply call a module — is a nonempty subset M of the integers $\mathbb{Z}$ which is closed under subtraction; i.e.: if $a, b \in M$ then $a - b \in M$. For example the set of all multiple of the positive integer m

is a module: if pm and qm are multiples of m then their difference $pm-qm = (p-q)m$ is also a multiple of m. We denote the module of all integer multiples of m by (m). Thus $(m) = \{0, \pm m, \pm 2m, \pm 3m, \ldots\}$. If $m = 0$ then $(0) = \{0\}$ obviously. (0) is the smallest module and is contained in every module, for if $a \in M$ then $a - a = 0 \in M$. At the other extreme $(1) = \mathbb{Z}$ and contains every module. The most important fact about modules of integers is the following theorem which asserts that every module is of the form (m) for some $m \in \mathbb{N}$.

Theorem 2.4.1 *If M is a module of integers, them either $M = (0)$ or M contains a least positive element m and $M = (m)$.*

Proof. First notice that a module is closed under sums as well as differences, for if $a, b \in M$ then $-b = 0 - b \in M$ so $a + b = a - (-b) \in M$. If $M \neq (0)$, then M contains some nonzero integer a and hence, as noted above $a - a = 0$ and hence $0 - a = -a$. Since one of a and $-a$ is positive, we conclude that M contains some positive integer and hence by the well-ordering principle, M contains a *least* positive integer m. Let a be any element of M. Then by the division algorithm, $a = qm + r$ where $0 \leq r < m$ and $q \in \mathbb{Z}$. But qm is just plus or minus the $|q|$- fold sum of m so $qm \in M$ and hence $r = a - qm \in M$. Since m is the least positive element of M we are forced to conclude that $r = 0$ and hence $a = qm$. Therefore $M = (m)$.

Now we shall give still another way of showing both the existence of $\gcd(a, b)$ and its representation in the form $ax + by$. We do this for arbitrary $a, b \in \mathbb{Z}$ using the extended sense of the gcd given at the end of Section 2.3. Suppose we are given $a, b \in \mathbb{Z}$. Define

$$M = \{ax + by : x, y \in \mathbb{Z}\}.$$

Observe that M is a module, for if $ax_1 + by_1$ and $ax_2 + by_2$ are members of M, then their difference $a(x_1 - x_2) + b(y_1 - y_2)$ is also in M. Now $M = \{0\}$ can occur only if both a and b are both zero and their gcd is the sole member of M. Otherwise $M \neq \{0\}$ and then certainly M contains positive integers so by Theorem 2.4.1 $M = (d)$ where $d = ax_0 + by_0$ is the least positive member of M, for some x_0 and y_0. From this representation it follows that d is divisible by any common divisor of a and b. Finally, $a = a \cdot 1 + b \cdot 0$ and $b = a \cdot 0 + b \cdot 1$ are in M so both a and b are divisible by d, i.e.: d is a common divisor. Hence $(a, b) = \pm d$.

Now let us compare the lattice $(\mathbb{N}, |)$ of nonnegative integers with respect to the order relation of divisibility with the ordered set $(\mathcal{M}, \subset)$ of all modules of $\mathbb{Z}$ with respect to the order relation of set containment $\subset$. Certainly $\mathcal{M}$ is partially ordered by containment — any collection of subsets of a set is partially ordered by containment. Is it a lattice? It is, and even more, as we now observe, for there is obviously a 1-1 correspondence

$$d \Longleftrightarrow (d)$$

between nonnegative integers $d \in \mathbb{N}$ and modules (d) of $\mathbb{Z}$. Moreover it is readily verified that for $d, d' \in \mathbb{N}$

$$d|d' \Longleftrightarrow (d') \subset (d).$$

This means that the "larger" the integer d, the "smaller" the module (d). This in turn means that d is an upper bound for $a, b \in \mathbb{N}$ (i.e.: is a common multiple of both a and b) iff (d) is a lower bound for $(a), (b) \in \mathcal{M}$ (i.e.: is contained in both (a) and (b)).Likewise d is a lower bound for a, b iff (d) is an upper bound for $(a), (b)$. From this we infer that $d = \gcd(a, b)$ is the glb of a and b with respect to divisibility (i.e.: is divisible by every divisor of both a and b) iff (d) is the lub of (a) and (b) (i.e.: is contained in every module containing both (a) and (b)). Likewise, interchanging lower bounds and upper bounds and glbs and lubs we see that $m = \text{lcm}(a, b)$ iff (m) is the $\text{glb}((a), (b))$ in $\mathcal{M}$. This shows that $(\mathcal{M}, \subset)$ is a lattice in which

$$\text{lub}((a), (b)) = ((a, b)) \text{ and } \text{glb}((a), (b)) = ([a, b]).$$

Moreover the lattices $(\mathbb{N}, |)$ and $(\mathcal{M}, \subset)$ are essentially the same, the difference being that the picture of one can be obtained by turning upside down the picture of the other: larger elements in one correspond to smaller elements in the other. In particular $(0) = \{0\}$ is the least module while 0 is the largest integer in the sense of divisibility. Likewise $(1) = \mathbb{Z}$ is the largest module while 1 is the least element in $(\mathbb{N}, |)$. We express the relation between these two lattices by say that they are *dually isomorphic*.

Exercises Section2.4

1. Show that a set of integers which is closed under the operation of addition need not consist of integer multiples of one fixed element of the set.

2. Which of the following sets are modules. Find the least positive member of those sets which are modules.
   a) $\{m \in \mathbb{Z} : \text{ for some } n \in N,\ 64 | m^n\}$
   b) $\{m \in \mathbb{Z} : (m, 7) = 1\}$
   c) $\{m \in \mathbb{Z} : m | 24\}$
   d) $\{m \in \mathbb{Z} : 6 | m \text{ and } 24 | m^2\}$
   e) $\{m \in \mathbb{Z} : 9 | 21m\}$

3. Prove that if (a) and (b) are modules, then so is their intersection $(a) \cap (b)$. Prove that $(a) \cap (b)$ is the glb of (a) and (b) and that $(a) \cap (b) = ([a, b])$.

4. Prove that the lattice $(\mathcal{M}, \subset)$ is distributive. See Section 2.2.

2.5 Counting; Euler's ϕ-function

In Section 1.8 we discussed the Principle of inclusion and exclusion and Example 2 of that section applied it to count the number of primes less than 100. Now we extend the method of that example to a more general counting problem. In particular we

are interested in some properties of the Euler ϕ-function (also called the Euler *totient* function). This important function is defined for each positive integer n and its value $\phi(n)$ is the number of integers $m \leq n$ which are relatively prime to n. It is easy to compute the following short table of values of $\phi(n)$:

$n =$	1	2	3	4	5	6	7	8	9	10	11	12
$\phi(n) =$	1	1	2	2	4	2	6	4	6	4	10	4

To investigate the ϕ-function we begin by recalling a few facts about divisibility used in Example 2 of Section 1.8. As in that example let $[x]$ = the largest integer $\leq x$. Also notice, as in that example, that if $n, a \in \mathbb{Z}^+$, then $n = aq + r,\ 0 \leq r < a$, so that the positive integers $x \leq n$ which are divisible by a are $x = a, 2a, \ldots, qa$;, i.e.: there are q of them. Since

$$\frac{n}{a} = q + \frac{r}{a},\ 0 \leq \frac{r}{a} < 1,$$

we see that

$\left[\frac{n}{a}\right] = q =$ the number of integers $\leq n$ which are divisible by a.

Also, an integer n is divisible by each of a and b iff n is divisible by $\text{lcm}(a, b)$ and hence if a and b are relatively prime this means that n is divisible by each of a and b iff $ab|n$. Hence if $(a, b) = 1$, $\left[\frac{n}{ab}\right]$ = the number of integers $\leq n$ which are divisible by both a and b. In general (by Exercise 7 at the end of this section) if $a_1, \ldots, a_r$ are pairwise relatively prime, then

$$\left[\frac{n}{a_1 \cdots a_r}\right] = \text{ the number of integers } \leq n \text{ which are divisible by each of } a_1, \ldots, a_r.$$

The Principle of inclusion and exclusion (Theorem 1.8.1) then directly yields the following result. (We are taking S to be the set of positive integers $\leq n$ and A_i to be those divisible by a_i.)

Theorem 2.5.1 *The number of positive integers $\leq n$ and not divisible by any of the pairwise relatively prime integers $a_1, \ldots, a_r$, is*

$$[n] - \sum_i \left[\frac{n}{a_i}\right] + \sum_{i<j} \left[\frac{n}{a_i a_j}\right] - \cdots + (-1)^r \left[\frac{n}{a_1 \cdots a_r}\right].$$

Now let $a_1, \ldots, a_r$ be the various prime factors $p_1, \ldots, p_r$ of n and make the following observations:

i) the number of positive integers $\leq n$ and not divisible by any of $p_1, \ldots, p_r$ is precisely the number of integers $m \leq n$ which are relatively prime to n and this is just the value of $\phi(n)$.

ii)

$$\left[\frac{n}{p_i \cdots p_k}\right] = \frac{n}{p_i \cdots p_k}.$$

Hence Theorem 2.5.1 yields

$$\begin{aligned}
\phi(n) &= n - \sum_i \frac{n}{p_i} + \sum_{i<j} \frac{n}{p_i p_j} - \cdots + (-1)^r \frac{n}{p_1 \cdots p_r} \\
&= n\left[1 - \sum_i \frac{1}{p_i} + \sum_{i,j} \frac{1}{p_i p_j} - \cdots + (-1)^r \frac{1}{p_1 \cdots p_r}\right] \\
&= n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right).
\end{aligned}$$

This formula, due to Legendre, is one of the most important of arithmetic. We state it compactly in the following form.

Theorem 2.5.2

$$\phi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

The subscript $p|n$ of the product indicates that the product is to be taken over all primes p dividing n.

Still another important property of the ϕ function is the following theorem, attributed to Gauss.

Theorem 2.5.3 *(Gauss' formula).*

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

For example, if $n = 12$ we have for the divisors d of 12:

$d =$	1	2	3	4	6	12
$\phi(d) =$	1	1	2	2	2	4

and the sum of the second row is 12.

We prove Gauss' formula by the following ingenious counting argument. Consider the list of fractions:

$$\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n}.$$

A fraction $\frac{k}{n}$ of this list is in "lowest terms" iff $(k,n)=1$ and hence $\phi(n)$ of the n fractions are in lowest terms. More generally, let d be any divisor of n. Then from the $\frac{n}{d}$ fractions

$$\frac{d}{n}, \frac{2d}{n}, \cdots, \frac{\left(\frac{n}{d}\right)d}{n}$$

of the original list, when we cancel the d we obtain

$$\frac{1}{\left(\frac{n}{d}\right)}, \frac{2}{\left(\frac{n}{d}\right)}, \cdots, \frac{\left(\frac{n}{d}\right)}{\left(\frac{n}{d}\right)}$$

and $\phi\left(\frac{n}{d}\right)$ of these are in lowest terms. Hence each divisor d of n counts exactly $\phi\left(\frac{n}{d}\right)$ of the n members of the original list. (We showed this above for $d=1$. For $d=n$ we obtain the list consisting only of $\frac{n}{n}=\frac{1}{1}$ and count $\phi\left(\frac{1}{1}\right)=1$ members of the original list.)

Next observe that each $\frac{k}{n}$ in the original list appears as $\frac{\frac{k}{d}}{\frac{n}{d}}$ in lowest terms for the divisor $d=(k,n)$, i.e.: each member of the original list is counted at least once as we add up $\phi\left(\frac{n}{d}\right)$ over all divisors d of n.

Finally, we must see that no fraction is counted twice. If this were so we would have d_1, d_2 distinct divisors of n and

$$\frac{k}{\left(\frac{n}{d_1}\right)} = \frac{l}{\left(\frac{n}{d_2}\right)}$$

both in lowest terms. But then

$$k\left(\frac{n}{d_2}\right) = l\left(\frac{n}{d_1}\right)$$

from which we have $k|l$ since $\left(k, \frac{n}{d_1}\right)=1$ and $l|k$ since $\left(l, \frac{n}{d_2}\right)=1$. Therefore $k=l$ so $d_1=d_2$, a contradiction. Hence each of the original fractions is counted exactly once so that $n=\sum_{d|n}\phi\left(\frac{n}{d}\right)$. This is equivalent to the statement of Theorem 2.5.3 since the integers $\frac{n}{d}$ run through all divisors of n as d runs through all divisors of n.

Arithmetical functions are integer valued functions defined on $\mathbb{Z}^+$ and which express some arithmetical property of the argument. The Euler ϕ function is just one of many interesting arithmetical functions. The exercises introduce some other arithmetical functions.

Exercises Section 2.5

1. An arithmetical function f is said to be multiplicative if $(m,n)=1 \Rightarrow f(mn)=f(m)f(n)$. Show that the Euler ϕ function is multiplicative.

2. Prove that if f is a multiplicative arithmetical function and $F(n) = \sum_{d|n} f(d)$, then F is a multiplicative function.

3. The function $d(n)$ is defined to be the number of divisors of n.
   a) Show that $d(n)$ is multiplicative.
   b) If $n = p_1^{n_1} \cdots p_k^{n_k}$ then show that $d(n) = (n_1 + 1)(n_2 + 1) \cdots (n_k + 1)$.
   (Cf. Exercise 1 of Chapter 1, section 4.)
   c) Find the smallest integer having 36 divisors.
   d) Show that a number n is a perfect square iff $d(n)$ is odd.

4. The function $\sigma_k(n) = \sum_{d|n} d^k$, the sum of the k-th powers of the divisors of n. Notice that $d(n) = \sigma_0(n)$. Show that $\sigma_k(n)$ is multiplicative.

5. A positive integer is said to be perfect if it is equal to the sum of its proper divisors, i.e.: divisors less than itself. This means that n is perfect if $\sigma_1(n) = 2n$. Prove the following theorem usually attributed to Euclid: If $2^n - 1$ is prime then $2^{n-1}(2^n - 1)$ is perfect. (Euler proved that every even perfect number is of the form $2^{n-1}(2^n - 1)$, i.e.: is of Euclid's type. Odd perfect numbers are either very rare or are nonexistent: none have ever been found.)

6. The *n-th roots of unity* are the complex numbers z which are roots of $z^n = 1$. Recall that these are the numbers

$$z_k = e^{2\pi ik/n}, \quad k = 0, 1, \ldots, n-1.$$

z_k is a *primitive* n-th root of unity if each of the n-th roots of unity is some power of z_k, i.e.: if z_k multiplicatively generates all of the n-th roots of unity. Show that there are $\phi(n)$ primitive n-th roots of unity.

7. Use induction to show that if $a_1, \ldots, a_r$ are pairwise relatively prime, then

$$\left[\frac{n}{a_1 \cdots a_r}\right] = \text{ the number of integers} \leq n \text{ which are divisible by each of } a_1, \ldots, a_r.$$

Note: It is important that $a_1, \ldots, a_r$ be pairwise relatively prime and not just relatively prime. For example, the integers 2,3, and 4 are relatively prime but not pairwise relatively prime and the number if integers ≤ 12 divisible by each of them is 1 while

$$\left[\frac{12}{2 \cdot 3 \cdot 4}\right] = 0.$$

2.6 Congruences

In Chapter 1 we made frequent use of the equivalence relations of the integers determined by congruence modulo m. Now we study these relations more systematically.

Recall that for $m \in \mathbb{N}$, integers $a, b \in \mathbb{Z}$ are said to be congruent modulo m iff $m|(a-b)$. We express this by writing $a \equiv b \ (m)$ or $a \equiv b \ mod(m)$. $a \not\equiv b \ (m)$ then means that $m \not|(a-b)$. Notice that if $m = 1$ then $a \equiv b \ (m)$ for any $a, b \in \mathbb{Z}$; at the other extreme if $m = 0$ then $a \equiv b \ (m)$ iff $a = b$.

By the division algorithm (Theorem 2.3.1), for $m \in \mathbb{Z}^+$ every integer a leaves a unique remainder r, $0 \leq r < m$, on division by m. We show that a and b are congruent modulo m iff they leave the same remainder on division by m. First suppose

$$a = mq + r, \quad b = mq' + r, \ \ 0 \leq r < m.$$

Then $a - b = m(q - q')$ so $m|a - b$ and hence $a \equiv b \ \ (m)$. Conversely, suppose $a \equiv b \ (m)$. Let

$$a = mq + r \ \ b = mq' + r', \ \ 0 \leq r, r' < m$$

where we may suppose $r' \leq r$. Then

$$a - b = m(q - q') + (r - r').$$

Since m divides $a - b$ it follows that $m|r - r'$. But $0 \leq r - r' < m$ and this means that $r - r' = 0$. Therefore $r = r'$. The remainder left by dividing a by m is frequently called the *residue of a modulo m*. It is often denoted by $a \bmod (m)$.

The following theorem gives the most important properties of congruence modulo m.

Theorem 2.6.1 *a) Congruence modulo m is an equivalence relation:*
$a \equiv a \ (m)$ *for all* $a \in \mathbb{Z}$ *(reflexivity)*
$a \equiv b \ (m) \Rightarrow b \equiv a \ (m)$ *(symmetry)*
$a \equiv b \ (m)$ *and* $b \equiv c \ (m) \Rightarrow a \equiv c \ (m)$ *(transitivity)*

b) Congruence modulo m has the "substitution property" with respect to addition and multiplication:
$a \equiv b \ (m) \Rightarrow a + c \equiv b + c \ (m)$
$a \equiv b \ (m) \Rightarrow ac \equiv bc \ (m)$

The proof of this theorem is an easy exercise. The following corollary is a simple consequence of it.

Corollary 2.6.1 *If* $p(x_1, \ldots, x_n)$ *is a polynomial in the variables* $x_1, \ldots, x_n$ *and having integer coefficients, then*

$$a_i \equiv b_i \ (m) \text{ for } i = 1, \ldots, n \ \ \Rightarrow p(a_1, \ldots, a_n) \equiv p(b_1, \ldots, b_n) \ \ (m).$$

Theorem 2.6.1 a) asserts that congruence modulo m determines an equivalence relation and hence a partition on $\mathbb{Z}$. Since two integers are incongruent modulo m iff they have different residues modulo m it follows that there are exactly m blocks of

the resulting partition, i.e.: the partition has index m. The elements of a particular block consist of all integers in $\mathbb{Z}$ having the same remainder on division by m. For example, congruence modulo 2 partitions the integers into two blocks: the odd and even integers respectively. The blocks of this partition are called *residue classes modulo* m and the residue class (block) containing the integer a is denoted by $a/(m)$, i.e.:

$$a/(m) = \{b \in \mathbb{Z} : a \equiv b \ \ (m)\}.$$

This means that we have the following obvious but important connection between congruence modulo m and equality of residue classes:

$$a/(m) = b/(m) \Leftrightarrow a \equiv b \ \ (m).$$

Thus we can translate statements about equality of residue classes into statements about congruence modulo m and visa versa. We will frequently make use of this simple observation.

It is frequently useful to select one representative from each of the m residue classes determined by congruence modulo m. Such a selection is called a *complete system of residues modulo* m. It follows that a set of m integers is a complete system of residues modulo m iff no two of the integers are congruent modulo m. For example the set $\{0, 1, 2, 3, 4, 5, 6\}$ constitutes the most obvious complete system of residues modulo 7, but $\{7, 8, -5, 10, 18, 40, 48\}$ is also a complete system modulo 7.

With reference to Theorem 2.6.1 b) the converse of the substitution property for addition also holds and is usually called the cancellation law for addition:

$$a + c \equiv b + c \ \ (m) \Rightarrow a \equiv b \ \ (m).$$

This is obtained by simply adding the integer $-c$ to both sides of $a \equiv b \ \ (m)$, which is allowed by the substitution property for addition.

It is not true in general that for $c \neq 0$, if $ac \equiv bc \ \ (m)$, then $a \equiv b \ \ (m)$. For example $2 \cdot 3 \equiv 4 \cdot 3 \ \ (6)$ but $2 \not\equiv 4 \ \ (6)$. We do however have the following very important special case where cancellation does work.

Theorem 2.6.2 *If* $(c, m) = 1$ *and* $ac \equiv bc \ \ (m)$ *then* $a \equiv b \ \ (m)$. *More generally, if* $(c, m) = d$ *and* $ac \equiv bc \ \ (m)$ *then* $a \equiv b \ \ (\frac{m}{d})$.

Proof. $ac \equiv bc \ \ (m) \Rightarrow m|(a-b)c \Rightarrow \frac{m}{d}|(a-b)\frac{c}{d}$. But $(\frac{m}{d}, \frac{c}{d}) = 1$ so by Euclid's divisibility theorem (Theorem 2.1.2), $\frac{m}{d}|(a-b)$.

We are all familiar with the following laws for computing with even and odd integers:

$$\begin{array}{rcrcrcrcr} \text{even} + \text{even} & = & \text{odd} + \text{odd} & = & \text{even}, & \quad \text{even} + \text{odd} & = & \text{odd} \\ \text{even} \cdot \text{even} & = & \text{even} \cdot \text{odd} & = & \text{even}, & \quad \text{odd} \cdot \text{odd} & = & \text{odd}. \end{array}$$

These identities may be regarded not as properties about the ordinary integers, but as properties of a new "addition" and 'multiplication" in a new algebra of the two elements "even" and "odd". This algebra can also be constructed as an algebra of residues modulo 2. The even integers are just those which leave a remainder of 0 on division by 2 while the odds are just those leaving a remainder of 1. These two remainders can be added and multiplied in the ordinary way, provided the results are then replaced by the appropriate remainders modulo 2. This gives operation tables

$$\begin{array}{rclcrcl} 0+0 & = & 1+1 & = & 0 & \quad 0+1 & = & 1 \\ 0\cdot 0 & = & 0\cdot 1 & = & 0 & \quad 1\cdot 1 & = & 1 \end{array}$$

which are essentially the same as the tables above for even and odd. Alternatively, we can say that the new equation $1+1=0$ is just another way of writing the congruence $1+1 \equiv 0 \pmod{2}$ which, from our basic relation connecting equality of residue classes with congruences, can be written as $1/(2)+1/(2)=0/(2)$, another way of writing odd + odd = even. Corresponding to the algebra of even and odd, for any $m \in \mathbb{Z}^+$ there is a similar algebra $\mathbb{Z}/(m)$ whose elements consist of the m residue classes $0/(m), 1/(m), \ldots, m-1/(m)$ and where we define addition and multiplication of residue classes by adding or multiplying arbitrary elements $a' \in a/(m)$ and $b' \in b/(m)$ and defining

$$\begin{array}{rcl} a/(m)+b/(m) & = & (a'+b')/(m), \quad \text{and} \\ a/(m)\cdot b/(m) & = & (a'\cdot b')/(m). \end{array}$$

This means that the sum (or product) of two residue classes is obtained by choosing representatives from each, computing their sum (or product) as usual in $\mathbb{Z}$, and taking the residue class containing the result of this computation to be the sum (or product) of the given classes. In order to justify these definitions we must show that the sum $(a'+b')/(m)$ and product $(a'\cdot b')/(m)$ are independent of the choices of $a' \in a/(m), b' \in b/(m)$. But if $a', a'' \in a/(m)$ and $b', b'' \in b/(m)$, then by the substitution property of Theorem 2.6.1 we have $a'+b' \equiv a'+b'' \equiv a''+b'' \pmod{m}$ and $a'b' \equiv a'b'' \equiv a''b'' \pmod{m}$. Hence the definitions are justified. Because the definition is independent of the choice of a' and b' we can take $a'=a$ and $b'=b$ and restate the definitions as

$$\begin{array}{rcl} a/(m)+b/(m) & = & (a+b)/(m), \quad \text{and} \\ a/(m)\cdot b/(m) & = & (a\cdot b)/(m). \end{array}$$

The resulting system $\mathbb{Z}/(m)$ is called the "integers modulo m" and generalizes the algebra $\mathbb{Z}/(2)$ of even and odd.

An alternate equivalent algebra, usually denoted by $\mathbb{Z}_m$ consists of the integers $0, 1, \ldots, m-1$ where we define $a+b$ to be the unique remainder $(a+b) \bmod (m)$ left on dividing the ordinary sum of a and b by m, and likewise the product $a\cdot b$ to be the unique remainder $(a\cdot b) \bmod (m)$ obtained by dividing the ordinary product by m. The resulting algebra generalizes the algebra of 0 and 1 as described above and aside

from using the complete system of residues $\{0, 1, \ldots, m-1\}$ modulo m instead of the residue classes $0/(m), 1/(m), \ldots, (m-1)/(m)$, is no different from $\mathbb{Z}/(m)$. Specifically this means that if $a \cdot b = c$ in $\mathbb{Z}_m$, then $a/(m) \cdot b/(m) = c/(m)$ in $\mathbb{Z}/(m)$ (and likewise for $+$).

Finally, notice that we are using the same symbols $+$ and $\cdot$ to represent operations in three different, though closely related, systems: $\mathbb{Z}$, $\mathbb{Z}/(m)$, and $\mathbb{Z}_m$. The context will always make clear which meaning is intended.

Either way, many of the usual arithmetical "laws" of the system of integers $\mathbb{Z}$ with the usual addition and multiplication are inherited by the integers modulo m. Among these are the following (stated for $\mathbb{Z}/(m)$:

$a/(m) + b/(m) = b/(m) + a/(m)$
$a/(m) \cdot b/(m) = b/(m) \cdot a/(m)$ (commutativity)
$a/(m) + (b/(m) + c/(m)) = (a/(m) + b/(m)) + c/(m)$
$a/(m) \cdot (b/(m) \cdot c/(m)) = (a/(m) \cdot b/(m)) \cdot c/(m)$ (associativity)
$a/(m) \cdot (b/(m) + c/(m)) = a/(m) \cdot b/(m) + a/(m) \cdot c/(m)$ (distributivity)
$0/(m) + a/(m) = a/(m)$
$0/(m) \cdot a/(m) = 0/(m)$
$1/(m) \cdot a/(m) = a/(m)$.

The proofs of these laws consist merely of observing how the inheritance takes place. For example, for the distributive law we have

$$\begin{aligned} a/(m) \cdot (b/(m) + c/(m)) & \\ &= a/(m) \cdot ((b+c)/(m)) \text{ by the definition of } + \\ &= (a(b+c))/(m) \text{ by the definition of } \cdot \\ &= (ab + ac)/(m) \text{ by the distributive law for } \mathbb{Z} \\ &= ab/(m) + ac/(m) \text{ by the definition of } + \\ &= a/(m) \cdot b/(m) + a/(m) \cdot c/(m) \text{ by the definition of } \cdot. \end{aligned}$$

Sometimes $\mathbb{Z}_m$ fails to have properties enjoyed by $\mathbb{Z}$. The most notable example is the cancellation law

$$c \neq 0 \text{ and } ac = bc \Rightarrow a = b$$

which we noted earlier fails in $\mathbb{Z}_6$. But by Theorem 2.6.2 we do have the following very important cases where cancellation works.

Theorem 2.6.3 *If $(c, m) = 1$ and $a/(m) \cdot c/(m) = b/(m) \cdot c/(m)$ then $a/(m) = b/(m)$.*
In particular in $\mathbb{Z}_p$, p a prime, the cancellation law holds.

In $\mathbb{Z}$ the only elements a having multiplicative inverses (i.e.: a solution to $ax = 1$) are ± 1. In $\mathbb{Z}_m$ we have the following situation.

Theorem 2.6.4 *In $\mathbb{Z}_m$ an element a has a multiplicative inverse iff $(a, m) = 1$. Hence $\phi(m)$ elements in $\mathbb{Z}_m$ have multiplicative inverses. The integer p is a prime iff each nonzero element of $\mathbb{Z}_p$ has a multiplicative inverse.*

PROOF. The equation $ax = 1$ is solvable in $\mathbb{Z}_m$ iff the congruence $ax \equiv 1 \ (m)$ is solvable, and this is the case iff $m|ax-1$ for some x, which is true iff there are $x, y \in \mathbb{Z}$ such that $ax + my = 1$ and this is true iff $(a, m) = 1$. The Euler ϕ function counts the number of $a \in \mathbb{Z}_m$ with $(a, m) = 1$. Finally, if $m = p$ a prime, then every nonzero element in $\mathbb{Z}_p$ is relatively prime to p, so all such elements have inverses. Conversely, if each nonzero element of $\mathbb{Z}_m$ has an inverse then each is relatively prime to m which means that m is prime.

It is important to observe that the Euclidean algorithm provides a method for computing the inverse of $a \in \mathbb{Z}_m$ provided $(a, m) = 1$, for it provides an algorithm for finding $x, y \in \mathbb{Z}$ for which

$$ax + my = 1.$$

The integer x found by the Euclidean algorithm gives the solution: we may take the solution to be either $x/(m) \in \mathbb{Z}/(m)$ or to be the least positive integer $x' \in x/(m)$ to be the solution in $\mathbb{Z}_m$. For example to compute the multiplicative inverse of 15,257 in $\mathbb{Z}_{23,134}$ the example in Section 2.3 tells us that $15257(2467) + 23134(-1627) = 1$ so we conclude that $15257 \cdot 2467 = 1$ in $\mathbb{Z}_{23134}$. It is important to note that this is *the* practical way to compute the inverse of an element in $\mathbb{Z}_m$. It is one of the most significant applications of the Euclidean algorithm. Public key encryption systems, discussed in Section 2.8 below, provide an example.

Exercises Section 2.6

1. Prove that $(a, m) = 1$ and $(b, m) = 1$ implies $(ab, m) = 1$.

2. a) Show that the congruence

$$ax \equiv b \ (m)$$

has a solution iff $(a, m)|b$.

b) Show that this congruence has a solution which is unique modulo m iff $(a, m) = 1$.

c) Show that, in general, this congruence has (a, m) solutions which are incongruent modulo m (assuming the condition for solvability $(a, m)|b$ is met).

3. Show that if p is a prime and $a, b \in \mathbb{Z}$, then $(a + b)^p \equiv a^p + b^p \ (p)$. (This means that in $\mathbb{Z}_p$, $(a + b)^p = a^p + b^p$ is a law.

4. Prove that $a \equiv b \ (m_1)$ and $a \equiv b \ (m_2)$ iff $a \equiv b \ ([m_1, m_2])$. More generally prove that for any $n > 1$, $a \equiv b \ (m_i)$ for $i = 1, \ldots, n$ iff $a \equiv b \ ([m_1, \ldots, m_n])$.

5. a) Prove that for every odd integer n, $n^2 \equiv 1 \ (8)$.

b) Prove that if n is an odd integer not divisible by 3, then $n^2 \equiv 1 \ (24)$.

6. Theorem 2.6.1 showed that congruence modulo m is an equivalence relation having the substitution property. In general, such a binary relation on $\mathbb{Z}$ is called a *congruence relation* on $\mathbb{Z}$. Show that every congruence relation R on $\mathbb{Z}$ is congruence modulo m for some $m \in \mathbb{N}$. To be specific let R be a binary relation on $\mathbb{Z}$ such that

aRa for all $a \in \mathbb{Z}$,

$aRb \Rightarrow bRa$ for all $a, b \in \mathbb{Z}$,

aRb and $bRc \Rightarrow aRc$ for all $a, b, c \in \mathbb{Z}$,

$aRb \Rightarrow a + c \; R \; b + c$ and $acRbc$ for all $a, b, c \in \mathbb{Z}$.

Show that for some $m \in \mathbb{N}$, $aRb \Leftrightarrow a \equiv b \; (m)$ for all $a, b \in \mathbb{Z}$. Hint: Begin by showing that the R equivalence class $0/R$ containing 0 is a module.

7. If R and S are equivalence relations on a set X, recall that $R \leq S$ means that for all $a, b \in X$, $aRb \Rightarrow aSb$. If $R \leq S$ we say that R is a *refinement* of S or that S is *coarser* than R.

a) For $m \in \mathbb{N}$ let R_m be the congruence relation of congruence modulo m (as in Exercise 6 above). Show that the system $(\mathcal{R}, \leq)$, where $\mathcal{R}$ is the set of all R_m for $m \in \mathbb{N}$, is a partially ordered set and that for all $m \in \mathbb{N}$, $R_1 \leq R_m \leq R_0$.

b) Show that for all $m, n \in \mathbb{N}$,

$$R_m \leq R_n \Leftrightarrow n|m.$$

Conclude that $(\mathcal{R}, \leq)$ is a lattice which is dually isomorphic to the lattice $(\mathbb{N}, |)$, and that $\text{lub}(R_m, R_n) = R_{(m,n)}$ and $\text{glb}(R_m, R_n) = R_{[m,n]}$. Hint: follow the reasoning used at the end of Section 2.4 to show that the collection of all modules of $\mathbb{Z}$ is a lattice dually isomorphic to the lattice $(\mathbb{N}, |)$.

8. Prove that for $a \not\equiv 0 \; (m)$, the congruence $ax \equiv 0 \; (m)$ has a solution $x \not\equiv 0 \; (m)$ iff $(a, m) \neq 1$.

9. Verify the associative law for addition in $\mathbb{Z}/(m)$.

10. For real numbers x, y let $x \equiv y \; (2\pi)$ mean that $x = y + 2n\pi$ for some $n \in \mathbb{Z}$. Show that this *congruence modulo* 2π is an equivalence relation on the real numbers $\mathbb{R}$. Show that addition of residue classes modulo 2π can be defined as in $\mathbb{Z}/(m)$ but that multiplication of residue classes cannot be so defined. (Hence congruence modulo 2π is *not* a congruence relation on $\mathbb{R}$.)

2.7 Classical theorems about congruences

In this section we discuss Fermat's so called "little theorem" (so called presumably to distinguish it from his famous "last theorem"), Wilson's theorem, and the Chinese remainder theorem. These are important classical theorems of elementary number

theory. Not only are their proofs instructive but they are useful in a variety of areas, for example in coding theory and in the theory of computation. First we prove a simple lemma.

Lemma 2.7.1 *If $a_1, \ldots, a_m$ is a complete system of residues modulo m and $(a, m) = 1$, then $aa_1, \ldots, aa_m$ is also a complete system of residues modulo m.*

PROOF. We need only show that no pair aa_i, aa_j are congruent modulo m and this is immediate from our cancellation theorem (Theorem 2.6.2) since a_i and a_j are incongruent modulo m.

Notice that the lemma does not assert that the residue classes $a_i/(m)$ and $aa_i/(m)$ are equal (unless $a_i \equiv 0\ (m)$). It does assert that the $m-1$ residue classes not containing 0 are rearranged, i.e.: if $a_1 \equiv 0\ (m)$, then the classes $aa_2/(m), \ldots, aa_m/(m)$ equal the classes $a_2/(m), \ldots, a_m/(m)$, but probably in different order.

Theorem 2.7.1 *(Fermat) If p is a prime and $(a, p) = 1$, then $a^{p-1} \equiv 1\ (p)$.*

PROOF. Observe that the set $0, 1, 2, \ldots, p-1$ is a complete system of residues modulo p. Hence, by the lemma, so is $0, a, 2a, \ldots, (p-1)a$ if $(a, p) = 1$. This means that each of the numbers $a, 2a, \ldots, (p-1)a$ is congruent to exactly one of the numbers $1, 2, \ldots, p-1$. Hence by Corollary 2.6.1 (taking $p(x_1, \ldots, x_{m-1}) = x_1 \cdots x_{m-1}$) we have

$$a \cdot 2a \cdots (p-1)a \equiv 1 \cdot 2 \cdots (p-1)\ \ (p),$$

that is

$$(p-1)!a^{p-1} \equiv (p-1)!\ \ (p),$$

so that $a^{p-1} \equiv 1\ \ (p)$ by the cancellation theorem, since $((p-1)!, p) = 1$.

Corollary 2.7.1 *If a is any integer and p is a prime, then $a^p \equiv a\ \ (p)$.*

The corollary is immediate from the theorem for if $(a, p) \neq 1$, then $a \equiv 0\ \ (p)$ so the corollary is trivial in this case. Otherwise multiply $a^{p-1} \equiv 1\ \ (p)$ through by a. Conversely the corollary implies the theorem since if $(a, p) = 1$ we can cancel an a from $a^p \equiv a\ (p)$ to obtain $a^{p-1} \equiv 1\ (p)$. Usually Fermat's theorem refers to either of the equivalent statements.

Now we prove Wilson's theorem by using somewhat more sophisticated reasoning of the same sort used to prove Fermat's theorem.

Theorem 2.7.2 *(Sir John Wilson) If p is a prime then $(p-1)! \equiv -1\ \ (p)$.*

PROOF. If $p = 2$ or 3 the conclusion is obvious so for the remainder of the proof we suppose $p > 3$. Let a be one of the numbers $1, 2, \ldots, p-1$ and examine the congruence $ax \equiv 1 \ \ (p)$. If x goes through the values $1, 2, \ldots, p-1$ then, by the lemma above, ax goes through a complete system of residues modulo p, except for a residue congruent to 0. Hence there is exactly one $x \in \{1, \ldots, p-1\}$ which satisfies the congruence.

Thus the numbers $1, 2, \ldots, p-1$ fall into pairs such that the product of any pair is congruent to 1 modulo p. If the members of a pair are equal, say to a, then $a^2 \equiv 1 \ \ (p)$ so $(a+1)(a-1) = a^2 - 1 \equiv 0 \ \ (p)$ and hence $p|(a+1)$ or $p|(a-1)$. Since $1 \leq a \leq p-1$ we have either $a = 1$ or $a = p-1$.

Excluding 1 and $p-1$ it therefore follows that from the $p-3$ numbers $2, \ldots, p-2$ we can form the product of the $(p-3)/2$ appropriate unequal pairs to obtain

$$2 \cdot 3 \cdot 4 \cdots (p-2) \equiv 1 \ \ (p).$$

Multiplying, we have $(p-1)! \equiv p-1 \equiv -1 \ \ (p)$.

The theorems of Fermat and Wilson were discovered in the seventeenth century. The Chinese remainder theorem is so old its discoverer is unknown, but it was first recorded by the Chinese mathematician Sun-Tsu in the first century A.D. The theorem gives necessary and sufficient conditions for the simplest simultaneous system of linear congruences to be solvable.

Theorem 2.7.3 *(Chinese remainder theorem) The system of simultaneous congruences*

$$\begin{aligned} x &\equiv a_1 \ \ (m_1) \\ &\cdots \\ x &\equiv a_n \ \ (m_n) \end{aligned}$$

is solvable iff $a_i \equiv a_j \ \ ((m_i, m_j))$ *for all* $i, j = 1, \ldots, n$. *(These are called the compatibility conditions of the system.)*

In particular the system is always solvable iff the moduli are pairwise relatively prime.

If the integers a_i all satisfy $0 \leq a_i \leq m_i$, then the theorem asserts that we can find a single integer x such that $x \bmod (m_i) = a_i$ for all i iff the compatibility conditions hold.

PROOF. We will give almost two proofs of the theorem. The first proof of the general theorem consists of showing that the theorem is true iff the lattice $(\mathbb{N}, |)$ is distributive. Since we have already established distributivity in Exercise 6 of Section 2.2 this will give one proof. The problem with this proof is that it gives us no way to actually solve a simultaneous system. On the other hand the importance of this proof is that it demonstrates that the ancient Chinese remainder theorem is actually another way of stating a fundamental property of the integers, namely the distributivity of

the lattice $(\mathbb{N}, |)$. Our second proof will give a constructive procedure for solving the system but we will present it only for the special case where the moduli are pairwise relatively prime.

Before beginning the detailed proof notice that if a system of congruences is solvable then for any i, j, $x \equiv a_i \ (m_i)$ and $x \equiv a_j \ (m_j)$. Hence $x \equiv a_i \ ((m_i, m_j))$ and $x \equiv a_j \ ((m_i, m_j))$ so by transitivity the compatibility conditions $a_i \equiv a_j \ ((m_i, m_j))$ are met. Further, if the system is solvable for arbitrary $a_1, \ldots, a_n$, then the compatibility conditions hold for all a_i and a_j. Taking $a_i = 1$ and $a_j = 0$ we see that m_i and m_j are relatively prime. Hence in the remainder of our proof we only need to establish the "if" direction of the Chinese remainder theorem.

First Proof. First we show that the distributivity of $(\mathbb{N}, |)$ implies the Chinese remainder theorem. We do this by induction on the number n of congruences to be solved. The case $n = 2$ is the first to be considered. Suppose we are given the congruences

$$\begin{aligned} x &\equiv a_1 \ \ (m_1) \\ x &\equiv a_2 \ \ (m_2) \end{aligned}$$

and that the compatibility condition $a_1 \equiv a_2 \ \ ((m_1, m_2))$ holds, which means that $(m_1, m_2) | (a_1 - a_2)$. Since (m_1, m_2) is an integer linear combination of m_1 and m_2, it follows that $a_1 - a_2 = m_1 x + m_2 y$ for some $x, y \in \mathbb{Z}$. Hence $x = a_1 - m_1 = a_2 + m_2$ is clearly a solution to the given pair of congruences.

Now suppose that $n > 2$ and that any system of fewer than n congruences is solvable provided the compatibility conditions are met, and that we are now given the system

$$\begin{aligned} x &\equiv a_1 \ \ (m_1) \\ &\cdots \\ x &\equiv a_n \ \ (m_n) \end{aligned}$$

and that $a_i \equiv a_j \ ((m_i, m_j))$ for all $i, j = 1, \ldots, n$. By the induction hypothesis, there is a solution x' of the first $n - 1$ congruences: $x' \equiv a_i \ \ (m_i) \ \ i = 1, \ldots, n - 1$. Then we need to solve the system

$$\begin{aligned} x &\equiv x' \ \ (m_i), \ \ i = 1, \ldots, n - 1 \\ x &\equiv a_n \ \ (m_n), \end{aligned}$$

which (using Exercise 4, Section 2.6) means we need to solve the pair

$$\begin{aligned} x &\equiv x' \ \ ([m_1, \ldots, m_{n-1}]) \\ x &\equiv a_n \ \ (m_n) \end{aligned}$$

and, by the case $n = 2$, these are solvable provided the compatibility condition

$$x' \equiv a_n \ \ ((m_n, [m_1, \ldots, m_{n-1}]))$$

is met. But by the distributivity of the gcd through the lcm, this condition becomes

$$x' \equiv a_n \ \ ([(m_n, m_1), \ldots, (m_n, m_{n-1})]).$$

To verify that this condition holds we use the facts that

$$x' \equiv a_i \ (m_i) \ \text{ and } \ a_i \equiv a_n \ ((m_i, m_n)) \text{ for all } i < n$$

to conclude that $x' \equiv a_n \ ((m_i, m_n))$ for all $i < n$, and hence (again by Exercise 4, Section 2.6) that the desired condition holds. This completes the proof that the distributivity of $(\mathbb{N}, |)$ implies the Chinese remainder theorem. Notice that from the proof we only used the distributivity to solve more than two simultaneous congruences.

Now we prove the converse, that the Chinese remainder theorem implies distributivity. By Exercise 9 of Section 2.2 it is enough to prove either of the two distributive laws. We will show that for any $a, b, c \in \mathbb{N}$

$$(a, [b, c]) = [(a, b), (a, c)].$$

Let $(a, [b, c]) = d$ and $[(a, b), (a, c)] = m$. First observe that by definition $d|a$ so $d|(a, b)|m$ and therefore $d \leq m$ in any case. Hence we need only show that the Chinese remainder theorem implies $m|d$. To do this it suffices to show that from the Chinese remainder theorem we can infer that congruence modulo $[(a, b), (a, c)]$ entails congruence modulo $(a, [b, c])$.

Thus suppose that for some integers $u, v \in \mathbb{Z}$, $u \equiv v \ ([(a, b), (a, c)])$, which means that u, v satisfy the following pair of congruences (A):

$$\begin{aligned} u &\equiv v \ \ ((a, b)), \\ u &\equiv v \ \ ((a, c)). \end{aligned}$$

To show that $u \equiv v \ ((a, [b, c]))$ it is enough to show that there is a solution z of the two congruences (B):

$$\begin{aligned} z &\equiv u \ \ (a), \\ z &\equiv v \ \ ([b, c]). \end{aligned}$$

(This is true since $(a, [b, c])$ is a common divisor of both a and $[b, c]$ so that if z solves the two congruences (B) then

$$\begin{aligned} z &\equiv u \ \ ((a, [b, c])) \text{ and} \\ z &\equiv v \ \ ((a, [b, c])), \end{aligned}$$

and $u \equiv v \ ((a, [b, c]))$ follows by transitivity.) But the two congruences (B) are equivalent to the three congruences

$$\begin{aligned} z &\equiv u \ \ (a) \\ z &\equiv v \ \ (b) \\ z &\equiv v \ \ (c) \end{aligned}$$

and the compatibility conditions for these are precisely the conditions (A) together with $v \equiv v \ ((b, c))$, which is trivial. This completes the proof that the Chinese remainder theorem implies distributivity of the lattice $(\mathbb{N}, |)$.

Now we give a simple constructive proof that the system

$$\begin{gathered} x \equiv a_1 \ \ (m_1) \\ \cdots \\ x \equiv a_n \ \ (m_n) \end{gathered}$$

always has a solution if $(m_i, m_j) = 1$ for all i, j. Indeed, in this case for each $i = 1, \ldots, n$ we let

$$M_i = m_1 \cdots m_{i-1} m_{i+1} \cdots m_n.$$

Then we have that $(M_i, m_i) = 1$ for all i, so that by Theorem 2.6.4 we can find integers μ_i such that

$$M_i \mu_i \equiv 1 \ \ (m_i).$$

We also observe that for $i \neq j$, $M_i \equiv 0 \ \ (m_j)$. From this we conclude immediately that the integer

$$x = a_1 M_1 \mu_1 + \cdots + a_n M_n \mu_n$$

solves the congruence. This is a constructive solution since the Euclidean algorithm computes the numbers $\mu_1, \ldots, \mu_n$ needed for the solution.

Exercises Section 2.7

1. Let $a_1, \ldots, a_k$ be the $k = \phi(m)$ positive integers less than m for which $(a_i, m) = 1$. If a is any integer such that $(a, m) = 1$ show that for each a_i there is one and only one a_j such that $aa_i \equiv a_j \ \ (m)$.

2. a) Prove Euler's theorem

$$(a, m) = 1 \Rightarrow a^{\phi(m)} \equiv 1 \ \ (m).$$

Hint: From Exercise 1 above observe that each of $a_1, \ldots, a_k$ is congruent modulo m to exactly one of the integers $aa_1, \ldots, aa_k$.

b) Derive Fermat's theorem (Theorem 2.7.1) from Euler's theorem.

c) let p and q be primes and $m = pq$. Prove that

$$(a, m) = 1 \Rightarrow a^{(p-1)(q-1)} \equiv 1 \ \ (m).$$

3. Give a different proof of Fermat's theorem by using induction on the integer a. Hint: Prove the corollary: for a prime p

$$a^p \equiv a \ \ (p)$$

for all integers a.

4. Prove that if $(m_1, m_2) = (a_1, m_1) = (a_2, m_2) = 1$, then the simultaneous congruences

$$\begin{aligned} a_1 x &\equiv b_1 \pmod{m_1} \\ a_2 x &\equiv b_2 \pmod{m_2} \end{aligned}$$

have a common solution.

5. For integers $b, c, i \in \mathbb{N}$, define Gödel's β-function by

$$\beta(b, c, i) = \text{ the remainder left on division of } b \text{ by } 1 + (i+1)c.$$

Thus $\beta(b, c, i) = k$ iff the division algorithm produces

$$b = q(1 + (i+1)c) + k, \quad 0 \leq k < 1 + (i+1)c.$$

The purpose of this exercise is to show that if $k_0, \ldots, k_n$ is an arbitrary finite sequence of nonnegative integers, then we can compute integers $b, c \in \mathbb{N}$ such that

$$\beta(b, c, i) = k_i \quad \text{for all } i = 0, \ldots, n.$$

Thus the β-function gives us a uniform way for encoding finite sequences of integers: for a given sequence we can determine the encoding parameters b and c so that for each i the $\beta(b, c, i)$ returns the i-th term of the sequence. This function, and the fact that for a given sequence, the parameters b, c and the function values $\beta(b, c, i)$ are easily computed, plays a fundamental role in the proof of Gödel's famous theorem which asserts that it is impossible to completely "axiomatize" the theory of elementary arithmetic. The function also plays a basic role in the equally famous proof of the "unsolvability" of the "halting problem" for Turing machines.

To prove that the β-function has these properties, let $k_0, \ldots, k_n$ be given. Then let $m = \max(n, k_0, \ldots, k_n)$ and let $c = m!$. Consider the numbers $u_i = 1 + (i+1)c$ for $i = 0, \ldots, n$.

a) Prove that the numbers $u_0, \ldots, u_n$ are pairwise relatively prime. Hint: Suppose p is a prime which divides both u_i and u_j for some $i \neq j$ and hence $p|(u_i - u_j)$. Show that this leads to a contradiction.

b) Show that $k_i < u_i$ for all i.

c) Show that there is an integer b with the property that for all i, $b \bmod (u_i) = k_i$, i.e.: the remainder on division of b by u_i is k_i, and thus conclude that $\beta(b, c, i) = k_i$ for all i.

2.8 The complexity of arithmetical computation

In Section 2.3 when we introduced the Euclidean algorithm for computing the gcd (and hence lcm), we suggested that this was a more efficient way to compute $\gcd(a, b)$ than to first factor a and b into their prime factors and then determine the minimum

values of the corresponding exponents. The aim of our present discussion is to make this perceived difference in efficiency a little more clear and to give an interesting application of the difference to so called *public key encryption systems*.

First we need to say something about how to measure the efficiency of an algorithm. Because this is not a completely well defined concept — indeed the concept of an algorithm does not appear to have a universally accepted mathematical meaning — we shall be rather informal in our discussion. Also, when a computer executes an algorithm it uses both time, which is usually measured in units consisting of the time required to execute some basic operations, and memory space. Now it is commonly understood that there is a trade off between time and space: we can often perform a computation in less time if we are prepared to use more memory. Also memory space is often not in short supply in many practical situations. More important, space is always reusable while time is not. For these reasons we frequently measure only the time required and assume that as much space as necessary is available. This is of course an artificial restriction but it illustrates why estimates of efficiency are necessarily somewhat inexact. In any case if an arithmetical algorithm A operates on input data of size n we shall want to express the complexity of A, $c(A)$ as some function of n which counts the number of arithmetical operations needed to execute A. More often we are willing to settle for a *worst case analysis* i.e.: an upper bound for the value of such a function given input of size n.

A familiar example from elementary linear algebra illustrates the common approach: if we wish to solve a system of n simultaneous linear equations in n unknowns, it is not difficult to show that using the familiar process of Gaussian elimination, the combined number of multiplications and divisions required is $\frac{1}{3}(n^3+3n^2+2n)$. (This count is obtained in most elementary linear algebra texts.) It is then argued that, relative to the time required for multiplications and divisions, all other operations (addition, subtraction, comparison, storing and retrieving temporary results, etc.) are negligible. It is also agreed that all multiplications and divisions take the same (maximum) time, and hence that for the Gaussian elimination algorithm G and a linear system of size n,

$$c(G)(n) = K[\frac{1}{3}(n^3 + 3n^2 + 2n)],$$

where K is some constant, dependent on the computing environment, which converts the operation count to actual time. Next it is observed that the fraction $\frac{1}{3}$ can be absorbed into the K and, more important, that for $n \geq 3$, the terms $3n^2$ and $2n$ are dominated by the term n^3. Consequently it follows that there is a new constant K such that for all n,

$$c(G)(n) \leq K(n^3).$$

This is usually written as

$$c(G)(n) = O(n^3)$$

and we say that the complexity function for Gaussian elimination is n^3. (In general we say that $f(n) = O(g(n))$ if there is a positive constant K such that $f(n) \leq Kg(n)$ for all n.)

Based on the example above we shall say that an algorithm A has (time) complexity function $g(n)$ if the length of time to execute A is $O(g(n))$ where n is some appropriate parameter which measures the size of the input. Since for sufficiently large n, we have

$$\log(n) \leq \sqrt{n} \leq n \leq n^2 \leq n^3 \leq \cdots \leq n^k \leq \cdots \leq 2^n,$$

it is clear that an algorithm with complexity function $\log(n)$ is considered more efficient than one with complexity function n, n^k, or any polynomial $p(n)$, and an algorithm with polynomial complexity is certainly considered more efficient than one with exponential complexity function. This is borne out by comparing execution times of typical computers for polynomial versus exponential complexity functions. Thus while it is often true that for small values of n, say less than 10, the execution times of both polynomial and exponential algorithms can be measured in a few seconds (with the exponential algorithm sometimes even faster than the polynomial), for even moderate values of n, say between 50 and 100, while polynomial algorithms require only a few minutes to execute, exponential algorithms require millions of centuries. Moreover, even if we greatly increase the speed of the computer, say a thousand fold, the advantage gained allows for no significant increase in n. For this reason we say that polynomial algorithms are practical for ordinary use, while those with only exponential algorithms are intractable, and are unsuitable for ordinary use either now or at any time in the future no matter how fast computers become.

Before we consider algorithms for factoring (as a means of computing gcds) and the Euclidean algorithm, there is one more issue we need to address. In analyzing an algorithm with input an integer n we also need to take into account the way in which n is presented to the computer. In our example above, of Gaussian elimination, this is not important since the system of n equations is simply presented to the computer and any computation actually done on the integer n is minor compared to the multiplications and divisions performed on the coefficients of the system. Thus a more complete analysis of Gaussian elimination might take into account the way in which the coefficients of the linear equations are presented (e.g.: binary, decimal, fixed point, floating point, etc.) as well as the size of the system. This is not ordinarily done since the analysis is usually made under the assumption that the coefficients will always be represented in the same way (usually as floating point binary numbers of fixed size).

In the case of factoring the integer n or of applying the Euclidean algorithm to a pair of integers m, n, the way in which we present these integers becomes more important. More precisely, since the computer works in binary, it is the number b of binary bits needed to represent the integer n which is the appropriate input parameter

for a careful analysis. Now the integer n is represented by precisely b binary bits iff

$$2^{b-1} \leq n < b^b.$$

Since the function $\log = \log_2$ is an increasing function, taking logs of the above gives

$$b - 1 \leq \log(n) < b$$

and hence $b = [\log(n)] + 1$ is the number of bits required to represent n.

Now let us return to the Euclidean algorithm E which computes (m, n) by successive divisions as described in Theorem 2.3.2. We will show in Section 3.4 that for integers $m < n$, if $c(E)(n)$ counts the number of divisions required to compute (m, n), then

$$c(E)(n) = O(\log(n)),$$

i.e.: the complexity is logarithmic in the maximum of the two integers m and n. Then if we take into account the relation between n and the number b of binary bits required to encode n we see that

$$c(E)(n) = O(b),$$

i.e.: the complexity is linear in b. Thus the Euclidean algorithm is regarded as a very efficient algorithm and there is little wonder that it is still so important today after such a long history.

On the other hand, factoring "appears" to be very inefficient: "appears" because there is no known proof that the best algorithms for factoring are exponential and hence intractable. However experience with the best current algorithms for factoring suggests that they are exponential. For example if we use a typical "state of the art computer and algorithm", factoring a 100 decimal digit number, in the worst case, requires about one month of computer time, for a 200 digit number more than 4 million years would be required, for a 300 digit number, over 10^{12} years, and for a 500 digit number over 10^{22} years. This appears to be due to the fact that, to the best of our knowledge, factoring seems to require repeated trial divisions and the number of these seems to at best exponential in the integer n.

Now we describe a popular public key encryption system which exploits the huge apparent gap in complexity between factoring and the Euclidean algorithm. The system is known as the RSA system after Rivest, Shamir, and Adleman, who devised it in 1979. The system envisions a receiver R who wishes to receive secret messages from any one S of a number of different senders, and wishes to make public the method by which they should encrypt their messages, but wishes to do this so that (practically) they cannot be decrypted by anyone other than R. As an example take R to be large company headquarters and S to be any one of the company branch offices which are required to periodically report proprietary information to headquarters and wish to do so without fear of decryption by any competitive companies. Alternatively R could

be the Chief of U. S. Naval operations (CNR) and S is any one of the thousands of ships and stations who wish to send secure reports to R. In case R is a large company the messages could be sent by some sort of secure mail or courier, but these are expensive alternatives which would not even be possible in case R is CNR, S is a ship at sea, and the message must be sent by radio, so that its encrypted version is accessible to all. Traditionally, before public key systems, the Navy used encryption systems in which the encoding procedure was as secret as the decryption procedure; so that so long as neither of these procedures was discovered by the "enemy", the encrypted messages could be sent by radio and interception by the enemy was not an issue. The problem was that sometimes naval ships were captured or their coding apparatus otherwise obtained by the enemy. With a public key system this would be no problem since with such a system *anyone* can send a message and only R can decrypt it.

To establish a particular RSA system the encoder R begins by selecting two large prime numbers p and q (usually close to each other, but this is not essential) and forms their product $n = pq$. From what we have said about factoring if the primes are each of about 50 digits so that n has about 100 digits, the resulting encryption system will be good for about one month since it can be shown that breaking the coding system requires about the same computation as factoring n. In any case forming $n = pq$ is easy to do. Second, R selects a number $e > 1$ relatively prime to $(p-1)(q-1)$. This is also easy to do: just select random numbers e less than $(p-1)(q-1)$ and apply the Euclidean algorithm until $(e,(p-1)(q-1)) = 1$ is obtained. The pair (e,n) is called the *encrypting pair* of the system, and the components of the encrypting pair are called the *exponent* and the *modulus*, respectively, of the system. Once they are determined by R, they are published as widely as desired. In applying the Euclidean Algorithm to find e, the encoder determines x and y such that

$$ex + (p-1)(q-1)y = 1.$$

Take d in $\mathbb{Z}_{(p-1)(q-1)}$ to be congruent to x. Therefore d is a multiplicative inverse of e in $\mathbb{Z}_{(p-1)(q-1)}$, i.e.: d has the property that it is a positive integer less than $(p-1)(q-1)$ such that

$$ed \equiv 1 \ \ ((p-1)(q-1)).$$

(See Theorem 2.6.4.) The pair (d,n) is called the *decrypting* pair. R can now discard both p and q if he wishes but the integer d he holds with the highest security.

Now suppose S wishes to send R a message. To do this he must employ a previous publicly agreed upon substitution system whereby the message text becomes a sequence of integers, called the *blocks* of the message,

$$C_1, \ldots, C_k.$$

The requirement is that the integers C_i are all less than n and are relatively prime to n. Since $\phi(pq) = (p-1)(q-1)$ by Theorem 2.5.2, the fraction of all integers less

than pq which are relatively prime to pq is

$$\frac{(p-1)(q-1)}{pq} = 1 - \frac{1}{p} - \frac{1}{q} + \frac{1}{pq}$$

which for large primes is very close to 1. Hence a large fraction of the integers less than n can be blocks of a message. Usually these blocks C_i are determined by first replacing the alphabetic-numeric characters of the message, including punctuation and blanks, by their 2 decimal digit ASCII equivalents. The resulting long string of decimal digits is then broken into blocks of decimal numbers C_i, each less than n. In the off chance that the blocks of the message are not prime to the modulus, S can attempt a change in the message (e.g.: in the punctuation), or he can notify R who can establish a new encrypting pair.

Next S encrypts his message by first computing C_i^e for each i and then replacing each of these integers by their residues R_i modulo n. Thus the encrypted version of the message $C_1, \ldots, C_k$, the one actually sent to R (and which is available for the public at large), is the sequence

$$R_1, \ldots, R_k \quad \text{where} \quad R_i \equiv C_i^e \ \ (n).$$

Computing the powers C_i^e is also not very difficult. This is due to the fact that by successive squaring we can compute $C_i^2, C_i^4, C_i^8, \ldots$, always reducing modulo n as we go.

Now for the decrypting process. Having received the encrypted message $R_1, \ldots, R_k$, R decrypts it by simply computing the sequence

$$R_1^d, \ldots, R_k^d, \ \text{ reducing modulo } n.$$

(As in the encryption process, the computation of these powers modulo n is not hard.) To justify this we need to verify that $R_i^d \equiv C_i \ \ (n)$ for all i. To do this, first note that since $de \equiv 1 \ \ ((p-1)(q-1))$, it follows that

$$de = 1 + r(p-1)(q-1) \quad \text{for some integer } r.$$

Also, by Exercise 2 b) of Section 2.7 we have

$$C_i^{(p-1(q-1)} \equiv 1 \ \ (n) \quad \text{for each } i = 1, \ldots, k.$$

(This is where we need $(C_i, n) = 1$.) Consequently we have, for each i, starting from $R_i \equiv C_i^e \ \ (n)$,

$$\begin{aligned} R_i^d &\equiv C_i^{de} \ \ (n) \\ &\equiv C_i^{1+r(p-1)(q-1)} \ \ (n) \\ &\equiv C_i (C_i^{(p-1)(q-1)})^r \ \ (n) \\ &\equiv C_i \cdot 1^r \ \ (n) \\ &\equiv C_i \ \ (n). \end{aligned}$$

This justifies the procedure. Of course R must now concatenate the blocks C_i and reverse the substitution process to obtain the original message.

In the encryption process, the function which carries C_i into C_i^e is an example of what is usually called a *trapdoor* function: it is easy to compute it but very difficult to compute its inverse (without special knowledge, in this case the integer d).

Let us consider a very simple example of the encryption process. We take the primes p and q to be 89 and 91 respectively and hence $n = 8099 = 89 \cdot 91$. Of course in practice we would choose much larger primes. We also choose $e = 7$. This is the smallest integer larger than 1 and relatively prime to $(89-1)(91-1) = 7920$. Based on the modulus 8099 we decide that the blocks will be four digit numbers less than 8099. Suppose our message is to be "HELLO". Using the two digit ASCII codes H=72, E=69, L=76, O=79, blank=32, we substitute to obtain the string 726976767932 (placing a blank at the end) and break this into blocks

$$C_1 = 7269, \quad C_2 = 7676, \quad C_3 = 7932.$$

Even with such small primes we see that $1 - 1/89 - 1/91 + 1/89 \cdot 91$ is about 0.98, i.e.: about 98% of the integers less than 8099 are relatively prime to 8099; in particular it is easily checked that the C_i just determined are prime to 8099. Finally we raise each of these integers to the 7-th power mod 8099 by using four multiplications to compute $C_i \cdot C_i^2 \cdot (C_i^2)^2$, reducing mod 8099 at each step, to obtain the R_i.

Finally, it is important to understand that the utility of the RSA system depends upon the "perceived" complexity in breaking the code. More precisely, it has been proven that the complexity of breaking this coding system is essentially the same as the complexity of factoring a number into primes. But also it is not known that there is no polynomial time algorithm for either factoring or for breaking the RSA system. Thus it is could turn out that, in time, faster algorithms may be developed which will render the RSA system ineffective.

Exercises Section 2.8

1. Prove that if $a^r \equiv 1\ (n)$ and $a^s \equiv 1\ (n)$, then $a^{(r,s)} \equiv 1\ (n)$.

2. Suppose the nonnegative integers m and n are represented in unary, i.e.: m is represented by m strokes: $||\cdots|$. Prove that in this representation the complexity of the Euclidean algorithm $c(E)$ for integers $m < n$ is $c(E)(n) = O(n)$.

3. In describing the RSA system we suggested that an efficient way to compute a power a^n of an integer a was to compute successively a^2, $a^{2^2} = (a^2)^2$, $a^{2^3} = (a^{2^2})^2, \ldots$ by successively squaring. Use this approach to show that a^n can be computed by $O(\log_2(n))$ multiplications.

4. In the example above of the RSA system show that $d = 2263$. How many multiplications are required to compute each of the $R_i^d \bmod (8099)$ in the decryption process?

Chapter 3

The Discrete Calculus

It is frequently observed that the world of mathematics is split into to two rather opposing cultures: the discrete and the continuous. The extent to which this is true is often experienced by students who are familiar with the useful intuition developed in studying calculus and differential equations and who find themselves somewhat at a loss when first confronted with problem solving in the discrete domain. On the other hand, another broad generality about the world of mathematics is the observation that analogy plays an important role in mathematical thinking. Nowhere is this more striking than in the discrete analog of the familiar continuous calculus. In the present chapter we shall selectively develop this discrete calculus, systematically seeking analogies with its familiar continuous counterpart. We hope that analogical presentation will help make the gap between the discrete and continuous become more apparent than real. We shall restrict ourselves to the discrete calculus of functions of a single variable. As in the more familiar continuous calculus we begin with the *calculus of finite differences*, the analog of differential calculus of functions of a single variable.

3.1 The calculus of finite differences

The finite difference calculus is the study of functions —in our case, of a single variable — defined at a discrete set of equally spaced points. For example if the constant space between points is h then

$$f(a), \quad f(a+h), \quad f(a+2h), \ldots, f(a+nh)$$

are $n+1$ such values of the function f. The basic operation of the difference calculus, the analog of differentiation, is the difference operator Δ defined by

$$\Delta f(x) = f(x+h) - f(x).$$

This operator is already familiar from calculus where it occurs in the definition of the derivative. Since h is supposed to be constant it is simpler to not divide this

expression by h. Also, unlike the continuous case we do not take limits and as one might expect this leads to major simplification. What are the basic properties of the derivative which we hope the difference to share? The first, not unexpectedly, is *linearity*:

$$\Delta[af(x) + bg(x)] = a\Delta f(x) + b\Delta g(x).$$

This property is easily verified from the definition and the reader should take the trouble to go through the formal verification. The consequences of linearity are many, not the least of which is that we can compute the differences of complicated expressions term by term. For example, from $\Delta x^2 = (x+h)^2 - x^2 = 2hx + h^2$, $\Delta x = (x+h) - x = h$, and $\Delta c = 0$, c a constant, we have

$$\Delta(ax^2 + bx + c) = 2ahx + ah^2 + bh.$$

The difference of a product is given by

$$\begin{aligned}
\Delta[f(x)g(x)] &= f(x+h)g(x+h) - f(x)g(x) \\
&= f(x+h)g(x+h) - f(x+h)g(x) + f(x+h)g(x) - f(x)g(x) \\
&= f(x+h)\Delta g(x) + g(x)\Delta f(x).
\end{aligned}$$

Notice that if we had added and subtracted $f(x)g(x+h)$ instead of $f(x+h)g(x)$, we would have obtained, equivalently,

$$\Delta[f(x)g(x)] = f(x)\Delta g(x) + g(x+h)\Delta f(x).$$

This little detail is one place where the difference calculus is *not* simpler than its continuous counterpart. For the difference of a quotient we have

$$\begin{aligned}
\Delta\left(\frac{f(x)}{g(x)}\right) &= \frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)} \\
&= \frac{g(x)f(x+h) - f(x)g(x+h)}{g(x+h)g(x)} \\
&= \frac{g(x)f(x+h) + g(x)f(x) - f(x)g(x+h) - g(x)f(x)}{g(x+h)g(x)} \\
&= \frac{g(x)\Delta f(x) - f(x)\Delta g(x)}{g(x+h)g(x)}.
\end{aligned}$$

The similarity between these and the corresponding formulas for the derivative is typical. Note, however, the presence of the $x + h$ terms.

The most important fact about the differential calculus is the Mean Value Theorem, which asserts that if a function f is differentiable on an interval $[a, b]$, then for $a \le x \le b$, $f(x) = f(a) + (x - a)f'(\xi)$ for some ξ between a and x. The importance of this theorem is largely because of its important consequence: if a function has derivative zero on an interval then the function is a constant on the interval. This

in turn implies the so called *law of parallel graphs:* if two functions have the same derivative they differ by an additive constant. For the difference calculus, while we obviously do not have an analog of the Mean Value Theorem, we do have the fact that if $\Delta f(x) = 0$ for the discrete values of x where the function is defined, then $f(x)$ is constant, trivially, by the definition of the difference. Likewise if $\Delta f(x) = \Delta g(x)$ then linearity gives $f(x) = g(x) + c$, analogous to the differential calculus case. The law of parallel graphs is important for evaluating definite integrals in the continuous calculus. Its discrete analog will have similar importance in the discrete analog of integral calculus — the *summation* calculus — which we develop in the next section.

The number e is important in the continuous differential calculus, in large part because it is (up to a constant factor) the unique function which is equal to its own derivative. This role is paralleled by the number 2 in the difference calculus, since

$$\Delta 2^x = 2^{x+h} - 2^x = 2^x(2^h - 1) = 2^x \text{ if } h = 1.$$

Partly for this reason we shall shortly specialize to the case $h = 1$.

Higher order differences are defined by

$$\Delta^2 f(x) = \Delta(\Delta(f(x)))$$

and generally by

$$\Delta^n f(x) = \Delta(\Delta^{n-1}(f(x))), \quad n > 1.$$

For example

$$\Delta^2(ax^2 + bx + c) = 2ah^2,$$
$$\Delta^3(ax^2 + bx + c) = 0.$$

These computations illustrate the following fundamental theorem:

Theorem 3.1.1 *If $f(x) = a_n x^n + \cdots + a_1 x + a_0$, $(a_n \neq 0)$ is a polynomial of degree $n \geq 0$, then $\Delta^n(f(x)) = a_n h^n n!$ and $\Delta^{n+1}(f(x)) = 0$.*

Proof By the binomial theorem we have

$$\begin{aligned} \Delta(x^n) &= (x+h)^n - x^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^{n-k} h^k - x^n \\ &= nhx^{n-1} + \left(\begin{array}{c} n \\ 2 \end{array} \right) h^2 x^{n-2} + \cdots + h^n. \end{aligned}$$

Hence when Δ is applied to $f(x)$ the term $a_n x^n$ becomes a polynomial of degree $n-1$ with leading coefficient $a_n nh$. Using the linearity property it is clear that the degree of each term is decreased by 1 and the term $a_n nhx^{n-1}$ cannot cancel. Applying Δ repeatedly $n-1$ more times gives $\Delta^n f(x) = a_n n! h^n$, and once more gives $\Delta^{n+1} f(x) = 0$.

Notice that if $h = 1$ then the formula of the preceding theorem for the n-th difference of a polynomial of degree n becomes simply $\Delta^n f(x) = a_n n!$. The absence of the factor h^n will considerably simplify many results. Also notice that since we are dealing with functions defined on equally spaced points, we can always rescale the spacing between points to be 1 without loss of generality. For this reason we will henceforth specialize to $h = 1$.

The significance of Theorem 3.1.1 can be appreciated by introducing the *difference table*, in which higher differences are laid out as indicated below:

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\cdots$
0	$f(0)$				
		$\Delta f(0)$			
1	$f(1)$		$\Delta^2 f(0)$		
		$\Delta f(1)$		$\Delta^3 f(0)$	
2	$f(2)$		$\Delta^2 f(1)$		$\cdot$
		$\Delta f(2)$		$\Delta^3 f(1)$	$\cdot$
3	$f(3)$		$\Delta^2 f(2)$		$\cdot$
		$\Delta f(3)$		$\Delta^3 f(2)$	
4	$f(4)$		$\Delta^2 f(3)$		
		$\Delta f(4)$			
5	$f(5)$				
$\cdot$	$\cdot$				
$\cdot$	$\cdot$				

From Theorem 3.1.1 we know that if $f(x)$ is a polynomial of degree n, then the n-th differences will be constant. Going backwards, if we know the top elements of each column of the difference table for some polynomial then we can reproduce the entire table, and specifically the column of function values, using only addition. For example, $\Delta^2 f(1) = \Delta^2 f(0) + \Delta^3 f(0)$ in the table above.

Another important observation is to consider the difference table determined by a *smooth* function (intuitively a continuous function with many bounded derivatives) like $\sin x$ which is approximated well over short intervals by a polynomial of low degree, say a cubic. This means that if we look at the column of third differences of $\sin x$ we see that the entries change slowly. Indeed any abrupt change would indicate an error in the original table of values. Historically, before computing machines displaced human table makers, this observation provided the practical means of checking for computation errors in tables of functions.

Just as in continuous calculus it is useful to think of differentiation as an *operator* on functions the same is true of the difference operator Δ. To increase our flexibility in applying Δ, we define the 0-th difference $\Delta^0 f(x) = f(x)$ for all $f(x)$. If we define the *identity* operator 1 by $1f(x) = f(x)$ for all $f(x)$, the we have the relation

$$\Delta^0 = 1$$

between these operators. We further define the *shift* operator E by

$$Ef(x) = f(x+1).$$

Then it is easy to see that both E and 1 are linear and the three operator satisfy the equation

$$\Delta = E - 1,$$

for $\Delta f(x) = f(x+1) - f(x) = Ef(x) - 1f(x) = (E-1)f(x)$. We also define

$$E^n f(x) = EE^{n-1}f(x), \;\; n \geq 1, \;\; E^0 f(x) = f(x)$$

i.e.:

$$E^0 = 1.$$

A principle advantage of this symbolism is that we can readily verify that

$$\Delta^2 = (E-1)^2 = (E-1)(E-1) = E^2 - 2E + 1,$$

and in general,

$$\Delta^n = (E-1)^n = \sum_{k=0}^{n} (-1)^{n-k} \left(\begin{array}{c} n \\ k \end{array} \right) E^k.$$

Applying this formula to $f(x)$ we obtain the following famous conclusion:

Theorem 3.1.2 *(Lagrange's formula for the n-th difference)*

$$\Delta^n f(x) = \sum_{k=0}^{n} (-1)^{n-k} \left(\begin{array}{c} n \\ k \end{array} \right) f(x+k).$$

For example, taking $n = 4$ we have

$$\Delta^4 f(x) = \left(\begin{array}{c} 4 \\ 0 \end{array} \right) f(x) - \left(\begin{array}{c} 4 \\ 1 \end{array} \right) f(x+1) + \\ \left(\begin{array}{c} 4 \\ 2 \end{array} \right) f(x+2) - \left(\begin{array}{c} 4 \\ 3 \end{array} \right) f(x+3) + \left(\begin{array}{c} 4 \\ 4 \end{array} \right) f(x+4).$$

As an application, if $f(x)$ is a cubic polynomial, so that $\Delta^4 f(x) = 0$, and say $f(x+2)$ is missing from a table of values, we can compute it from the formula above using the four nearby values.

In differential calculus the formula

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

for the derivative of x^n plays a special role. In the difference calculus with $h = 1$ the corresponding role is played by the *factorial polynomials*

$$\begin{aligned} x^{(n)} &= x(x-1)(x-2)\cdots(x-n+1),\ n \geq 1, \quad (n \text{ factors}). \\ x^{(0)} &= 1, \end{aligned}$$

(What would be the appropriate definition for $h \neq 1$?) We justify this by the following computation for $n \geq 1$:

$$\begin{aligned} \Delta x^{(n)} &= (x+1)^{(n)} - x^{(n)} \\ &= [(x+1)(x)(x-1)\cdots(x-(n-2))] - [x(x-1)\cdots(x-n+1)] \\ &= x(x-1)\cdots(x-(n-2))[(x+1)-(x-n+1)] \\ &= nx^{(n-1)}. \end{aligned}$$

Note that the definition $x^{(0)} = 1$ implies $0^{(0)} = 1$ which corresponds to $0! = 1$. Also notice the important general relation

$$\frac{x^{(n)}}{n!} = \left(\begin{array}{c} x \\ n \end{array} \right)$$

connecting the factorial polynomials and the binomial coefficients.

The special role played by the functions $x^n;\ n = 0, 1, 2, \ldots$ in differential calculus is primarily due to Taylor's Theorem which tells us how to approximate functions with enough derivatives near $x = 0$ by polynomials, i.e.: by linear combinations of $x^n,\ n = 0, 1, 2, \ldots$. Thus the functions $\{x^0 = 1, x, x^2, \ldots, x^n\}$ are a basis for the vector space $\mathcal{P}_n$ of polynomials of degree n or less (including the polynomial 0). But it is easy to verify that the factorial polynomials $\{x^{(0)} = 1, x^{(1)} = x, x^{(2)} = x(x-1), \ldots\ x^{(n)}\}$ also form a basis for $\mathcal{P}_n$, so that for any polynomial $f(x)$ of degree n, there are unique sets of numbers $\{a_k\}$ and $\{b_k\}$ such that

$$f(x) = \sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n} b_k x^{(k)}.$$

By taking successive derivatives at $x = 0$ to compute the a_k the left equality above yields Taylor's theorem for polynomials:

$$f(x) = \sum_{k} \frac{f^{(k)}(0)}{k!} x^k.$$

This process is familiar from elementary continuous calculus. To obtain the corresponding formula for the difference calculus we proceed in analogous fashion: first observe that setting $x = 0$ in the formula on the right yields $f(0) = b_0$. Next compute the difference of the formula on the right,

$$\Delta f(x) = \sum_{k} b_k k x^{(k-1)};$$

taking $x = 0$ then yields $\Delta f(0) = b_1$. Taking the second difference and the setting $x = 0$ yields $\Delta^2 f(0) = 2b_2$. Continuing in this way we obtain for each k, $\Delta^k f(0) = k!b_k$ so that we have another famous formula:

Theorem 3.1.3 *(Newton's interpolation formula) If $f(x)$ is a polynomial of degree n,*

$$f(x) = \sum_{k=0}^{n} \frac{\Delta^k f(0)}{k!} x^{(k)} = \sum_{k=0}^{n} \binom{x}{k} \Delta^k f(0).$$

Notice that the factors $\Delta^k f(0)$ appearing in the second expression are just the right-to-left diagonal entries from the difference table (obtained from the tops of the columns).

Now let us return to the relation between the two set of polynomials $\{x^n : n = 0, 1, 2, \ldots\}$ and $\{x^{(n)} : n = 0, 1, 2, \ldots\}$ as different bases for the vector space $\mathcal{P}_n$. Since both sets are bases of $\mathcal{P}_n$, each x^k is a unique linear combination of the $x^{(k)}$ and vise-versa, that is, for each integer n we have

$$x^{(n)} = \sum_{k=0}^{n} s(n,k)x^k \text{ and } x^n = \sum_{k=0}^{n} S(n,k)x^{(k)}$$

for suitable constants $s(n,k)$ and $S(n,k)$. These numbers are called the *Stirling numbers* of the first and second kind, respectively. The $S(n,k)$ (the second kind) were introduced in Chapter 1, Section 7 where they counted the number of partitions of index k of an n-set. Now we see them in a totally different context. Since they are now used to express each of two vector space bases in terms of the other it is clear why there are two kinds of Stirling numbers. To show that the $S(n,k)$ are the same Stirling numbers as before we first observe that they satisfy the same boundary conditions as in Chapter 1, namely

$$S(n,n) = 1 \text{ for } n = 0, 1, 2, \ldots, \text{ and } S(n,0) = 0 \text{ for } n > 0.$$

This is easy to see by equating coefficients in the equation above expressing x^n as a linear combination of the $x^{(k)}$. To complete the proof that the $S(n,k)$ are the same Stirling numbers as before it suffices to show that they satisfy the same recursive formula as before (Theorem 1.7.2), namely

$$S(n,k) = S(n-1,k-1) + k \cdot S(n-1,k),\ 1 \le k \le n.$$

To do this we start by observing that $x^n = x \cdot x^{n-1}$ implies that

$$\begin{aligned} \sum_{k=0}^{n} S(n,k)x^{(k)} &= x \sum_{k=0}^{n-1} S(n-1,k)x^{(k)} \\ &= \sum_{k=0}^{n-1} S(n-1,k)(x-k)x^{(k)} + \sum_{k=0}^{n-1} kS(n-1,k)x^{(k)} \end{aligned}$$

$$\begin{aligned} &= \sum_{k=0}^{n-1} S(n-1,k)x^{(k+1)} + \sum_{k=0}^{n-1} kS(n-1,k)x^{(k)} \\ &= \sum_{k'=1}^{n} S(n-1,k'-1)x^{(k')} + \sum_{k=1}^{n-1} kS(n-1,k)x^{(k)} \quad \text{where } k' = k+1 \\ &= \sum_{k=1}^{n-1} [S(n-1,k-1) + kS(n-1,k)]x^{(k)} + S(n-1,n-1). \end{aligned}$$

Equating coefficients of $x^{(k)}$ we obtain the desired formula.

A similar derivation yields the following recursive relation for the Stirling numbers of the first kind:

$$s(n,k) = -(n-1)s(n-1,k) + s(n-1,k-1) \quad \text{for } 1 \leq k \leq n.$$

Not surprisingly there is a corresponding combinatorial interpretation, in terms of partitions, of the Stirling numbers of the first kind: $s(n,k)$ counts the number of ways of partitioning n objects into k disjoint *circular* permutations. (Consult most any text on combinatorics for more details.)

Before leaving our brief description of the difference calculus we extend the factorial polynomials to negative exponents. The resulting functions will prove to be useful in the next section. At first we might be tempted to define $x^{(-n)}$ for positive n by setting

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

The trouble is that with this definition it is easy to verify that we do *not* have

$$\Delta x^{(-n)} = -nx^{(-n-1)}$$

which is what we would like. To see how to proceed observe that for $0 \leq m \leq n$,

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

which can be easily verified by writing out the definitions of $x^{(m)}$ and $(x-m)^{(n-m)}$. If we formally set $n = 0$ in this equation, we get

$$x^{(0)} = 1 = x^{(m)}(x-m)^{(-m)},$$

that is,

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

Replacing $x - m$ by y we obtain

$$y^{(-m)} = \frac{1}{(y+m)^{(m)}} = \frac{1}{(y+m)(y+m-1)\cdots(y+1)}$$

which we take as the definition of $y^{(-m)}$ for $m \geq 0$. Notice that for $m \neq 0$ we do *not* have

$$x^{(m)}x^{(-m)} = x^{(0)} = 1.$$

On the other hand we *do* have the important formula for the difference:

$$\begin{aligned}
\Delta y^{(-m)} &= \frac{1}{(y+m+1)(y+m)\cdots(y+2)} - \frac{1}{(y+m)\cdots(y+2)(y+1)} \\
&= \left[\frac{1}{y+m+1} - \frac{1}{y+1}\right] \frac{1}{(y+m)\cdots(y+2)} \\
&= -m\frac{1}{(y+m+1)(y+m)\cdots(y+1)} \\
&= -my^{(-m-1)}.
\end{aligned}$$

Exercises Section 3.1

1. Verify each of the following formulas:
   a) $\Delta \sin(ax+b) = 2\sin\frac{ah}{2}\cos[a(x+\frac{h}{2})+b]$.
   b) $\Delta a^x = a^x(a^h - 1)$.
   c) $\Delta \log x = \log(1+\frac{h}{x})$.

2. Compute a table of cubes of integers from 10 to 20, using a difference table.

3. Using Newton's interpolation formula, find a polynomial $f(x)$ where for $n = 1, 2, 3, 4, 5, 6$, $f(n) =$ the n-th prime. (The first prime is 2.)

4. Verify the following:
   a) $3^{(-3)} = \frac{1}{120}$
   b) $1^{(-m)} = \frac{1}{(m+1)!}$.

5. Explain why the following formula is true:

$$\sum_{k=0}^{n} S(n,k)s(k,m) = \delta(m,n),$$

where $\delta(m,n)$, the *Kronecker delta function*, is defined as usual by

$$\delta(m,n) = \begin{cases} 1 & \text{if } m = n, \\ 0 & \text{if } m \neq n. \end{cases}$$

6. In differential calculus the function $\log(x)$ has derivative $1/x$ and thus plays a special role since the formula $d/dx(x^n) = nx^{n-1}$, $(n \in \mathbb{Z})$ never produces a result of degree -1. Show that for the difference calculus the so called *digamma* function

$$F(x) = \sum_{k=1}^{x} \frac{1}{k}$$

plays a role analogous to $\log(x)$. (For reasons unrelated to our present pursuits the digamma function is usually defined by

$$F(x) = \sum_{k=1}^{x} \frac{1}{k} - \gamma$$

where γ is an irrational number called Euler's constant and which is approximately $0.5772156\cdots$. Appropriately interpreting the summation, this means that $F(0) = -\gamma$).

7. This is a slightly open ended question. Since the discrete analog of the continuous function e^x is 2^x as noted earlier, why does it not follow (by the principle of analogy!) that the proper analog of $\log(x) = \log_e(x)$ is not $\log_2(x)$ instead of the digamma function introduced in the last exercise? Hint: Think about how the functions $\log(x)$ and e^x are related and how their derivatives are consequently related. Why does this relationship not have an analog in the discrete calculus?

3.2 The summation calculus

The (finite) summation calculus, which we discuss in this section, stands in the same relation to the finite difference calculus as the familiar (continuous) integral calculus stands to the (continuous) differential calculus. The immediate practical importance of this topic is the remarkable fact that we shall be able to employ analogs of a number of the various strategies for evaluating integrals (by finding antiderivatives) to the solve the problem of expressing finite sums in "closed form", that is, without using the $\sum$ summation symbol. As before we shall restrict ourselves to the case $h = 1$.

Recall that the *Fundamental Theorem of integral calculus* asserts that

$$\int_a^b df(x) = f(b) - f(a).$$

From the definition

$$\Delta f(x) = f(x+1) - f(x)$$

we calculate immediately that

$$\begin{aligned} \sum_{x=a}^{b-1} \Delta f(x) &= f(b) - f(b-1) + f(b-1) - f(b-2) + \cdots + f(a+1) - f(a) \\ &= f(b) - f(a), \end{aligned}$$

and hence this formula is the analog of the *Fundamental Theorem of integral calculus*. It is very important to notice that to obtain the right hand side $f(b) - f(a)$ we must

take the upper index of the sum on the left to be $b-1$ and not b as in the case of the integral. Forgetting this fact is a common source of errors in computing closed form formulas for finite sums.

Skill in evaluating elementary definite integrals $\int_a^b g(x)\,dx$ depends upon being able to recognize the integral as a derivative, i.e.: finding a function f such that

$$df(x) = g(x)\,dx.$$

Likewise, evaluating sums of the form $\sum_{x=a}^{b-1} g(x)$ by the Fundamental Theorem

$$\sum_{x=a}^{b-1} \Delta f(x) = f(b) - f(a)$$

depends upon finding a function f such that

$$\Delta f(x) = g(x).$$

A first and most important example is to exploit the formula

$$\Delta x^{(n)} = n x^{(n-1)}$$

for factorial polynomials. From this we have

$$\sum_{x=a}^{b-1} x^{(n)} = \sum_{x=a}^{b-1} \frac{\Delta x^{(n+1)}}{n+1} = \frac{b^{(n+1)} - a^{(n+1)}}{n+1}, \quad n \neq -1.$$

Hence if we wish to obtain a closed formula for $\sum_{x=a}^{b-1} x^n$, the sum of the n-th powers of the successive integers x from $x = a$ to $x = b-1$, we need to express x^n in terms of factorial polynomials. For example, it is easy to see by direct substitution that $x^2 = x^{(2)} + x^{(1)}$. From this and the last formula we have

$$\begin{aligned}
\sum_{k=1}^{n} k^2 &= \sum_{k=1}^{n} k^{(2)} + k^{(1)} = \sum_{k=1}^{n} k^{(2)} + \sum_{k=1}^{n} k^{(1)} \\
&= \frac{(n+1)^{(3)} - 1^{(3)}}{3} + \frac{(n+1)^{(2)} - 1^{(2)}}{2} \\
&= \frac{(n+1)^{(3)}}{3} + \frac{(n+1)^{(2)}}{2} = \frac{(n+1)(n)(n-1)}{3} + \frac{(n+1)(n)}{2} \\
&= \frac{1}{6}(n+1)(n)(2n+1).
\end{aligned}$$

As another illustration, notice that

$$\sum_{k=1}^{n} 1 = \sum_{k=1}^{n} k^{(0)} = \frac{(n+1)^{(1)} - 1^{(1)}}{1} = n$$

as should be expected.

At this point it is suggested that the reader use this method to obtain the formula

$$\sum_{k=1}^{n} k = \frac{(n+1)(n)}{2}$$

for the sum of the first n integers. This formula is often derived by the following *ad hoc* method usually attributed to Gauss during his precocious childhood. Write the desired sum in two ways as

$$\sum_{k=1}^{n} k = 1 + 2 + \cdots + (n-1) + n$$

and

$$\sum_{k=1}^{n} k = n + (n-1) + \cdots + 2 + 1;$$

then add columnwise to obtain

$$2\sum_{k=1}^{n} k = (n+1) + (n+1) + \cdots + (n+1) = n(n+1).$$

However this ingenious method does not appear to generalize in any easy way to higher powers. Hence the power and advantage of the summation calculus.

To express x^n in terms of factorial polynomials the formula

$$x^n = \sum_{k=1}^{n} S(n,k)x^{(k)}$$

involving the Stirling numbers (of the second kind), $S(n,k)$, can be used provided we have the $S(n,k)$ available. (A table can be computed recursively from the recursion formula for the Stirling numbers, Theorem 1.7.2.) Another approach, usually simpler, is based on the process of synthetic division. Let us review this process now.

If we divide a polynomial $p(x)$ of degree $n \geq 1$ by $x - a$ we obtain

$$p(x) = (x-a)q(x) + r$$

where the quotient $q(x)$ has degree $n-1$ and the remainder $r = p(a)$ is a constant which is zero just in case a is a root of the equation $p(x) = 0$.

Synthetic division simplifies the process for finding the coefficients of $q(x)$ and the remainder r by noticing that the powers of x need not be written provided zero coefficients are supplied for the missing powers. Also we need not write $x - a$, but only a. Thus the form for carrying out the division becomes, for say $x - 2$ and

$p(x) = x^4 + 6x^2 - 7x + 8$,

$$\begin{array}{rcccccccccccc}
 & - & -- & - & -- & - & -- & - & -- & - & -- & & \\
2 & | & 1 & & 0 & & 6 & & -7 & & 8 & & \\
 & & & & 2 & & 4 & & 20 & & 26 & & \\
 & - & -- & - & -- & - & -- & - & -- & - & -- & & \\
 & & 1 & & 2 & & 10 & & 13 & | & 34 & = r = p(2) & \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & & & \\
\text{so } q(x) & = & x^3 & + & 2x^2 & + & 10x & + & 13 & & & &
\end{array}$$

and the remainder $r = p(2) = 34$.

To write a given polynomial $p(x)$ as a sum of terms in $x^{(n)}$ we need to divide, in turn, the polynomial by x, the quotient by $x - 1$, the next quotient by $x - 2$, etc., for n steps. For example if $p(x) = x^4 + 6x^2 - 7x + 8$ we want to express $p(x)$ in the nested form

$$p(x) = (\{[a_4(x-3) + a_3](x-2) + a_2\}(x-1) + a_1)x + a_0,$$

for then, multiplying out, we have

$$\begin{aligned}
p(x) &= a_4(x-3)(x-2)(x-1)x + a_3(x-2)(x-1)x + a_2(x-1)x + a_1x + a_0 \\
&= a_4x^{(4)} + a_3x^{(3)} + a_2x^{(2)} + a_1x^{(1)} + a_0x^{(0)}.
\end{aligned}$$

From this it is clear that a_0 is the constant term of $p(x)$ and also can be obtained as the remainder obtained by dividing $p(x)$ by x. a_1 is the remainder obtained by dividing the corresponding quotient by $x - 1$, etc., so that successive divisions by x, $x - 1$, $x - 2$, $x - 3$ yield the nested form of $p(x)$. Since division by x is trivial we can arrange the work for $p(x) = x^4 + 6x^2 - 7x + 8$ as follows:

$$\begin{array}{rcccccccccc}
 & - & -- & - & -- & - & -- & - & -- & - & -- & \\
1 & | & 1 & & 0 & & 6 & & -7 & | & 8 & = a_0 \\
 & & & & 1 & & 1 & & 7 & & & \\
 & - & -- & - & -- & - & -- & - & -- & & & \\
2 & | & 1 & & 1 & & 7 & | & 0 & = a_1 & & \\
 & & & & 2 & & 6 & & & & & \\
 & - & -- & - & -- & - & -- & & & & & \\
3 & | & 1 & & 3 & | & 13 & = a_2 & & & & \\
 & & & & 3 & & & & & & & \\
 & - & -- & - & -- & & & & & & & \\
 & & 1 & | & 6 & = a_3 & & & & & & \\
 & & \downarrow & & & & & & & & & \\
 & & a_4 & & & & & & & & &
\end{array}$$

Therefore $x^4 + 6x^2 - 7x + 8 = 1x^{(4)} + 6x^{(3)} + 13x(2) + 0x^{(1)} + 8x^{(0)}$.

As a final example of this process, to obtain a closed formula for $\sum_{k=0}^{n}(k^3+k)$ we take $p(x) = 1x^3 + 0x^2 + 1x^1 + 0x^0$ and obtain

$$\begin{array}{cccccccccc}
 & - & -- & - & -- & - & -- & - & -- \\
1 & | & 1 & & 0 & & 1 & | & 0 \\
 & & & & 1 & & 1 & & \\
 & - & -- & - & -- & - & -- & & \\
2 & | & 1 & & 1 & | & 2 & & \\
 & & & & 2 & & & & \\
 & - & -- & - & -- & & & & \\
 & & 1 & | & 3 & & & &
\end{array}$$

so $x^3 + x = x^{(3)} + 3x^{(2)} + 2x^{(1)} + 0x^{(0)}$; therefore

$$\begin{aligned}
\sum_{k=0}^{n}(k^3+k) &= \sum_{k=0}^{n}(k^{(3)} + 3k^{(2)} + 2k^{(1)}) \\
&= \frac{(n+1)^{(4)} - 0^{(4)}}{4} + 3\frac{(n+1)(3) - 0^{(3)}}{3} + 2\frac{(n+1)^{(2)} - 0^{(2)}}{2} \\
&= \frac{1}{4}(n+1)^{(4)} + (n+1)^{(3)} + (n+1)^{(2)} \\
&= \frac{1}{4}(n+1)(n)[(n-1)(n-2) + 4(n-1) + 4].
\end{aligned}$$

The summation formula

$$\sum_{x=a}^{b-1} x^{(n)} = \frac{b^{(n+1)} - a^{(n+1)}}{n+1}$$

is of course also valid for negative exponents $n \neq -1$. We merely need to recall that our definition

$$y^{(-m)} = \frac{1}{(y+m)^{(m)}} = \frac{1}{(y+m)(y+m-1)\cdots(y+1)} \quad (m > 0)$$

implies that

$$\Delta y^{(-m)} = -m y^{(-m-1)}.$$

For example, from the difference formula

$$(x-1)^{(-3)} = \frac{1}{(x-1+3)^{(3)}} = \frac{1}{(x+2)(x+1)(x)},$$

we obtain the summation formula

$$\begin{aligned}
\sum_{x=1}^{m} \frac{1}{(x+2)(x+1)x} &= \sum_{x=1}^{m}(x-1)^{(-3)} = \frac{[(m+1)-1]^{(-2)} - 0^{(-2)}}{-2} \\
&= -\frac{1}{2}\left[\frac{1}{(m+2)^{(2)}} - \frac{1}{(0+2)^{(2)}}\right] = -\frac{1}{2}\left[\frac{1}{(m+2)(m+1)} - \frac{1}{2\cdot 1}\right].
\end{aligned}$$

This can be applied to the corresponding infinite series:

$$\begin{aligned}\sum_{x=1}^{\infty}\frac{1}{(x+2)(x+1)x} &= \lim_{m\to\infty}\sum_{x=1}^{m}\frac{1}{(x+2)(x+1)x}\\ &= \lim_{m\to\infty}-\frac{1}{2}\left[\frac{1}{(m+2)(m+1)}-\frac{1}{2\cdot 1}\right]=1/4.\end{aligned}$$

In a similar way both finite and infinite sums of the general form $\sum\frac{1}{(x+k)\cdots(x+1)x}$ can be evaluated.

As in elementary integral calculus experience with difference formulas leads to formulas for finite sums, analogous to a table of integrals. In elementary integral calculus there are really only two other general methods for applying the Fundamental Theorem: change of variable (e.g.: trigonometric substitution) and integration by parts. Since our development of the discrete calculus depends on functions defined at equally spaced arguments — usually $h=1$ — the only changes of variable allowable are translations $x=x'+k$ which are very limited. On the other hand, summation by parts, the analog of integration by parts, is just as powerful in the discrete as in the continuous case. In the continuous case integration by parts is based on the formula

$$d(uv)=udv+vdu.$$

Likewise summation by parts uses the formulas

$$\begin{aligned}\Delta(uv) &= u\Delta v+v(x+1)\Delta u\\ &= u(x+1)\Delta v+v\Delta u\end{aligned}$$

established in Section3.1. The first of these directly implies

$$\sum_{x=0}^{m-1}u(x)\Delta v(x)=u(m)v(m)-u(0)v(0)-\sum_{x=0}^{m-1}v(x+1)\Delta u(x).$$

Exercises Section 3.2

1. Find formulas for the sum of the first n odd integers, and for the sum of the first n even integers.

2. Find a closed formula for $\sum_{k=1}^{n}(2k-1)^2$.

3. Use the summation calculus to obtain the formula $\sum_{n=0}^{m}a^n=\frac{a^{m+1}-1}{a-1}$ for the sum of a geometric series.

4. Find a formula for the sum of the cubes of the first n integers.

5. Establish the formula

$$\sum_{n=0}^{m-1} na^n = \frac{a}{(1-a)^2}[(m-1)a^m - ma^{m-1} + 1].$$

6. Show that $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$.

3.3 Difference Equations

The principle of analogy having been well established, the next topic in the development of the discrete calculus is naturally finite difference equations. As is customary in the elementary treatment of the continuous case, we restrict our treatment to *linear* difference equations.

First we need to point out that at least in notation, our principle of analogy can be somewhat misleading. For example we might expect that the discrete analog of the simple linear differential equation

$$y' + 2y = x$$

is the difference equation

$$\Delta y + 2y = x.$$

But if we substitute

$$\Delta y = y(x+1) - y(x)$$

the difference equation becomes

$$y(x+1) + y(x) = x$$

and in this form it turns out that the equation is more easily understood and solved. In fact in this form the equation says simply that the unknown function y is to be determined by the property that if we add together the function values for two successive arguments we obtain the first argument — this is not so clear from the form $\Delta y + 2y = x$. For a more extreme example consider the equation

$$\Delta^2 y + 2\Delta y + y = f(x).$$

We might think that this is a second order difference equation, but after substituting the meanings of Δ^2 and Δ we have

$$y(x+2) - 2y(x+1) + y(x) + 2[y(x+1) - y(x)] + y(x) = f(x)$$

which collapses into

$$y(x+2) = f(x)$$

which is completely trivial. What is important in a linear difference equation is the maximum difference in the arguments occurring in the unknown function. Hence the example $y(x+1) - y(x) = x$ is a first order equation while

$$y(x+2) + y(x-1) = 0$$

is third order. Thus the general second order linear difference equation can be expressed as

$$y(x+2) + a(x)y(x+1) + b(x)y(x) = f(x)$$

where $a(x), b(x)$ and $f(x)$ are given functions of integers x. In accepting this as the general form we are assuming, as is customary in the case of differential equations, that the coefficient of $y(x+2)$ can be taken to be 1, which is always possible if we can divide through by the given coefficient function, i.e.: if it is never zero in the range of values of x under consideration. From here on we will consider only second order linear difference equations; the theory extends naturally to the n-th order case and except for being much simpler, is completely analogous to the theory of linear differential equations. Further we shall develop solution techniques for the constant coefficient case, again by an almost perfect analogy with differential equations.

As in the study of differential equations, for a function y we write the left side of the equation as

$$L(y) = y(x+2) + a(x)y(x+1) + b(x)y(x)$$

and observe that L is a linear operator on the (real) vector space of functions defined on the integers, i.e.: for real numbers c and d and functions y_1 and y_2,

$$L(cy_1 + dy_2) = cL(y_1) + dL(y_2).$$

In this notation we can write our general difference equation as

$$L(y) = f(x).$$

This is usually called the complete equation. The equation

$$L(y) = 0$$

is the corresponding homogeneous equation. From the linearity of L it follows immediately that the solutions of the homogeneous equation form a vector space — a subspace of the space of all real valued functions defined on the integers. Further, if y_1 and y_2 are any pair of solutions of the complete equation then their difference is a solution of the homogeneous equation; this has the important consequence that if y_p is *any* (particular) solution of the complete equation, then the *general* solution is $y_p + g$ where g is the general solution of the homogeneous equation. Thus we need to solve two problems:

a) find the general solution of the homogeneous equation, and

b) find a particular solution of the complete equation.

Regarding problem a) we first have the following simple fact:

Theorem 3.3.1 *(Existence and uniqueness theorem) For an arbitrary pair of real numbers c_0 and c_1, there exists exactly one solution $y(x)$ of the equation*

$$L(x) = 0$$

with the property that $y(0) = c_0$ and $y(1) = c_1$.

The proof of the theorem is immediate if we write the equation $L(x) = 0$ as

$$y(x+2) = -a(x)y(x+1) - b(x)y(x),$$

for from this we can compute $y(2)$ uniquely as $y(2) = -a(0)y(1) - b(0)y(0) = -a(0)c_1 - b(0)c_0$. having done this we can then compute $y(3)$ uniquely from $y(2)$ and $y(1)$, and so on for the succeeding values of x. (For simplicity we have chosen the initial values to be $y(0)$ and $y(1)$; we could take any other pair of successive integers, positive or negative, if we wished, so long as we are in a range where the coefficients $a(x), b(x)$ are defined.)

We should mention at this point that the corresponding existence and uniqueness theorem for differential equations is far less trivial to prove. In any case, however, the important consequence is the same:

Theorem 3.3.2 *The solutions of the second order homogeneous linear difference equation*

$$L(y) = 0$$

form a two dimensional subspace of the vector space of all real valued functions on the integers.

We imitate the proof for differential equations. First we show that the *uniqueness* statement of the preceding theorem implies that the dimension of the solution space is ≤ 2: suppose y_1, y_2, y_3 are any three solutions of $L(y) = 0$, each satisfying the initial conditions, i.e.:

$$y_i(0) = c_0,\ y_i(1) = c_1 \text{ for } i = 1, 2, 3.$$

Next let (k_1, k_2, k_3) be any non-trivial solution (i.e.: not all zero) of the system of two homogeneous linear algebraic equations

$$\begin{aligned} y_1(0)x_1 + y_2(0)x_2 + y_3(0)x_3 &= 0 \\ y_1(1)x_1 + y_2(1)x_2 + y_3(1)x_3 &= 0 \end{aligned}$$

in the three unknowns x_1, x_2, x_3. Again by linearity the function

$$y(x) = k_1y_1(x) + k_2y_2(x) + k_3y_3(x)$$

is a solution of $L(y) = 0$ and, moreover, has the properties

$$y(0) = 0, \text{ and } y(1) = 0.$$

But the zero function is also a solution of $L(y) = 0$ having these same properties. Thus it follows from the uniqueness part of the Existence and uniqueness theorem that $y(x) = 0$ for all x, that is that

$$k_1 y_1(x) + k_2 y_2(x) + k_3 y_3(x) = 0$$

for all x. Hence the solutions y_1, y_2, y_3 are linearly dependent. Hence the dimension of the solution space is no greater than 2.

To complete the proof we show that the *existence* statement of the preceding theorem implies that the dimension of the solution space is ≥ 2. To do this we invoke existence to obtain solutions y_1, y_2 with the properties

$$\begin{aligned} y_1(0) = 1, &\quad y_1(1) = 0, \\ y_2(0) = 0, &\quad y_2(1) = 1. \end{aligned}$$

Then suppose k_1, k_2 are real numbers such that

$$k_1 y_1(x) + k_2 y_2(x) = 0$$

for all integers x. Taking $x = 0$ we have $k_1 = k_1 y_1(0) + k_2 y_2(0) = 0$ and $k_2 = k_1 y_1(1) + k_2 y_2(1) = 0$. We conclude that y_1 and y_2 are linearly independent, so the dimension of solution space is at least 2. Combining this with the preceding paragraph we see that the dimension is exactly 2.

How do we explicitly find two linearly independent solutions of the homogeneous equation? In general we can't. However in the case of constant coefficients we can. Recall that in the case of differential equations we substitute $y(x) = e^{mx}$ in the differential equation and obtain a polynomial equation, the characteristic equation in m, whose roots are such that $y(x) = e^{mx}$ is a solution. For difference equations we do the same thing except for a convenient change of variable, namely we substitute $y(x) = \rho^x$ in the difference equation

$$y(x+2) + ay(x+1) + by(x) = 0$$

to obtain

$$\rho^{x+2} + a\rho^{x+1} + b\rho^x = 0.$$

Canceling ρ^x we obtain the quadratic *characteristic equation*

$$\rho^2 + a\rho + b = 0$$

whose roots, the *characteristic roots*, are the values for which ρ^x is a solution. In general we expect two roots ρ_1 and ρ_2. If these are distinct then the solutions ρ_1^x and ρ_2^x are linearly independent (Exercise 6 below) so that

$$y(x) = C_1 \rho_1^x + C_2 \rho_2^x$$

is the general solution, where C_1 and C_2 are constants to be determined by the initial conditions. If the roots are equal then it turns out that both ρ^x and $x\rho^x$ are solutions and are linearly independent. Hence in this case the general solution is

$$y(x) = C_1\rho^x + C_2 x\rho^x.$$

We leave the verification of these facts as exercises and illustrate with some examples.

Example 1. In case the equation is first order, we obtain a linear characteristic equation. Consider the equation

$$y(x+1) + by(x) = 0,$$

which we wish to solve for $x = 0, 1, 2, \ldots$ and $y(0) = C$, any constant. Writing the equation as $y(x+1) = -by(x)$ it is clear that

$$y(1) = C(-b), y(2) = C(-b)^2, \ldots, y(n) = C(-b)^n, \ldots$$

is the solution. On the other hand, using the general method and substituting $y(x) = \rho^x$ we have

$$\rho^x(\rho + b) = 0,$$

so $\rho = -b$ and the general solution is $y(x) = C(-b)^x$, $x = 0, 1, 2, \ldots$ as expected.

Example 2. The Fibonacci sequence is the sequence of integers $0, 1, 1, 2, 3, 5, \ldots$ where each term is the sum of the two preceding terms, i.e.: for $n = 1, 2, \ldots$, the terms y_n satisfy the difference equation

$$y_{n+1} - y_n - y_{n-1} = 0.$$

(Changing the notation from $y(x)$ to y_n just reminds us that x, or n, is supposed to be an integer, and conforms to the common subscript notation for sequences.) Substituting $y_n = \rho^n$, we have

$$\rho^{n-1}(\rho^2 - \rho - 1) = 0$$

so $\rho = (1 \pm \sqrt{5})/2$. Thus the general solution is

$$y_n = C_1\left(\frac{1+\sqrt{5}}{2}\right)^n + C_2\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

Using $y_0 = 0$ we obtain

$$0 = C_1 + C_2$$

and from $y_1 = 1$

$$1 = C_1\left(\frac{1+\sqrt{5}}{2}\right) + C_2\left(\frac{1-\sqrt{5}}{2}\right).$$

These equations have the solution $C_1 = -C_2 = 1/\sqrt{5}$, so

$$y_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

Since y_n, the n-th Fibonacci number, is an integer, y_n is actually an integer, even though the above expression contains radicals. Hence, for example, since $y_6 = 8$ we have

$$8 = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^6 - \left(\frac{1-\sqrt{5}}{2}\right)^6\right].$$

This somewhat remarkable formula for the Fibonacci numbers was discovered by de Moivre in the eighteenth century. It does not appear to be a very practical way to compute the n-th Fibonacci number. (The most obvious way appears to be to just compute recursively, using the definition given by the difference equation.) On the other hand, and perhaps most remarkable of all, this formula does have at least one timely practical application. This will be the topic of the next section.

Example 3. Consider the difference equation

$$y_{n+2} - y_n = n^2 + n.$$

The homogeneous equation $y_{n+2} - y_n = 0$ has characteristic polynomial $\rho^2 - 1$ so its general solution is $C_1 + (-1)^n C_2$ where C_1 and C_2 are the constants which can be computed for given initial conditions. To find a particular solution of the complete equation we use undetermined coefficients, i.e.: we substitute a polynomial in n into the difference equation and attempt to solve the linear equations resulting from equating coefficients of like powers of n. If we take the trial polynomial to be

$$y_n = An^2 + Bn + C$$

we find that the resulting linear equations are inconsistent. This is due to the fact that any constant C already solves the homogeneous equation. Hence, just as we do in the case of differential equations, (when the trial solution contains terms which are solutions of the homogeneous equation), we try

$$y_n = An^3 + Bn^2 + Cn$$

instead. Substitution into the original equation then leads to a consistent system of three linear equations in A, B, C. The details of the solution are left as an exercise.

Exercises Section 3.3

1. Find a particular solution of the difference equation of example 3, above.

In exercises 2,3,4 below, find the general solution of the given difference equation.

2. $y_{n+2} - y_{n+1} - 2y_n = n$

3. $y_{n+1} - y_n = 1$

4. $y_{n+2} - 2y_{n+1} + y_n = 0$

5. Find the solution of

$$y_{n+2} - y_n = n^2 + n$$

satisfying the initial conditions $y_0 = 0, y_1 = 1$.

6. a) Show that if $\rho_1 \neq \rho_2$ then ρ_1^x, ρ_2^x are linearly independent.
b) Show that ρ^x and $x\rho^x$ are linearly independent.

3.4 Application: the complexity of the Euclidean algorithm

In Section 2.8 we discussed the RSA public key encryption system. The utility of this encryption system depended on two things:

a) the (apparent) *inefficiency* of factoring large integers into primes, and

b) the *efficiency* of the Euclidean algorithm.

In particular, in Section 2.8 we promised to justify b) in the present section by proving the following important theorem which implies that the Euclidean algorithm is extremely efficient.

Theorem 3.4.1 *If $a > b$ and $c(E)(a)$ is the number of divisions required to compute the greatest common divisor of a and b using the Euclidean algorithm, then*

$$c(E)(a) = O(\log(a)).$$

The key to the proof of this theorem is the following theorem proved in 1845 by the mathematician Lamé and which makes a surprising connection between the number of divisions required by the Euclidean algorithm and the Fibonacci sequence $F_0 = 0, F_1 = 1, \ldots$ where $F_{n+2} = F_{n+1} + F_n, \;\; n = 0, 1, 2, \ldots$.

Theorem 3.4.2 *If execution of the Euclidean algorithm on the pair of integers a and b with $a > b$ requires k divisions, then*

$$a \geq F_{k+2} \text{ and } b \geq F_{k+1}.$$

PROOF For a pair of integers a, b with $a > b$, if execution of the Euclidean algorithm requires k divisions, this means we apply the Division algorithm k times to form the sequence

$$\begin{aligned} a &= bq_1 + r_1 \quad 0 \le r_1 < b \\ b &= r_1 q_2 + r_2 \quad 0 \le r_2 < r_1 \\ &\cdots \\ r_{k-3} &= r_{k-2} q_{k-1} + r_{k-1} \quad 0 \le r_{k-1} < r_{k-2} \\ r_{k-2} &= r_{k-1} q_k + 0 = r_k \end{aligned}$$

where r_k is the first zero remainder (and r_{k-1} is the gcd). Thus if $k = 1$ division is required, then $r_1 = 0$ so $a = bq_1$ and $a > b$ obviously imply that

$$a \ge 2 = F_3 \quad \text{and} \quad b \ge 1 = F_2.$$

Now we proceed by induction and, assuming that the Theorem is true for some $k \ge 1$, suppose that computation of $\gcd(a, b)$ with $a > b$ requires $k+1$ divisions. This means that the sequence begins with

$$a = bq_1 + r_1$$

and that computation of $\gcd(b, r_1)$ with $b > r_1$ requires k divisions. By the induction hypothesis this means that

$$b \ge F_{k+2} \quad \text{and} \quad r_1 \ge F_{k+1}.$$

To complete the proof we need to see that

$$a \ge F_{k+3} \quad \text{and} \quad b \ge F_{k+2}.$$

We already have $b \ge F_{k+2}$ by the induction hypothesis. For the other inequality, since $a > b$ implies $q_1 \ge 1$ we have

$$\begin{aligned} a &= bq_1 + r_1 \\ &\ge b + r_1 \\ &\ge F_{k+2} + F_{k+1} = F_{k+3} \end{aligned}$$

from the induction hypothesis and the definition of F_{k+3}. This proves the theorem.

To complete the proof of Theorem 3.4.1 we let

$$\phi = \frac{1+\sqrt{5}}{2} \quad \text{and} \quad \hat{\phi} = \frac{1-\sqrt{5}}{2}.$$

Then from Example 2 of Section 3.3 above, we have that

$$F_k = \frac{\phi^k}{\sqrt{5}} - \frac{\hat{\phi}^k}{\sqrt{5}}.$$

Now using a pocket calculator we observe that $\hat{\phi} = -0.618\cdots$ so $\hat{\phi}/\sqrt{5} = -0.276\cdots$, $\hat{\phi}^2 = 0.171\cdots$ and thus

$$\left|\frac{\hat{\phi}^k}{\sqrt{5}}\right| \to 0 \quad \text{as} \quad k \to \infty$$

in a strictly decreasing fashion. It follows that for all k,

$$F_k = \frac{\phi^k}{\sqrt{5}} \quad \text{rounded to the nearest integer.}$$

This is a multiplicative (even exponential) formula for computing the n-th Fibonacci number. The original recursive definition gave us an additive method. (Notice that upward rounding is required for odd k and downward rounding for even k.)

Finally, if computation of $\gcd(a, b)$ with $a > b$ requires k divisions, Theorem 3.4.2 asserts that

$$a \geq F_{k+2} > F_{k+1},$$

and therefore, since F_{k+1} and F_{k+2} differ by at least 1,

$$\frac{\phi^{k+1}}{\sqrt{5}} < a.$$

Hence

$$k + 1 < \log_\phi(\sqrt{5}a)$$

and from this it is easy to see that

$$k = O(\log(a)).$$

This completes the proof of Theorem 3.4.1.

The Fibonacci numbers and, in particular the number $\phi = \frac{1}{2}(1 + \sqrt{5})$ occur in remarkably many seemingly unrelated mathematical contexts. The number ϕ is often called the *golden ratio*; it is supposed to be the ratio of the sides of a rectangle which is most pleasing to the eye. In fact in experiments, large numbers of persons have been asked to adjust the sides of a rectangle until the size is most pleasing; there is statistical evidence that they tend to prefer ratios near the golden ratio!

Chapter 4

Order and Algebra

4.1 Ordered sets and lattices

In our discussion of greatest common divisor and least common multiple in Section 2.2 we introduced the concept of a *lattice*, namely a poset (partially ordered set) in which each pair of elements possess both a least upper bound and a greatest lower bound. In that context the order relation was divisibility (denoted by $|$ and for non-negative integers a and b, $\text{glb}(a, b) = (a, b)$, their greatest common divisor, and $\text{lub}(a, b) = [a, b]$, their least common multiple.

In a general discussion of posets and lattices as abstract axiomatically defined mathematical systems it is customary to denote the order relation by the symbol $\leq$ and a poset by $(P, \leq)$ where P is non-empty and $\leq$ satisfies the axioms

for all $a \in P$, $a \leq a$ (reflexivity)

for all $a, b \in P$, if $a \leq b$ and $b \leq a$, then $a = b$ (antisymmetry)

for all $a, b, c \in P$, if $a \leq b$ and $b \leq c$, then $a \leq c$ (transitivity)

Thus the symbol $\leq$ is not to be confused with, for example, its ordinary meaning on the real numbers. We often denote a poset simply by the set P of its elements, provided there is no confusion about what the order relation is.

Also in a general discussion it is customary to call the least upper bound of a pair a, b of elements (if it exists) the *join* of a and b and denote it by $a \vee b$. Likewise the greatest lower bound of a pair is called their *meet* and is denoted by $a \wedge b$. This notation suggests that if the least upper bound u of two elements exists, it is unique. This is indeed the case, for suppose u' were another least upper bound for the same pair. Then since each of u and u' is both *an* upper bound for the pair and is less than or equal to *any* upper bound for the pair we have both $u \leq u'$ and $u' \leq u$, so that $u = u'$.

In any poset we define $a < b$ to mean that $a \leq b$ but $a \neq b$. Partially ordered sets can be thought of as hierarchies and in the study of abstract posets, as in any hierarchy it is useful to know when one individual is another's immediate superior. The concept of immediate superior can be defined abstractly in any poset as follows.

If a, b are elements of a poset, we say that a *covers* b if $a > b$ and $a > x > b$ is not satisfied by any element x in the poset. If a covers b we denote this by $a \succ b$.

This leads naturally to a graphical representation of any finite poset P. Small circles are drawn to represent the elements of P so that a is higher than b whenever $a > b$. A line segment is then drawn from a to b whenever a covers b. The resulting picture of P is called a *diagram* for P. (Often they are called Hasse diagrams, since their use was popularized by the mathematician H. Hasse.) The following are some examples of diagrams.

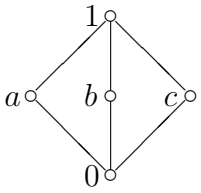


Figure 4.1 (M_3)

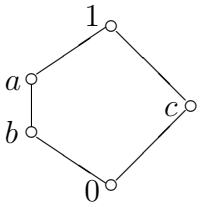


Figure 4.2 (N_5)

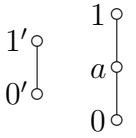


Figure 4.3

The posets represented in Figures 4.1 and 4.2 are lattices while the poset represented in Figure 4.3 is not: for example the elements $1'$ and 1 have neither a meet nor a join. Recall that a chain is a poset in which each pair of elements a, b are comparable, i.e.: either $a \leq b$ or $b \leq a$. Obviously any chain is a lattice; the meet and join of a pair of elements is the lessor and greater of the pair. Thus Figure 4.3 represents a five element poset which is simply the union of two disjoint chains. A poset in which no two elements are comparable is called an *antichain*. Thus the elements a, b, c of the lattice of Figure 4.1, considered as a poset by themselves are an antichain.

As is suggested by the examples above if a poset has a unique largest element it is customary to denote it by 1, and likewise to denote a unique least element of the poset by 0. (The notation used in Figure 4.3 is an exception to this justified by the fact that the disjoint components do have unique greatest and least elements.)

From the axioms for an order relation it is obvious that a subset S of a poset $(P, \leq)$ together with the relation $\leq$ restricted to S is also a poset. Such a poset is called a *subposet* of P. A subposet of a lattice may or may not be a lattice. If it is a lattice and if its meets and joins are the same as in the original lattice it is called a *sublattice* of the original. All subposets of the diagrams above which are lattices are sublattices, but in general this is not necessarily true. Exercise 1 below asks for an example.

While we can actually draw (complete) diagrams only for finite posets, it is also frequently useful to draw finite subposets of infinite posets and also, to the extent we are able, to at least imagine the diagrams of infinite posets. We are already familiar with doing this for the real numbers relative to the usual ordering.

4.2 Isomorphism and duality

Consider the set consisting of the cartesian plane $\mathbb{R}^2$, the origin O and any three distinct lines passing through the origin. This set of five subspaces of $\mathbb{R}^2$ together with the order relation of set inclusion $\subset$ is obviously a poset; in fact it is a lattice, for the least subspace of $\mathbb{R}^2$ (in the sense of set inclusion) which contains both of two distinct lines through the origin is $\mathbb{R}^2$ itself. Likewise the largest subspace (in the sense of inclusion) which is contained in both of two such lines is the subspace $\{O\}$. Thus $\mathbb{R}^2$ and $\{O\}$ are the 1 and 0 elements of this lattice. Evidently this lattice is, except for the nature of its elements and the order relation, no different than the poset described by Figure 4.1; i.e.: although they are totally different structures, *as posets* they are indistinguishable. We say they are isomorphic. Formally, two posets $(P_1, \leq_1)$ and $(P_2, \leq_2)$ are *isomorphic* if there is a 1-1 correspondence between P_1 and P_2 with the property that if $a_1 \leftrightarrow a_2$ and $b_1 \leftrightarrow b_2$ in this correspondence, then $a_1 \leq_1 b_1$ holds in P_1 *if and only if* $a_2 \leq_2 b_2$ holds in P_2. The double implication *if and only if* in this definition is essential. (See Exercise 2 below.) From the definitions of meet and join it is easy, but important, to see that if P_1 is a lattice and is isomorphic as a poset to P_2, then P_2 is also a lattice. (Take a moment to prove this.) Obviously in a general discussion of posets and lattices we usually do not want to distinguish between isomorphic posets. The situation here is essentially the same as in most axiomatically defined areas of mathematics, for example in linear algebra, where isomorphic vector spaces are frequently identified with each other.

Now let us show that a finite poset is uniquely determined (up to isomorphism) by its Hasse diagram; from this it will follow that two finite posets are isomorphic iff they have the same diagrams. To see this simply observe that for elements a, b the relation $a < b$ holds if and only if there is a sequence of elements $x_0, x_1, \ldots, x_n$ such that $a = x_0$, $b = x_n$, and x_{i-1} is covered by x_i for all $i = 1, \ldots, n$. Graphically this means that $a < b$ holds just in case one can move from a to b upward along a broken line. Using this fact one can easily check to see if two finite posets are isomorphic or not by simply comparing their diagrams. Any isomorphism must be 1-1 between lowest elements, between elements covering lowest elements, and so on up to the greatest elements. Corresponding elements must be covered by equal numbers of different elements, etc.. Of course the diagram of a poset is itself a poset, one whose elements are circles on paper and in which $a \leq b$ means that either $a = b$ or that b lies above a along a broken line.

In a poset with order relation $\leq$ the relation $\geq$, where $a \geq b$ means $b \leq a$, is called the *converse* of $\leq$. Hence the converse of the inclusion relation $\subset$ among sets is the containment relation $\supset$. The converse of the relation $\leq$ among numbers is $\geq$. From the reflexive, symmetric, and transitive properties defining posets, it is very easy to check that if $(P, \leq)$ is a poset, the so is $(P, \geq)$. For example to verify antisymmetry in $(P, \geq)$, suppose that $a \geq b$ and $b \geq a$. Then by the definition of $\geq$ we have $b \leq a$ and $a \leq b$ and by the antisymmetry of $\leq$ we conclude that $a = b$. Reflexivity and

transitivity are equally trivial to check. The poset $(P, \geq)$ is called the dual of $(P, \leq)$, we often denote it by $(P, \leq)^d$, or briefly by P^d. Notice that the dual of the dual of P is just P again. This observation has the simple but important consequence that the set of all posets is precisely the same as the set of all duals of all posets. We will make use of this consequence in the proof of Theorem 4.2.1 below. If P is finite with a particular diagram, then the diagram of its dual is obtained by simply turning its diagram upside down.

Now suppose that Φ is some "sentence" about posets. By this we mean that Φ makes assertions about elements and the relation $\leq$ among them. If in Φ we replace all occurrences of $\leq$ by $\geq$ we obtain another sentence, the dual of Φ (which we can denote by Φ^d). Then we have the following theorem, called the "Duality Principle".

Theorem 4.2.1 *(Duality Principle) If Φ is a sentence which is true for all posets then its dual is also true for all posets.*

PROOF The sentence Φ is true in $(P, \leq)$ iff its dual is true in $(P, \geq)$, which is also a poset. Hence if Φ is true for all posets then its dual is true for all duals of all posets, which as observed above is the same as the set of all posets! This completes the proof.

Consider a simple application of the Principle of duality. Let Φ be the sentence "If a set S of elements has a least upper bound, then the least upper bound is unique". Since a least upper bound of S is an element u such that $x \leq u$ for all $x \in S$, and such that if u' is any other upper bound of S then $u \leq u'$, it follows that the dual Φ^d is the sentence "If a set S has a greatest lower bound, then the greatest lower bound is unique." Since Φ is true for all posets it follows that Φ^d is also true for all posets. Further, by the definition of lattices it follows from the fact that each pair of elements of a lattice has a *unique* join together with the Duality Principle, that each pair also has a *unique* meet. This illustrates how in the theory of posets, one has to prove only half of the theorems; the others come for free by the Duality Principle!

When a poset Q is isomorphic to the dual of the poset P, we often say that Q and P are *dually isomorphic* or *anti-isomorphic*. This simply means that there is a 1-1 correspondence between Q and P such that comparable members of Q correspond to members of P which are comparable via the converse. In terms of diagrams an element of Q lies above another iff it corresponds to an element of P below the correspondent of the other. Some important posets are *self-dual*, i.e.: are isomorphic to their duals (anti-isomorphic to themselves). All of the posets described in Figures 4.1, 4.2, 4.3 have this property.

Finally, since both the meet and join of pairs of elements in a lattice are defined in terms of the order relation and since the definition of the meet is just dual to the definition of join, we can profitably specialize the Duality Principle to lattices as follows. Let Φ now be a sentence about lattices, i.e.: a sentence which makes assertions about elements, the relation $\leq$ among them, and also about $\vee$ and $\wedge$. Let the dual Φ^d of Φ be obtained by interchanging the occurrences of $\leq$ and $\geq$

and likewise interchanging the occurrences of $\vee$ and $\wedge$. Then we have the following immediate corollary to Theorem 4.2.1.

Corollary 4.2.1 *(Duality Principle for lattices) If Φ is a sentence which is true for all lattices then its dual is also true for all lattices.*

The following examples illustrate the concepts just discussed.

a) Let $(\mathcal{P}(X), \subset)$ be the poset consisting of all subsets of set X (i.e.: consisting of the power set of X) ordered by inclusion. This is a lattice where the join of two subsets is their union and the meet is the their intersection. This lattice is self-dual: the correspondence which carries each subset of X into its complement is 1-1 and inverts inclusion.

b) Recall the discussion in Section 2.4 of the lattices $(\mathbb{N}, |)$ (non-negative integers under divisibility) and $(\mathcal{M}, \subset)$, (modules of integers under set inclusion) where it was already observed that these lattices are dually isomorphic. On the other neither of these is self-dual. From the standpoint of $(\mathbb{N}, |)$ this is easily seen since each of the primes covers the integer 1, while the integer 0 covers no integer. This corresponds to the fact that the modules (p), p a prime, are covered by $\mathbb{Z}$ in $(\mathcal{M}, \subset)$, but in this lattice no module covers (0).

c) The poset of all subspaces of the n-dimensional Euclidean vector space $\mathbb{R}^n$, ordered by set inclusion, is a lattice in which the meet of two subspaces is their intersection while the join of two subspaces is the subspace generated by their union, also known as the linear span of the two subspaces. This lattice has a largest element $1 = \mathbb{R}^n$ and a least element $0 = \{o\}$ where o is the null vector. (The lattice of Figure 4.1 is a sublattice of the subspace lattice of $\mathbb{R}^2$.) This lattice is also self-dual: the correspondence which carries each subspace S into its orthogonal complement $S^\perp$ is 1-1 and inverts inclusion. (Recall that $S^\perp$ is the subspace consisting of all vectors which are orthogonal to each vector in S.)

In a lattice, or any poset, any cover of a least element is called an *atom*. Dually, any element covered by a maximum element is called a *coatom*. In the lattice M_3 of Figure 4.1. the elements a, b, c are both atoms and coatoms. In Figure 4.2 a is a coatom while b is an atom. In example a) above the atoms are just the one-element subsets of X (singletons) while the coatoms are the subsets of X of the form $X - \{a\}$ for each $a \in X$. In example b) in the lattice $(\mathbb{N}, |)$ the primes are the atoms and there are no coatoms. In the dual, the modules (p), p a prime, are the coatoms and there are no atoms. In c) the lines (1-dimensional subspaces) are the primes while the $n-1$ dimensional subspaces are the coatoms of the lattice of subspaces of $\mathbb{R}^n$.

Exercises Sections 4.1 and 4.2

1. Draw the diagram of a lattice and a subposet which is a lattice but which is not a sublattice.

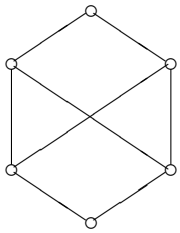
2. Give an example of two non-isomorphic posets $(P_1, \leq_1)$ and $(P_2, \leq_2)$ such that there is a 1-1 correspondence between P_1 and P_2 satisfying "if $a_1 \leftrightarrow a_2$ and $b_1 \leftrightarrow b_2$ and $a_1 \leq b_1$ then $a_2 \leq b_2$".

3. Let F consist of all real valued functions continuous on the interval $[0, 1]$ and define $f \leq g$ to mean that $f(x) \leq g(x)$ for all $x \in [0, 1]$. Show that $(F, \leq)$ is a lattice and describe the meet and join of pairs of elements.

4. Show that special relativity partially orders space-time $\mathbb{R}^4 = \{(x, y, z, t)\}$ by defining $(x_1, y_1, z_1, t_1) \leq (x_2, y_2, z_2, t_2)$ iff $((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)^{\frac{1}{2}} \leq t_1 - t_2$. What is the physical meaning of this relation?

5. Show that there are exactly five non-isomorphic posets of three elements, three of which are self-dual and one of which is a lattice.

6. Is the following poset a lattice?



7. Describe a poset containing three elements x, y, z which satisfy the following conditions:

a) $\{x, y, z\}$ is an antichain,
b) none of the pairs $\{x, y\}$, $\{x, z\}$, $\{y, z\}$ has a least upper bound,
c) the set $\{x, y, z\}$ has a least upper bound.

8. Verify that each of the following sentences is true in all lattices. Then write down the dual of each of the sentences.

a) If z is an upper bound of $\{x, y\}$ then $x \vee y \leq z$.
b) For all x and y, $x \wedge y \leq x \leq x \vee y$ and $x \wedge y \leq y \leq x \vee y$.

4.3 Lattices as algebras

In Section 4.1 we observed that in any poset, least upper bounds and greatest lower bounds, when they exist, are unique. In particular in a lattice both $x \vee y$ and $x \wedge y$ exist and are unique for each pair of elements x, y. Thus we may think of $\vee$ and $\wedge$ as binary operations in any lattice, and this suggests that we think of lattices as

algebraic systems. This idea has had a rich history and, in fact, prior to 1850 George Boole began to exploit the close analogy between laws governing the properties of expressions like $x \vee y$, $x \wedge (y \vee z)$, etc. (i.e.: terms formed from $\vee$ and $\wedge$) and terms like $x + y$, $x(y + z)$ etc., in ordinary algebra. (See Exercise 4 below for a precise definition of lattice terms.) Indeed it is easy to verify that in any lattice the following "laws"(or "identities") hold for all elements x, y, z.

$$\begin{aligned}
&x \wedge x = x \text{ and } x \vee x = x,\\
&x \wedge y = y \wedge x \text{ and } x \vee y = y \vee x,\\
&x \wedge (y \wedge z) = (x \wedge y) \wedge z \text{ and } x \vee (y \vee z) = (x \vee y) \vee z,\\
&x \wedge (x \vee y) = x \text{ and } x \vee (x \wedge y) = x.
\end{aligned}$$

These identities are known respectively as the *idempotent*, *commutative*, *associative*, and *absorption* laws. Each of these laws is listed together with its dual. By the Duality Principle we need verify only one of the dual pairs. The reader should take a few moments to do this carefully.

Now let us show that these identities completely characterize lattices as algebraic systems. More precisely we have the following theorem.

Theorem 4.3.1 *Let L be a nonempty set, and $\vee$ and $\wedge$ be binary operations defined on L and satisfying the idempotent, commutative, associative, and absorption laws (together with their duals). Define the relation $\leq$ on L by stipulating that*

$$x \leq y \iff x \wedge y = x \iff x \vee y = y.$$

Then $\leq$ is a partial order on L and $(L, \leq)$ is a lattice in which $lub(x, y) = x \vee y$ and $glb(x, y) = x \wedge y$.

Thus the theorem says that we can view lattices either as posets with each pair of elements having a meet and join (as we have so far), or equivalently as algebraic systems having two binary operations satisfying the identities listed above. The choice is a matter of convenience and in fact we often use the two operations together with the partial order. Examples of this will follow.

PROOF First observe that the two conditions $x \vee y = y$ and $x \wedge y = x$, used to define $x \leq y$, are equivalent: if $x \wedge y = x$ then $x \vee y = (x \wedge y) \vee y = y$ using absorption and commutativity. The converse follows by using the dual absorption and commutative laws.

Now we show that $\leq$ as defined is a partial order on L. Reflexivity $x \leq x$ is immediate since $x \wedge x = x$. For antisymmetry, suppose $x \leq y$ and $y \leq x$ and thus $x \wedge y = x$ and $y \wedge x = y$ so that $x = y$ by commutativity of $\wedge$. For transitivity suppose $x \leq y$ and $y \leq z$ which means that $x \wedge y = x$ and $y \wedge z = y$. Hence $x \wedge z = (x \wedge y) \wedge z = x \wedge (y \wedge z) = x \wedge y = x$, using associativity of $\wedge$. Hence $x \leq z$. Therefore $\leq$ is a partial order.

Next observe that $x \leq x \vee y$ and $y \leq x \vee y$ since

$$x \vee (x \vee y) = (x \vee x) \vee y = x \vee y, \text{ and}$$

$$y \vee (x \vee y) = y \vee (y \vee x) = (y \vee y) \vee x = y \vee x = x \vee y.$$

Hence $x \vee y$ is *an* upper bound for both x and y. Finally, suppose that $x \leq z$ and $y \leq z$, that is, that z is any upper bound for both. Then $x \vee z = z$ and $y \vee z = z$ so

$$(x \vee y) \vee z = x \vee (y \vee z) = x \vee z = z$$

which means that $x \vee y \leq z$ so $x \vee y = \text{lub}(x, y)$. Using the dual laws we likewise determine that $x \wedge y = \text{glb}(x, y)$. This completes the proof.

Exercises Section 4.3

1. The idempotent, commutative, associative, and absorption identities (and their duals) are not independent laws for lattices. For example prove that the absorption identities imply the idempotency of $\wedge$ and $\vee$. (Hint: simplify $a \vee [a \wedge (a \vee a)]$ in two ways to yield $a = a \vee a$.) Note: This exercise shows that the number of identities needed to define lattices in the theorem of this section can certainly be reduced from 8 to 6, and raises the possibility of making a further reduction. In fact, in 1970 Ralph McKenzie showed that lattices can be described by a single (not very interesting) identity involving $\wedge$ and $\vee$.

2. Suppose that $(L, \wedge, \vee)$ is a lattice and that S is a subset of L which is closed under the operations $\wedge$ and $\vee$. Show that $(S, \wedge, \vee)$ is a sublattice of L (in the sense in which this concept was defined in Section 4.1.)

3. Suppose that $(L, \wedge, \vee)$ is a lattice and that $(M, \wedge^*, \vee^*)$ is a system consisting of the non-empty set M which is closed under two binary operations $\wedge^*$ and $\vee^*$.
a) If $f : L \to M$ is a function from L *onto* M with the properties

$$f(x \wedge y) = f(x) \wedge^* f(y), \text{ and}$$

$$f(x \vee y) = f(x) \vee^* f(y),$$

for all $x, y \in L$, then $(M, \wedge^*, \vee^*)$ is also a lattice. (f is called a *lattice homomorphism* and M is a homomorphic image of L. Thus you have shown that homomorphic images of lattices are again lattices.)
b) Show that a lattice homomorphism is order preserving but that an order preserving function from a lattice onto a lattice is not necessarily a lattice homomorphism.

4. Define lattice *terms* in the variables $x_1, x_2, \ldots$ inductively (recursively) to be any of

a) the variables $x_1, x_2, \ldots$, or
b) $(t \vee s)$ or $(t \wedge s)$, where t and s are lattice terms.

Thus lattice terms are (similar to ordinary terms in algebra) compositions of the x_i obtained by applying $\wedge$ and $\vee$. Show that lattice terms determine order preserving functions on lattices. i.e.: suppose $t(x_1, \ldots, x_n)$ is a lattice term and for $i = 1, \ldots, n$, a_i, b_i are elements of a lattice L with $a_i \leq b_i$. Show that $t(a_1, \ldots, a_n) \leq t(b_1, \ldots, b_n)$ in L. (Hint: Since terms are defined recursively (inductively) you must use induction.)

5. As an application of Exercise 4, above, prove that every lattice satisfies the so called one-sided distributive laws:

$$x \wedge (y \vee z) \geq (x \wedge y) \vee (x \wedge z),$$

$$x \vee (y \wedge z) \leq (x \vee y) \wedge (x \vee z).$$

(Prove the first law by carefully applying Exercise 4; then invoke the duality principle for the second.)

4.4 Modular and distributive lattices

In Section 2.4 we observed that the lattice $(\mathbb{N}, |)$ satisfied the distributive law

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

where $\wedge$ and $\vee$ denoted gcd and lcm respectively. The same law also holds in the power set lattice $(\mathcal{P}(X), \subset)$ where $\wedge$ and $\vee$ are intersection and union respectively. It is also easy to see that the duals of these laws also holds in each of these lattices. This is explained by the following lemma.

Lemma 4.4.1 *In any lattice the following dual distributive laws are equivalent:*

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z),$$

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z).$$

Proof Let us denote the first of the laws by D and the second by D^d. Suppose D holds in a lattice. Then

$$\begin{aligned}
(x \vee y) \wedge (x \vee z) &= ((x \vee y) \wedge x) \vee ((x \vee y) \wedge z) \\
&= x \vee (z \wedge (x \vee y)) \\
&= x \vee ((z \wedge x) \vee (z \wedge y)) \\
&= (x \vee (z \wedge x)) \vee (z \wedge y) \\
&= x \vee (y \wedge z)
\end{aligned}$$

where the first and third equalities are obtained using D and the other equalities use other basic lattice laws. Hence in any lattice the sentence

$$\text{for all } x, y, z \;\; D \Longrightarrow D^d$$

is true. By the duality principle for lattices it follows that in any lattice the sentence

$$\text{for all } x, y, z \;\; D^d \Longrightarrow D$$

is true. hence D and D^d are equivalent.

Definition A lattice L is distributive if either (and hence both) of the dual distributive laws holds in L.

The lemma tells us that a lattice is distributive if and only if its dual is distributive.

Exercise 5 of the preceding section shows that every lattice satisfies the so called one sided distributive laws. Hence, we see that a lattice L is distributive if and only if either one of the following inequalities holds in L:

$$x \wedge (y \vee z) \leq (x \wedge y) \vee (x \wedge z),$$

$$x \vee (y \wedge z) \geq (x \vee y) \wedge (x \vee z).$$

Finally, any sublattice of a distributive lattice is again distributive. This is because the distributive law is an identity which if it holds "globally" in a lattice necessarily holds in any sublattice. Likewise, it takes not much more effort to show that any homomorphic image of a distributive lattice is distributive. Even more generally if a lattice satisfies *any* identity or inequality, then the same is inherited by any sublattice or any homomorphic image. (These are very simple observations but they turn out to be very important.)

Some of the most important lattices which occur in mathematics are distributive; the most familiar example are the lattices $(\mathcal{P}(X), \subset)$ and $(\mathbb{N}, |)$. Another important class of distributive lattices are chains, e.g.: the lattice of real numbers with respect to the usual order relation. This is easy to see using the fact that in a chain lub and glb become simply max and min respectively. Distributivity is also attractive because one of the defining laws is so much like the familiar distributive law of arithmetic. On the other hand not all of the most important lattices which occur in mathematics are distributive. The most important and conspicuous example of the failure of distributivity is in the lattice of all subspaces of a vector space. For example if a, b, c are distinct lines through the origin in $\mathbb{R}^2$, and o is the null vector, then

$$a \wedge (b \vee c) = a \wedge \mathbb{R}^2 = a > \{o\} = \{o\} \vee \{o\} = (a \wedge b) \vee (a \wedge c).$$

This failure is also pictured in the lattice M_3 of Figure 4.1, and indeed M_3 is isomorphic to the sublattice (of the whole subspace lattice) consisting of $a, b, c, \{o\}, \mathbb{R}^2$. Also

it is just as easy to check that the lattice N_5 of Figure 4.2 is also non-distributive.(Do this.)

The question then arises, is there some weaker identity than the distributive law, which is not just trivial (i.e.: holds in all lattices) and which holds in the important lattice of subspaces of a vector space. The answer is yes and the law is that which is obtained by requiring distributivity to hold for elements, x, y, z, not always, but whenever any pair of them are comparable. For example if $x \geq z$ (so that $x \wedge z = z$) we require that

$$x \wedge (y \vee z) = (x \wedge y) \vee z.$$

This leads to the following definition.

Definition A lattice L is *modular* if the sentence

$$\text{for all } x, y, z \in L \;\; x \geq z \Longrightarrow x \wedge (y \vee z) = (x \wedge y) \vee z$$

holds in L. This sentence is called the modular law. It was first recognized and studied by Dedekind before 1900.

From the definition it is clear that every distributive lattice is modular. You should stop and check immediately (considering a few cases) that the lattice M_3 is modular. Since it is not distributive, this shows that modularity, as a property of lattices is strictly weaker than distributivity. Thus the class of modular lattices is intermediate between the class of distributive lattices and the class of all lattices. For reasons we will suggest below modular lattices are the most important general class of lattices.

In the notation M_3, M stands for "modular" and 3 indicates that this lattice has 3 atoms (=coatoms). M_3 is also often called the "diamond". In general M_k is used to denote the modular lattice which looks just like M_3 except that it has k atoms. The lattice of all subspaces of $\mathbb{R}^2$ can be visualized as M_c where c (for the continuum) counts the distinct one-dimensional subspaces. The lattice N_5 of Figure 4.2, is not modular, since $a > b$ but

$$a \wedge (c \vee b) = a > b = (a \wedge c) \vee b.$$

Thus the N stands for "non-modular" and the 5 just indicates that it has 5 elements. N_5 is often called the "pentagon".

A few simple remarks about the definition of modular lattices are in order:

a) The modular law is self-dual: just interchange x and z and observe that except for notation the dual appears! This means that as in the case of distributivity, a lattice is modular iff its dual is modular and to test a lattice for modularity we need to check only one of the dual laws.

b) In any lattice the following one-sided modular laws hold:

$$x \geq z \Longrightarrow x \wedge (y \vee z) \geq (x \wedge y) \vee z,$$

$$x \leq z \Longrightarrow x \vee (y \wedge z) \leq (x \vee y) \wedge z.$$

This is an application of Exercise 4 of Section 4.3, for by that exercise, $x \wedge (y \vee z)$ is an upper bound for both $x \wedge y$ and z, and hence is $\geq$ their join. The second inequality follows from the first by duality. Hence as in the case of distributivity to check for modularity we need only check one inequality.

c) Another way to remember the modular law, and which is helpful when performing computations with lattice terms in a modular lattice, is to observe that formally, modularity allows one to re-parenthesize $x \wedge (y \vee z)$ as $(x \wedge y) \vee z$ if one knows that $x \geq z$.

d) While the modular law, as presented above, is logically in the form of a universally quantified (for all x, y, z,...) implication, (and not a universally quantified identity as is distributivity) it is actually equivalent to a (universally quantified) identity, namely

$$(x \vee z) \wedge (y \vee z) = [(x \vee z) \wedge y] \vee (x \wedge z).$$

We leave the verification of this as an exercise. In this counter intuitive form modularity is much less easy to apply, so we seldom use it. It is however important that modularity is expressible as an identity.

e) Finally, the universal quantifiers (for all x, y, z,...) in the definitions of both distributivity and modularity are critical. For example it is not true that if particular elements a, b, c in an arbitrary lattice satisfy $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$, then they also satisfy $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$. Also, without the universal quantifiers we would not be able to prove that sublattices and homomorphic images of distributive, (respectively modular), lattices, are distributive, (respectively modular).

In the discussion above we have observed that the lattice M_3 is modular but not distributive and N_5 is not modular, and hence is not distributive. It follows that no modular lattice can contain a sublattice isomorphic with N_5 and no distributive lattice can contain an isomorphic copy of either M_3 or N_5. This proves half of the following attractive characterization theorem for modular and distributive lattices:

Theorem 4.4.1 *For a lattice L,*
a) L is modular iff L does not contain a sublattice isomorphic to N_5.
b) L is distributive iff L does not contain a sublattice isomorphic to either N_5 or M_3.

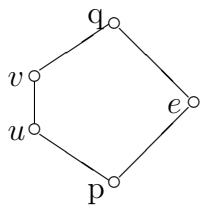
This theorem, sometimes called the M_3-N_5 theorem, gives a way of determining if a lattice is modular or distributive by inspecting the diagrams of its finite sublattices, rather than by checking the modular or distributive laws for all choices of elements.

Statement a) was proven by Dedekind in 1900 and b) was proven by Garrett Birkhoff in 1934.

PROOF We already know the "only if" directions of both a) and b). To prove the "if" direction of a) we must show (contrapositively) that if L is a non-modular lattice, then L contains a sublattice isomorphic to N_5. So suppose that L is non modular. Then L contains elements d, e, f such that $d > f$ and $v > u$ where

$$v = d \wedge (e \vee f) \quad \text{and} \quad u = (d \wedge e) \vee f.$$

We claim that $e \wedge u = e \wedge v$ ($= p$ say) and $e \vee u = e \vee v$ ($= q$ say). If we prove this claim then, using it, it is easy to see that the five elements u, v, e, p, q are distinct, and constitute a sublattice isomorphic with N_5. The reader should check these details.



To prove the claim we first have, using associativity and absorption,

$$v \wedge e = (d \wedge (e \vee f)) \wedge e = d \wedge ((e \vee f) \wedge e) = d \wedge e,$$

and

$$u \vee e = ((d \wedge e) \vee f) \vee e = ((d \wedge e) \vee e) \vee f = e \vee f.$$

Also we have, using Exercise 4 of Section 4.3, and the first of the equations above,

$$d \wedge e = (d \wedge e) \wedge e \leq u \wedge e \leq v \wedge e = d \wedge e,$$

and likewise, using the second equation above,

$$e \vee f = u \vee e \leq v \vee e \leq (e \vee f) \vee e = e \vee f.$$

Combining these we establish the claim and hence the proof of a).

The proof of a) just given, while certainly correct, is not very well motivated. How did we know how to choose the elements p and q? In order to give a completely motivated proof we would have to introduce the idea of a *free lattice*, which is beyond the scope of the present text. For this reason we will not bother with the proof of part b) which in the unmotivated form we would have to give is even less instructive. (For an excellent discussion of this topic consult, G. Grätzer, *Lattice theory: first concepts and distributive lattices*, W. H. Freeman and Co., San Francisco, 1971.)

Finally, why are modular lattices so important? The key is to observe a critical difference between M_3 and N_5, namely that in N_5 there are two *maximal* chains between 0 and 1, i.e.: the chain $0 \prec b \prec a \prec 1$ has 4 elements while $0 \prec c \prec 1$ has

only 3. (The chain $0 \prec b \leq 1$ has 3 elements but is not maximal, since the element a can be inserted in it.) On the other hand all maximal chains joining 0 and 1 in M_3 have 3 elements. In fact, using induction together with the M_3-N_5 theorem it can be proven that modular lattices can be characterized as lattices which satisfy the Jordan-Dedekind chain condition:

Jordan Dedekind chain condition All finite maximal chains between two elements have the same number of elements. (If a chain has $n+1$ elements it is said to have *length* n.)

In a modular lattice with 0 in which all chains from 0 to any element x are finite, the length of (any) maximal chain from 0 to x is called the dimension of x, denoted $d(x)$. Thus in any such modular lattice the dimension function $d(x)$ is well defined.

As an example of the significance of the Jordan-Dedekind chain condition, we begin by showing that the lattice of subspaces of any vector space is a modular lattice. To show this observe that for subspaces X and Y of the vector space V, $X \wedge Y = X \cap Y$ while $X \vee Y = \{x + y : x \in X, y \in Y\}$, the latter is just the linear span of the union of X and Y. Now suppose that X, Y, Z are subspaces of V and that $X \supset Z$. We need to show that

$$X \wedge (Y \vee Z) \subset (X \wedge Y) \vee Z.$$

Let v be a vector in the left hand side. Then

$$x \in X \quad \text{and} \quad x \in Y \vee Z \quad \text{so} \quad x = y + z$$

for some $y \in Y$ and $z \in Z$. Since $X \supset Z$ it follows that $z \in X$ and hence $y = x - z \in X$ so $y \in X \wedge Y$. Therefore $x = y + z \in (X \wedge Y) \vee Z$. Hence the modular law holds in this lattice. Now let V be a vector space having some finite basis $v_1, ..., v_n$. Let $\{o\}$ be the subspace consisting of the null vector alone and consider the chain

$$\{o\} \subset s(v_1) \subset s(v_1, v_2) \subset \ldots \subset s(v_1, \ldots, v_n) = V,$$

where $s(v_1)$ denotes the subspace spanned by v_1, $s(v_1, v_2)$ the subspace spanned by v_1 and v_2, etc.. From the linear independence of the basis it is easy to show that this chain is maximal. From this it follows from the Jordan Dedekind chain condition that the numbers of elements in any two bases are the same. Hence this fundamental property of vector spaces follows directly form the fact that the subspace lattice of V is modular. This example explains, in large part why modular lattices are so important.

Exercises Section 4.4

1. Show that a lattice L is distributive iff

$$(x \wedge y) \vee (x \wedge z) \vee (y \wedge z) = (x \vee y) \wedge (x \vee z) \wedge (y \vee z)$$

is an identity of L.

2. In any lattice L define the function m by either

$$m(x, y, z) = (x \wedge y) \vee (x \wedge z) \vee (y \wedge z)$$

or

$$m(x, y, z) = (x \vee y) \wedge (x \vee z) \wedge (y \vee z).$$

(Unless the lattice is distributive Exercise 1 asserts that these two functions will be different.) Show that with either definition, m is a *majority* function, i.e.:

$$m(x, x, z) = m(x, y, x) = m(w, x, x) = x$$

for all x, y, z, w. With either definition m is often called a *median* term. Justify this by describing how to compute $m(x, y, z)$ if L is a chain.

3. Prove directly from the definitions that every distributive lattice is modular.

4. Show that M_3 and N_5 are the only non-distributive lattices of five elements.

5. An element a of a lattice is said to be *join irreducible* if

$$\text{for all } x, y, \quad a = x \vee y \Rightarrow a = x \quad \text{or} \quad a = y,$$

i.e.: a is not the join of any pair of strictly smaller elements.

a) Show that if a is an atom of a lattice then a is join irreducible.

b) Describe the join irreducible elements of $(\mathbb{N}, |)$.

c) Dually, an element of a lattice is *meet irreducible* if

$$\text{for all } x, y, \quad a = x \wedge y \Rightarrow a = x \quad \text{or} \quad a = y,$$

i.e.: a is not the meet of any pair of strictly larger elements. Show that any coatom is meet irreducible. What are the meet irreducible elements of $(\mathbb{N}, |)$?

d) Describe the join and meet irreducible elements of the lattice of all subspaces of a finite dimensional vector space. Can you say anything about infinite dimensional spaces?

6. Prove that in any modular lattice each element of finite dimension is the join of a finite set of join irreducible elements. Using Exercise 5 b) above apply this to the lattice $(\mathbb{N}, |)$.

7. Suppose L is a distributive lattice and

$$a \leq x_1 \vee \cdots \vee x_n$$

where $a, x_1, \ldots, x_n \in L$ and a is join irreducible. Prove that $a \leq x_i$ for some $i = 1, \ldots, n$. (Hint: Observe that $a \leq x_1 \vee \cdots \vee x_n$ implies $a = a \wedge (x_1 \vee \cdots \vee x_n)$.)

8. In a distributive lattice suppose the element a is represented as

$$a = x_1 \vee \cdots \vee x_n$$

where $x_1, \ldots, x_n$ are join irreducible. We say that this representation is *irredundant* if a is *not* the join of any proper subset of $\{x_1, \ldots, x_n\}$. Prove that in a distributive lattice any representation of an element a as an irredundant join of join irreducible elements is unique. (Hint: Suppose

$$a = x_1 \vee \cdots \vee x_n = y_1 \vee \cdots \vee y_m$$

where the x_i and y_i are join irreducible. Show that $m = n$ and that each $x_i \leq y_j$ for some j.)

Apply this result to the lattice $(\mathbb{N}, |)$.

9. Verify that the identity in the remark d) following the definition of modularity is equivalent to the modular law.

4.5 Boolean algebras

Historically, the mid-nineteenth century mathematician George Boole was the first to study lattices. Boole was concerned with formulating an "algebra of (propositional) logic" and hence he considered only lattices which were distributive and in addition each element had a complement which had properties like those of the complement of a set or the negation of a proposition. Consequently these lattices are named in his honor. Thus Boolean algebras are closely connected to both elementary propositional logic and to the algebra of sets and it is in these contexts and their applications (for example to electronic circuit design) that most students will have encountered them before. Our approach is to develop the study of Boolean algebras by approaching the subject via general lattice concepts.

We begin with the notion of a complement.

Definition Let L be a lattice with both a 0 and a 1. For $a \in L$, an element $b \in L$ is a *complement* of a if $a \vee b = 1$ and $a \wedge b = 0$.

Referring to both M_3 and N_3 it is clear that complements in a non-distributive lattice are not unique. Thus the following theorem expresses an important property of distributive lattices.

Theorem 4.5.1 *In a distributive lattice, complements, when they exist, are unique.*

For example in a finite chain $0 < a_1 < \cdots < a_n < 1$ only 0 and 1 have complements.
PROOF Suppose the element $a \in L$ has both x and y as complements. Then we have both

$$a \vee x = 1, \;\; a \wedge x = 0, \text{ and } \;\; a \vee y = 1, \;\; a \wedge y = 0.$$

From $a \vee x = 1$ and $a \wedge y = 0$ we have

$$x = 0 \vee x = (a \wedge y) \vee x = (a \vee x) \wedge (y \vee x) = 1 \wedge (y \vee x) = (y \vee x).$$

From $a \vee y = 1$ and $a \wedge x = 0$ we interchange the roles of x and y in the equalities above to obtain $y = x \vee y$. Hence $x = y$ which proves uniqueness.

Definitions In a distributive lattice if an element a has a complement denote it by a'. Thus for any complemented a, $a \vee a' = 1$ and $a \wedge a' = 0$.

A lattice $(L, \vee, \wedge)$ is a *Boolean lattice* if it is both distributive and *every* element of L has a complement. (Thus a Boolean lattice necessarily has both a 0 and a 1.)

In a Boolean lattice it is often more natural to regard the complement $'$ as a unary operation (since x' is uniquely determined for each element x), and in addition to take the distinguished elements 0 and 1 as *nullary* (i.e.: constant) operations. Accordingly a *Boolean algebra* B is a system $(B, \vee, \wedge, ', 0, 1)$ where the system $(B, \vee, \wedge)$ is a Boolean lattice and $'$ is the complement operation, and 0 and 1 are the uniquely determined distinct least and greatest elements of the lattice. Notice that this forces a Boolean algebra two have at least two elements 0, and 1; on the other hand the one element lattice is a Boolean lattice.

Suppose a subset of a Boolean algebra is closed under the operations $\vee, \wedge, '$, and 0 and 1. Such a subset is called a subalgebra and will again be a Boolean algebra with least and greatest elements 0 an 1. This is because of the universal quantification of the identities defining Boolean algebras. e.g.: $x \vee x' = 1$ is such a law. In particular, in any Boolean algebra the two elements 0 and 1 by themselves are a Boolean algebra, which we usually refer to as the two element Boolean algebra. In contrast to this situation, it is perfectly possible to have a sublattice of a Boolean lattice which is itself a Boolean lattice but with least and greatest elements which are not the same as those in the enveloping Boolean lattice. Thus, in spite of the fact that the distinction between Boolean lattices and Boolean algebras appears to be solely notational, the result of adding complementation and 0 and 1 as new operations is that the two classes of systems are different in important ways.

Just as in the case of lattices we often refer to a Boolean algebra by naming just its set B of elements.

The complement operation on a Boolean algebra B has as one of its most important roles the fact that the mapping $x \mapsto x'$ determines an isomorphism of B onto its

dual. To prove this first observe that from the definition of the complement of the elements a and a', we have, using commutativity, both

$$a' \vee a = 1,\ a' \wedge a = 0 \quad \text{and} \quad a' \vee (a')' = 1,\ a' \wedge (a')' = 0,$$

which implies that $(a')' = a$ by the uniqueness of complements. From this it follows first that $x' \mapsto x$ so that the mapping ia function from B onto B. Also $x' = y'$ implies $x = (x')' = (y')' = y$ so the mapping is a 1-1 correspondence of B onto itself. Next, if $x \leq y$ then $x \wedge y' \leq y \wedge y' = 0$ so $x \wedge y' = 0$. Therefore, as in the proof of Theorem 4.5.1,

$$x' = x' \vee 0 = x' \vee (x \wedge y') = (x \vee x') \wedge (x' \vee y' = 1 \wedge (x' \vee y') = x' \vee y'$$

and from this we conclude that $y' \leq x'$. Repeating this reasoning if we start with $y' \leq x'$, and using $(x')' = x$ we conclude that $x \leq y$ iff $y' \leq x'$, which means that $x \mapsto x'$ is a dual isomorphism of the lattice $(B, \vee, \wedge)$ onto itself. In Section 4.2 we saw that a dual isomorphism carries joins to meets and meets to joins. Hence we conclude from this directly that

$$(x \vee y)' = x' \wedge y' \quad \text{and} \quad (x \wedge y)' = x' \vee y',$$

as laws of any Boolean algebra. These are usually called DeMorgan's formulas.

Examples of Boolean algebras.

a) The power set lattice $(\mathcal{P}(X), \subset)$ of all subsets of a set X is obviously a Boolean lattice. It leads to the Boolean algebra $(\mathcal{P}(X), \cup, \cap, ^c, \emptyset, X)$ where A^c denotes the complement of any element A of $\mathcal{P}(X)$, i.e.: A^c consists of the elements of X not in A. In particular if X is finite with say n elements, then $\mathcal{P}(X)$ has 2^n elements. We will see later that every finite Boolean algebra is actually isomorphic with some finite power set lattice and hence has 2^n elements for some n.

Based on what we have discussed so far we might reasonably ask if every subalgebra of the power set algebra $\mathcal{P}(X)$ is a power set algebra of some subset of X. The following example shows that this is not true, and even more shows that not every Boolean algebra is even isomorphic to a power set algebra. First, for any set X, a subset $A \subset X$ is said to be *cofinite* if its complement in X, $X - A$ is finite. Let $Z(X)$ be the set of all finite subsets and all cofinite subsets of X. In particular consider the case where X is, say the set of all integers, a countable set. In any case $Z(X)$ is a subalgebra of the power set algebra. To see this we need to show that $Z(X)$ is closed under the operations. We illustrate with the case A finite and B cofinite. Then $(A \cup B)^c = A^c \cap B^c$ is finite since B is cofinite. The verification is similar for the other cases. Now the power set of a set X is either finite (if X is) or uncountable (is X is infinite). Hence the power set algebra $\mathcal{P}(X)$ is either finite or uncountable. But $Z(X)$is countable if X is, since the collection of finite subsets of a countable set

is countable. For background and details on countable and uncountable sets refer to Appendix A.

The example just given suggests the following definition: A *field of sets* is a collection F of some subsets of a set X, which includes both the empty set $\emptyset$ and the whole set X, and which is closed under unions, intersection, and complementation. Every field of sets is a Boolean algebra obviously, and in fact is a subalgebra of a power set algebra. More generally a *ring of sets* is a collection R of some subsets of a set X which is closed under just unions and intersection. Notice that a ring of sets is certainly a distributive lattice, but for example the set of all finite subsets of the set of all integers is obviously a ring of sets which is not a field of sets.

b) Example a) made the connection between Boolean algebras and the algebra of sets. Now let us turn to the connection with propositional logic. We have already noted that in every Boolean algebra the two elements 0, and 1 constitute a Boolean algebra. Often we denote these elements as F and T (or $\bot$ and $\top$ (thinking of "false" and "true" and construe $\vee$ to denote logical "or" (disjunction), $\wedge$ to denote logical "and" (conjunction), and $'$ to denote logical "negation" (it is false that...). Thus the operations are defined on the elements by

$$F \vee F = F \wedge F = F \wedge T = T' = F,$$

$$T \wedge T = F \vee T = T \vee T = F' = T.$$

Now let us generalize this example and obtain what is called the *Lindenbaum* or *Lindenbaum-Tarski* algebra. We start with the set of all "terms" of propositional logic. These are similar to the set of all lattice terms except we include the operations $'$, 0, and 1. Propositional terms (also often called propositional *formulas* are thus defined inductively to be any of

a) the constants 0 and 1 or

b) the variables $p_1, p_2, \ldots$, (propositional variables) or

c) $(P \vee Q)$, $(P \wedge Q)$, or (P'), where P and Q are propositional terms.

Now we want to make the set F of all propositional formulas a Boolean algebra with operations $\vee, \wedge$, and $'$. Notice that as it stands F is not a Boolean algebra; for example what are the unique 0 and 1 elements? We might reasonably want the 0 to be a false formula and the 1 to be a true formula, the trouble is that there are many false formulas and many true formulas. For this reason we partition the set F into equivalence classes, where two formulas P and Q are put in the same equivalence class just in case P and Q always have the same truth values. To be precise suppose each of P and Q are constructed using variables from among $p_1, \ldots, p_n$. Then by the familiar truth table method, for each assignment of True of False sentences for $p_1, \ldots, p_n$, each of P and Q has a certain truth value T of F. Now we define the binary relation θ on the set F by

$$P\theta Q \iff P \text{ and } Q \text{ have the same truth value for each such assignment.}$$

Thus for example the DeMorgan formula for propositional logic tells us that the formulas $(p_1 \vee p_2)'$ and $p_1' \wedge p_2'$ are related by θ. In particular all formulas which are True for all such assignments fall in one class which we denote by 1 (often called the class of *tautologies*) and all formulas which are False under all assignments fall in a single class denoted by 0. Obviously θ is an equivalence relation on F (check reflexivity, symmetry, and transitivity), and in general if P is a formula we denote, as usual, the equivalence class containing P by P/θ. Thus $P/\theta = Q/\theta$ iff $P\theta Q$. Denote by F/θ the collection of all θ-equivalence classes.

Further we can easily verify that the system of equivalence classes has the "substitution property" with respect to the operations $\vee, \wedge$, and $'$. For example, the substitution property for $\vee$ is the property that for any formulas P, P_1, Q, Q_1,

$$P\theta Q \text{ and } P_1\theta Q_1 \Rightarrow (P \vee P_1) \vee (Q \vee Q_1).$$

(Cf Section 2.6 where we did this for the equivalence relation of "congruence modulo m on $\mathbb{Z}$ for addition and multiplication of integers). We leave the actual verification of the substitution property to the reader. Having done this we can then unambiguously define the operations $\vee, \wedge$, and $'$ on the θ-equivalence classes and thus describe the Lindenbaum algebra as the algebra $(F/\theta, \vee, \wedge, ', 0, 1)$ where for $P/\theta, Q/\theta \in F/\theta$,

$$P/\theta \vee Q/\theta = (P \vee Q)/\theta, \quad P/\theta \wedge Q/\theta = (P \wedge Q)/\theta, \quad (P/\theta)' = P'/\theta.$$

It is easy to check that the Lindenbaum algebra is a Boolean algebra. As one might expect it plays an important role in the theory of propositional logic. Incidentally, since it is easy to see that the Lindenbaum algebra is countable, it is not isomorphic to a power set algebra.

Exercises Section 4.5

1. Prove that if L is a distributive lattice with 0 and 1, then the complemented elements of L form a sublattice (which is a Boolean lattice).

2. If a positive integer is squarefree show that the set of all of its divisors forms a Boolean sublattice of $(\mathbb{N}, |)$.

3. Show that the following hold in all Boolean algebras:

(a) $(a \wedge b) \vee (a' \wedge b) \vee (a \wedge b') \vee (a' \wedge b') = 1$,

(b) $(a \wedge b) \vee (a' \wedge c) = (a \wedge b) \vee (a' \wedge c) \vee (b \wedge c)$,

(c) $a = b \iff (a \wedge b') \vee (a' \wedge b) = 0$,

(d) $a \wedge b \leq c \vee d \iff a \wedge c' \leq b' \vee d$,

(e) $x \leq y \iff x \wedge y' = 0$.

4. Show that the class of Boolean algebras can be defined "equivalently" as systems $(B, \vee, ')$ where $\vee$ and $'$ have certain properties of lattice join and Boolean complementation. Describe the properties which $\vee$ and $'$ should have and explain the sense in which the resulting systems are equivalent to the Boolean algebras defined in the text.

5. In any Boolean algebra define a new operation $\rightarrow$ called *implication* by $x \rightarrow y = x' \vee y$ (thinking of the meaning of implication in propositional logic). Show that in any Boolean algebra $x \leq y \Leftrightarrow x \rightarrow y = 1$.

6. Recall from Section 4.2 that an atom of a lattice is a cover of a least element. Hence an atom of a Boolean algebra is an element which covers 0. A Boolean algebra is *atomic* if for each no-zero element x of the algebra there is an atom a of the algebra such that $a \leq x$, i.e.: each non-zero element "contains" an atom.

a) Prove that in a Boolean algebra any non-zero element is an atom iff it is join irreducible.(The "only if" direction was established in Exercise 4 of Section 4.4.)

b) Describe the atoms in the Boolean algebra of all finite and cofinite subsets of a countable set X and show that this algebra is atomic.

c) Show that the Lindenbaum algebra contains no atoms, i.e.: is *atomless*. (Hint: Think of propositional logic; if P is a propositional formula a p_i is a propositional variable not occurring in P, then $P \wedge p_i \rightarrow P$ is a tautology but its converse $P \rightarrow P \wedge p_i$ is not.

d) Prove that every finite Boolean algebra is atomic.

4.6 The representation of Boolean algebras

In this section we shall explicitly show that every finite Boolean algebra is isomorphic with a field of sets and that every finite distributive lattice is isomorphic to a ring of sets and then indicate what is involved in showing that the same result is true even if we drop finiteness. The details of the general case would take more space than we have here. Suffice it to say, however, that the general result—the representation theorem for Boolean algebras and distributive lattices (proven in 1930 by Marshall Stone), is one of the seminal results of twentieth century mathematics. Because of this result the theory of distributive lattices (and thus of Boolean algebras) stand quite apart from the more general theory of modular lattices. For modular lattices there is no representation theorem resembling the distributive case.

The main pieces of the proof of the finite case for distributive lattices can already be found in the exercises of the last two sections. We shall however first present the proof for finite Boolean algebras here independently, since the ideas are so important and perhaps more suggestive of how to approach the general representation problem. Let B be a finite Boolean algebra B.

1) If x is any non-zero element of B then for some atom a of B, $a \leq x$. For if x is not already an atom then for some $y \in B$, $0 < y < x$. Then if y is not an atom there

must be a $z \in B$ such that $0 < z < y$. Continuing in this way we obtain a chain in B: $0 < \cdots < z < y < x$, which by the finiteness of B must eventually terminate in an atom.

2) Let X be the (finite) set of all atoms of B. For any element x of B let $A(x)$ be the set of all atoms $a \leq x$. From 1) above, notice that $A(x) = \emptyset$ iff $x = 0$ while $A(1) = X$. We will show that B is isomorphic to the power set algebra $\mathcal{P}(X)$, via the correspondence $x \mapsto A(x)$.

3) Obviously $x \mapsto A(x)$ is a well defined function from B into $\mathcal{P}(X)$. We show that it is 1-1. Assume $x \neq y$. Then either $x \not\leq y$ or $y \not\leq x$, say $x \not\leq y$. Hence (by Exercise 3, Section 4.5) $x \wedge y' \neq 0$ and therefore there is an atom $a \leq x \wedge y'$. Then $a \leq x$ so $a \in A(x)$. Also, however, $a \leq y'$ so $a \leq y$ would imply that $a \leq y \wedge y' = 0$, contradicting the fact that a is an atom. Hence $a \notin A(y)$ and hence $A(x) \neq A(y)$.

4) Now we show that $x \mapsto A(x)$ is onto $\mathcal{P}(X)$. Let C be any subset of the atoms of B. Since $0 \mapsto \emptyset$ we may assume $C = \{a_1, \ldots, a_k\}$ for some $k > 0$. Let $x = a_1 \vee \cdots \vee a_k$. We claim that $A(x) = C$. First notice that each $a_i \leq x$ so certainly $C \subset A(x)$. Next, if $a \in A(x)$ then $a \leq x = a_1 \vee \cdots \vee a_k$ so

$$\begin{aligned} a = a \wedge x &= a \wedge (a_1 \vee \cdots \vee a_k) \\ &= (a \wedge a_1) \vee \cdots \vee (a \wedge a_k). \end{aligned}$$

Now if a and a_i are distinct atoms then $a \wedge a_i = 0$. Hence the equation above implies that $a = a_i$ for some i. Therefore $A(x) \subset C$. Hence $x \mapsto A(x)$ is a 1-1 correspondence between B and $\mathcal{P}(X)$.

This already proves that if B has n atoms (so $|X| = n$), then $|\mathcal{P}(X)| = 2^n$ and hence $|B| = 2^n$.

5) Finally we must see that $x \mapsto A(x)$ is an isomorphism of the given Boolean algebra B and the power set Boolean algebra $\mathcal{P}(X)$. Since we have just seen that $A(x)$ is the set of all atoms $\leq x$ and x is the join of all such atoms it follows immediately that

$$x \leq y \Leftrightarrow A(x) \subset A(y)$$

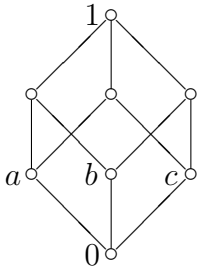
which proves that $x \mapsto A(x)$ is a lattice isomorphism. Last, we need to see that the mapping preserves complements, i.e.: that $A(x)^c = A(x')$. But on the one hand, if a is an atom and $a \leq x'$ then $a \not\leq x$, and this means that $a \in A(x)^c$. On the other hand, if a is an atom and $a \notin A(x)$ then $a \not\leq x$, so $a \neq a \wedge x$. But $a = a \wedge (x \vee x') = (a \wedge x) \vee (a \wedge x')$ and since a is an atom it must therefore equal one of $a \wedge x$ and $a \wedge x'$, hence it must equal $a \wedge x'$. Thus $a \in A(x')$.

The steps in 1) through 5) above establish the following theorem.

Theorem 4.6.1 *Every finite Boolean algebra B is isomorphic to a field of sets (the power set algebra of all subsets of the set of atoms of B).*

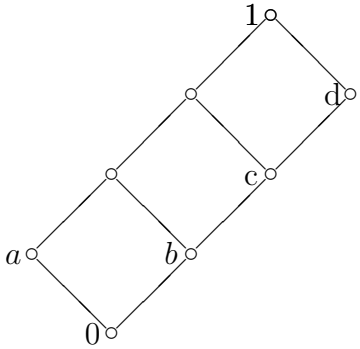
A simple consequence of this theorem is that two finite Boolean algebras are isomorphic iff they have the same number of elements.

The following diagram describes the 8 element Boolean algebra with atoms a, b, c (of dimension 1). The elements of dimension 2 are $a \vee b, a \vee c$, and $b \vee c$. The element $1 = a \vee b \vee c$.



As indicated earlier the theorem remains true for arbitrary non-finite Boolean algebras though, as we have seen, the isomorphic field of sets is no longer a full power set algebra.

Now let us turn to the corresponding representation theorem for finite distributive lattices. As the following diagram for an 8 element distributive lattice illustrates, since the 8 element Boolean algebra has the same diagram as a certain 8 element distributive lattice, the order of a distributive lattice no longer uniquely determines it up to isomorphism (as it did in the Boolean case). Also in the diagram below the only atoms are the elements a and b.



To obtain a ring of subsets of a set X which isomorphic to a given distributive lattice D we must therefore consider a larger collection than just the atoms. The answer is to take the set X to be the set of join irreducible elements in D. In the lattice pictured above the join irreducible elements are $0, a, b, c$ and d. Recall, by the way, from Exercise 5 of Section 4.4, that in any lattice every atom is join irreducible, and from Exercise 6 of Section 4.5, that for Boolean algebras the two concepts coincide for non-zero elements.

Now let us establish the representation theorem for a finite distributive lattices D.

1) Every element $x \in D$ is a join of join irreducibles. The proof of this is by induction on the dimension $d(x)$ of the element x. For $d(x) = 0$ we have $x = 0$ which is certainly join irreducible. For the induction step assume $d(x) > 0$ and that for all y with $d(y) < d(x)$, y is the join of join irreducibles. If x is already join irreducible we are done; otherwise $x = y \vee z$ where both $y < x$ and $z < x$ so both $d(y), d(z) < d(x)$ and hence each of y and z is the join of join irreducibles. Therefore x is also the join of join irreducibles. (We are doing Exercise 6 of Section 4.4 and also essentially repeating the argument (Theorem 2.1.1,a)) establishing the prime factorization of the integers.)

2) For each $x \in D$ let

$$J(x) = \{b : \ b \text{ is join irreducible} \leq x\}.$$

Since $x = b_1 \vee \cdots \vee b_k$, $k > 0$, where the b_i are join irreducible, certainly each $b_i \leq x$. On the other hand, if b is join irreducible and $b \leq x$ then we have

$$b = b \wedge x = b \wedge (b_1 \vee \cdots \vee b_k) = (b \wedge b_1) \vee \cdots \vee (b \wedge b_k)$$

which implies that $b \leq b_i \leq x$ for some i and hence $b \in J(x)$. Hence for any $x \in D$, x is the join of the members of $J(x)$. Hence $J(x) = J(y)$ implies $x = y$ so the mapping $x \mapsto J(x)$ is a 1-1 function of B into $\mathcal{P}(X)$. Notice that $1 \mapsto J(1) = X$ while $0 \mapsto J(0) = \{0\}$.

3) Denote the range of the mapping $x \mapsto J(x)$ by R, a subset of $\mathcal{P}(X)$. Then the mapping is a 1-1 correspondence between D and the poset $(R, \subset)$. We need to see that R is a ring of sets isomorphic to D. (Since it is immediate from the definition of $J(x)$ that $x \leq y \Leftrightarrow J(x) \subset J(y)$, we know that $(R, \subset)$ is a lattice isomorphic with D, but this is not enough. Why?) But if b is a join irreducible and $b \in J(x) \cup J(y)$, then $b \leq x$ or $b \leq y$ so $b \leq x \vee y$ and hence $b \in J(x \vee y)$. Conversely, if $b \in J(x \vee y)$ then $b \leq x \vee y$ from which we conclude as before that $b \leq x$ or $\leq y$ so $b \in J(x) \cup J(y)$. Hence in the correspondence $x \vee y \mapsto J(x) \cup J(y)$. A similar argument shows that $x \wedge y \mapsto J(x) \cap J(y)$. Hence R is closed under both $\cup$ and $\cap$ so it is a ring of sets and is isomorphic to D.

Steps 1) through 3) prove the following theorem.

Theorem 4.6.2 *Every finite distributive lattice is isomorphic to a ring of sets (a sub-ring of the power set algebra of all subsets of join irreducibles of D).*

In the lattice of the diagram above, the lattice is isomorphic to a ring of subsets of $X = \{0, a, b, c, d\}$. The isomorphism corresponds 0 to $\{0\}$, a to $\{0, a\}$, b to $\{0, b\}$, $a \vee b$ to $\{0, a, b\}$, etc.

How can these theorems be extended to their infinite versions? Lets focus on Theorem 4.6.1 for Boolean algebras; the situation for distributive lattices is quite similar. First we should observe precisely where the barrier to extension lies. If B

is an atomic Boolean algebra, then it is clear that part 3) of the proof of Theorem 4.6.1 is still valid so that even in the infinite case it is still true that $x \mapsto A(x)$ is a 1-1 correspondence of B *into* $\mathcal{P}(X)$. Then it is also not hard to show that this correspondence preserves joins, meets, and complementation so that we have the following theorem (the details of the proof are left as an exercise):

Theorem 4.6.3 *If B is an atomic Boolean algebra and X is the set of all atoms of B, then $x \mapsto A(x)$ is an isomorphism of B into the power set algebra $\mathcal{P}(X)$.*

Since we have seen that there are examples of atomless Boolean algebras (e.g.: the Lindenbaum algebra, see Exercise 6, Section 4.5) to prove a general representation theorem we apparently need to find some set X of entities which can play the role of the atoms, even when atoms don't exist. The answer is to construct "ideal" objects which will do the job for us and the idea behind this is very similar to the construction of the "ideal" or complex number i when we needed to, for example, find a root of a polynomial like x^2+1 when none was available in the real numbers. The ideal objects most convenient for our present situation are called *filters* and are defined as follows:

Definition A *filter* of a lattice L is a subset $F \subset L$ with the properties
a) $a, b \in F$ implies that $a \wedge b \in F$, and
b) $a \in L$, $b \in F$ and $a \geq b$ imply that $a \in F$.

The simplest example of a filter in any lattice is a so called *principal* filter (a filter determined by a single element): for any element $f \in L$ let

$$F = \{x \in L : x \geq f\}.$$

It is easy to see that this set is always a filter. Now suppose that a is an atom in a Boolean algebra and consider the filter $F = \{a \in B : x \geq a\}$. This filter has the property that it is *maximal*, meaning that the only filter which contains it is the entire Boolean algebra. To see this suppose there is a filter $F' \supset F$ and that $y \in F' - F$. Then $a \not\leq y$ so $a \wedge y' \neq 0$. But $a \wedge y' \leq a$ and a is an atom. Hence we must have $a = a \wedge y'$ so $a \leq y'$ and hence $y' \in F$. Therefore both y and y' are in F' and hence so is $0 = y \wedge y'$ and this means that $F' = B$. This illustrates how the atoms of a Boolean algebra are very much like the primes in $\mathbb{Z}$; in particular recall from the end of Section 2.4 of Chapter 2 that the lattice $(\mathbb{N}, |)$ is dually isomorphic to the lattice $(\mathcal{M}, \subset)$ of all modules, and hence a positive integer p is a prime iff the module (p) is maximal. Indeed the modules of $\mathbb{Z}$ are very much like filters and are always principal, i.e.: of the form (m) for some integer m.

The important thing for our representation problem is that under the appropriate assumptions about the foundations of mathematics it is possible to show that in any distributive lattice (and hence Boolean algebra), there are many maximal filters which are *not* principal and hence certainly not of the form $\{x : x \geq a\}$ where a

is an atom, but that nonetheless, each non-zero element is contained in a maximal filter and furthermore, for each pair of distinct elements x and y in a distributive lattice it turns out that one is contained in at least one maximal filter which does not contain the other. This implies that if we let X be the set of all maximal filters of the distributive lattice and $A(x)$ be the set of all maximal filters containing x, then the mapping $x \mapsto A(x)$ is obviously 1-1. From here essentially the same proof as in Theorems 4.6.1 and 4.6.2 shows that this mapping is an isomorphism of either a field or a ring of subsets of X.

Finally, what is this "appropriate" assumption about the foundations of mathematics which enables us to find the maximal filters we need? It is an axiom of set theory which is very close to, but slightly weaker than, the famous "Axiom of Choice". To say much more than this in this text would take us far from our central purpose. For more details consult E. Mendelson, *Introduction to Mathematical Logic*, 2nd ed., Van Nostrand, New York. Incidentally, maximal filters are usually called *ultrafilters*. The principle of duality correctly suggests that instead of using filters we could instead use dual objects called which are called *ideals*. See the following exercises for some details.

Exercises Section 4.6

1. With reference to step 3) of the proof of Theorem 4.6.2 prove that the mapping $x \mapsto J(x)$ has the property $x \wedge y \mapsto J(x) \wedge J(y)$.

2. Complete the proof of Theorem 4.6.3.

3. Diagrams of two 8 element distributive lattices are presented in the preceding section. Draw the diagrams of another two 8 element distributive lattices (including the 8 element chain) and describe the representation of each as a ring of sets.

4. a) Show that in a Boolean algebra $a \leq b$ iff $b' \leq a'$.

b) In any Boolean algebra show that a complement of an atom is a coatom.

5. In the definition of a filter show that the condition b) $a \in L$, $b \in F$ and $a \geq b$ imply that $a \in F$, can be replaced by the condition b') $a \in L$, $b \in F$ implies $a \vee b \in F$.

6. a) The dual of a filter is called an *ideal*. Dualize the definitions of a filter given in the text and in Exercise 5, to obtain the definition of an ideal.

b) Show that if a is an element of a lattice L, then $\{x \in L : x \leq a\}$ is an ideal of L.

c) Show that if a is an atom of a Boolean algebra B, then $\{x \in B : x \leq a'\}$ is a maximal ideal of B.

d) If a is an atom of a Boolean algebra, $F = \{x : x \geq a\}$, and $J = \{x : x \leq a'\}$, show that $F \cap J = \emptyset$ and $F \cup J = B$.

7. a) Suppose $(L', \vee', \wedge')$ and $(L'', \vee'', \wedge'')$ are both lattices. Let $L' \times L''$ be the cartesian product of L' and L'', i.e.: the set of ordered pairs (a', a'') where $a' \in L'$ and $a'' \in L''$. Show that $(L' \times L'', \vee, \wedge)$ is a lattice if we define $\vee$ and $\wedge$ by

$$(a', a'') \vee (b', b'') = (a' \vee' b', a'' \vee'' a'') \text{ and } (a', a'') \wedge (b', b'') = (a' \wedge' b', a'' \wedge'' b'').$$

This lattice is called the *direct product* of L' and L''. Briefly speaking, we say that the operations on the direct product are defined "componentwise" on the factor lattices.

b) Show that the direct product of two lattices is distributive (respectively modular) iff each of the factors is distributive (respectively modular).

c) Show that the direct product of two Boolean algebras, where complementation, as well as meet and join, is defined componentwise, is again a Boolean algebra.

d) Let B be the 2 element Boolean algebra with elements 0 and 1. Draw the diagrams of both $B \times B$ and $B \times B \times B$.

8. Let the set X have n elements and for convenience denote its elements by the integers $1, 2, \ldots, n$. For any subset $S \subset X$ let $(s_1, \ldots, s_n)$ be the n-uple obtained by taking each s_i to be either 1 or 0, 1 if $i \in S$ and 0 if $i \notin S$. Show that the mapping $S \mapsto (s_1, \ldots, s_n)$ is an isomorphism of the power set algebra $\mathcal{P}(X)$ and the direct product $B \times \cdots \times B$ (n factors) where B is the two element Boolean algebra. (Therefore every Boolean finite Boolean algebra with n atoms is isomorphic to the direct product $B \times \cdots \times B$ (n factors.)

Chapter 5

Finite State Machines

5.1 Machines-introduction

In this chapter we turn to an important contemporary area of discrete mathematics, the theory of computers. Computer science has two major components: first the fundamental ideas and mathematical models underlying computing, and second, engineering techniques for the design and implementation of computing systems, both hardware and software. In this chapter we are going to introduce the first component and examine one of the basic mathematical models of a computer, a *finite state machine* also known as a *finite state automaton*. This model of computation is not the most general model in that its finiteness does not allow for the possibility of computations or storage of arbitrarily large size, and hence it does not capture the essence of what we think of as a general purpose computer. Nonetheless it is the most realistic model of any real computer in that every real world computer, as a practical matter, is finite in the same sense as our model.

A finite state machine is a mathematical model of a system with discrete inputs and outputs. The system may be in any one of a finite number of internal configurations or *states*. The current state of the system, in effect, amounts to a summary of the information concerning the past inputs that is needed to determine the behavior of the system on subsequent inputs. The behavior of the machine on subsequent input consists of two aspects: first the machine changes into a new internal state, and second it produces output. In the model we shall present, the output will be restricted to one of two possible reports which we designate as Acceptance and Rejection. In fact we shall accomplish this by simply designating some of the internal states of the machine as accepting states and the others as rejecting states. This simplification of the output turns out to be not a very great restriction in terms of the theory we shall develop in this chapter. The reason for this will become clearer as we proceed.

In reading its input any given machine is construed to understand characters which come from some fixed finite alphabet. This is just like any ordinary computing system. The restriction here, as is usual, is that the alphabet have at least two

distinct characters. The simplest alphabet then has just two characters, say 0 and 1. A more common alphabet would have the 26 letters of the English alphabet together with the numeric characters $0, 1, \ldots, 9$. In any case if Σ denotes the finite alphabet, then an input to the machine consists of a character string or a *word* $x = \sigma_1 \cdots \sigma_n$ consisting of $\sigma_i \in \Sigma$. Sometimes it is useful to think of a word physically as a piece of tape with the characters printed on it. The machine then reads the word or tape, say from left to right, starting in some initial state, and moves from state to state as it reads each character. After reading the last character in the word if the machine is in one of its accepting states we say that the machine has accepted the word x, otherwise the machine rejects the word. The reading of the word is accomplished by some sort of "read head" and the internal states and the mechanism (probably solid state electronics) is embodied in what is often called the "finite control". From the standpoint of our mathematical model, the operation of these components is unimportant and simply takes place in a black box. Figure 5.1 schematically describes the action of the machine.

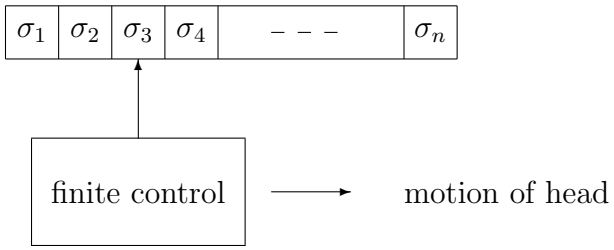


Figure 5.1

5.2 Semigroups and monoids

The main results of this chapter can be most profitably phrased in terms of concepts related to algebraic systems called semigroups. Thus we take a little space to discuss their basic properties.

Definitions a) A *semigroup* is a system $(S, \cdot)$ where S is a nonempty set and $\cdot$ is a binary operation (called multiplication). If x and y are elements of S their product $x \cdot y$ (which we will usually write as xy) is supposed to again be in S. Also the product is supposed to be associative. Thus if $x, y, z \in S$, then $(xy)z = x(yz)$. As a consequence of associativity any un-parenthesized finite product $xy \cdots z$ is well defined, since it can be re-parenthesized in any way without changing its value in the semigroup. A special case of this consequence of associativity is that we can define exponents: for any positive integer n we define

$$x^n = x \cdots x \quad (\text{ the product of } n \text{ } x\text{'s}).$$

b) A *monoid* is a semigroup which contains an identity element $e \in S$ with the defining property $ex = xe = x$ for all $x \in S$. If both e and e' are identity elements of a monoid then the definition yields $e = e'e = e'$ so the identity is unique. In a monoid we can consequently define the zero-th power of an element as $x^0 = e$. With our definition of exponentiation observe that the usual laws of exponents

$$x^{m+n} = x^m x^n, \quad (x^m)^n = x^{mn},$$

hold for any nonnegative integers m and n.

Examples.

1) Let S consist of the positive integers greater than 1 and let $\cdot$ be ordinary integer multiplication. Then $(S, \cdot)$ is a semigroup. If we replace S by $\mathbb{Z}^+$, the system $(\mathbb{Z}^+, \cdot)$ is monoid with 1 as an identity. On the other hand, the system $(\mathbb{Z}^+, +)$ where $+$ is integer addition, is a semigroup which is not a monoid, while $(\mathbb{N}, +)$ is a monoid with 0 as an identity element. For each of these examples the operation is commutative: $x \cdot y = y \cdot x$ in $(\mathbb{Z}^+ . \cdot)$ and $x + y = y + x$ in $(\mathbb{N}, +)$.

2) For some fixed integer n, the set of all $n \times n$ real matrices with ordinary matrix multiplication is a (noncommutative) monoid with the $n \times n$ identity matrix as an identity element.

3) Let Σ be any finite alphabet and let Σ^* be the set of all words over Σ, i.e.: the set of all finite character strings. This includes the empty word which we denote by ϵ. Let $\cdot$ be the operation of concatenation of words or, if we think of words as tapes, then concatenation becomes the operation of splicing two tapes together. If x is any word then concatenating it with the empty word in either order obviously gives $\epsilon x = x\epsilon = x$ so the system $(\Sigma^*, \cdot)$ is a monoid with identity element ϵ. Notice that concatenation is a noncommutative operation.

Now let θ be an equivalence relation on the elements S of a semigroup S. We say that θ has the *substitution property* (with respect to the semigroup multiplication) if for all $x, x', y, y' \in S$,

$$x\theta x' \text{ and } y\theta y' \Rightarrow xy\theta x'y'.$$

As we have in the past (in the integers and in Boolean algebras) we call an equivalence relation with the substitution property a *congruence relation.* When θ is a congruence relation we call the θ equivalence classes congruence classes. If an equivalence relation has the substitution property then, (just as in the case of the integers modulo m) this allows us to unambiguously define the semigroup operation among the congruence classes. Thus if the θ class containing the element x is as usual denoted by x/θ, defined by

$$x/\theta = \{x' \in S : x'\theta x\} \text{ (and hence } x/\theta = x'/\theta \Leftrightarrow x\theta x' \text{),}$$

then multiplication of congruence classes is defined by

$$x/\theta \cdot y/\theta = (x \cdot y)/\theta$$

and it is precisely the substitution property which justifies the definition. The set of all θ congruence classes is denoted by S/θ and the system $(S/\theta, \cdot)$ is the *quotient semigroup* called "S mod θ".

We shall also be interested in two notions that are weaker than the substitution property:

Definition An equivalence relation θ on a semigroup S called *right stable* (respectively *left stable*) if for all $x, y, z \in S$

$$x\theta y \Rightarrow xz\theta yz \quad \text{(respectively } \quad zx\theta zy \text{)}.$$

Lemma 5.2.1 *An equivalence relation is a congruence relation iff it is both left and right stable.*

PROOF If θ is a congruence relation and $x\theta y$ then for any z, $z\theta z$ so $zx\theta zy$ and $xz\theta yz$. Conversely, if θ is both left and right stable and $x\theta x'$ and $y\theta y'$, then $xy\theta x'y$ by right stability and $x'y\theta x'y'$ by left stability. Hence $xy\theta x'y'$ by transitivity.

Exercises Section 5.2

1. For the monoid $(\mathbb{N}, +)$ of Example 1 above, find a congruence relation of index 3 (i.e.: which has 3 congruence classes) and which is not congruence modulo 3.

2. For the monoid of all $n \times n$ real matrices under multiplication, of Example 2 above find
   a) an example of a congruence relation;
   b) describe the quotient monoid determined by your example from a);
   c) an example of a right stable equivalence relation which is not a congruence relation.
3. For the monoid $(\Sigma^*, \cdot)$ of Example 3 above find
   a) an example of a congruence relation of finite index,
   b) an example of a congruence relation of infinite index,
   c) an example of a right invariant equivalent relation of finite index which is not a congruence relation.

5.3 Machines - formal theory

Now let us formalize the idea of a finite state machine discussed in Section 5.1. As indicated there we visualize a finite state machine as a "black box" M which reads a

word $x = \sigma_1 \cdots \sigma_n$, chosen from some fixed alphabet Σ, symbol by symbol from left to right. The machine starts reading in some designated "initial" state and enters a new state after reading a character of the word solely on the basis of the current state and the character just read. M either accepts or rejects the word; acceptance occurs just in case the state the machine enters after reading the last character σ_n is one of the designated accepting states and rejection occurs otherwise. This leads to the following mathematical model:

Definition A finite state machine (or automaton) over the alphabet Σ is a system

$$M = (S, m, s_0, F)$$

where

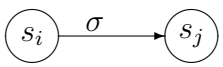
S is a finite set, the set of *states*;
m is a function of two variables, the *state transition function* $m : S \times \Sigma \to S$; if $s \in S$ is a state and $\sigma \in \Sigma$ is a character, then $m(s, \sigma)$ is the "next" state;
$s_0 \in S$ is the initial state;
$F \subset S$ is the set of accepting states.

It is convenient to extend the function m from $S \times \Sigma \to S$ to $S \times \Sigma^* \to S$ by the following inductive definition:

$m(s, \epsilon) = s$ for all $s \in S$,
$m(s, x\sigma) = m(m(s, x), \sigma)$ for all $x \in \Sigma^*$ and $\sigma \in \Sigma$.

This definition means first that if M is in state s and reads the empty word it remains in state s, and in general, M starts in state s and reads through the symbols of a word from left to right, changing states as prescribed by the function m; the value $m(s, x)$ is the last state reached. From the definition it follows (by induction of course) that for any words $x, y \in \Sigma^*$, $m(s, xy) = m(m(s, x), y)$. (Exercise 8 below.)

While a finite state machine (S, m, s_0, F) over Σ is completely described by a table giving the values of the transition function m, this description can be a little opaque; it is much easier to grasp the nature of the machine by drawing a certain directed graph called a "state diagram". To do this we first draw disjoint circles representing the states in S, labeling each by exactly one of the states. Label the accepting states in F, say by indicating them by double circles. Then connect the circle labeled s_i to the circle labeled s_j by an arrow labeled σ thus,



just in case the definition of m prescribes that $m(s_i, \sigma) = s_j$.

As an example let the alphabet $\Sigma = \{0, 1\}$ and let $S = \{s_0, s_1, s_2, s_3\}$ be the states of the machine $M = (S, m, s_0, \{s_0\})$ (i.e.: the only accepting state consists of just the

initial state). Figure 5.2 describes the transition function m and the corresponding state diagram.

m	0	1
s_0	s_2	s_1
s_1	s_3	s_0
s_2	s_0	s_3
s_3	s_1	s_2

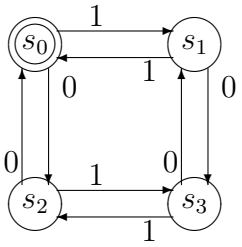


Figure 5.2

The machine M accepts a word $x \in \Sigma^*$ (a string of 0's and 1's) just in case the sequence of transitions corresponding to the symbols of x leads from s_0 back to s_0. How can we simply characterize the accepted words? We claim M accepts precisely those words which contain both an even number of 0's and of 1's. To make this clear visualize control as traveling from state to state in the diagram. Control starts at s_0 and must finish at s_0 if the input word is to be accepted. In the diagram partitioned by dotted lines of Figure 5.3 we see that each 0 input causes control to cross the horizontal partitioning line while a 1 input causes control to cross the vertical partitioning line. Thus control is above the horizontal line iff the number of 0's read so far is even and is to the left of the vertical line iff the number of 1's read so far is even. Thus control is at s_0 iff the number of 0's and 1's read so far are both even.

Incidentally, this partitioning also tells us, for example, that if instead of labeling s_0 as the only accepting state we had also included s_1, then the resulting machine would accept precisely those words containing an even number of 0's and ignore the number of 1's. Indeed in that case we could simplify the machine by combining s_0 and s_1 into a single state, call it q_0, and likewise s_2 and s_3 into a single state, call it q_1. Then the transition function m for the new two state machine would be defined by

$$m(q_0, 0) = q_1, \;\; m(q_1, 0) = q_0, \;\; m(q_0, 1) = q_0, \;\; m(q_1, 1) = q_1.$$

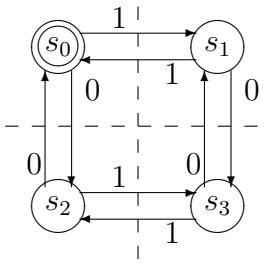


Figure 5.3

Now let Σ be some fixed finite alphabet. For convenience let us call a subset L of the set Σ^* of all words over Σ a *language* over Σ. If $M = (S, m, s_0, F)$ is a finite state machine let $L(M)$ denote the language of words in Σ^* accepted by M, that is

$$L(M) = \{x \in \Sigma^* : m(s_0, x) \in F\}.$$

(Sometimes we say that the language $L(M)$ is *defined* by M.) Finally a language $L \subset \Sigma^*$ is *regular* if there exists a finite state machine M such that $L = L(M)$. Often a regular language is simply called a *regular set*. The preceding example suggests that a particular regular language may generally be definable by more than one finite state machine. We shall see later that each regular language is in fact definable by infinitely many different machines, thus there is a many-one correspondence between machines and regular languages.

Our principal concern for the remainder of this chapter will be to give some simple mathematical characterizations of the class of all regular languages. Before we start it is worth noticing that not every language over Σ is regular. To see this recall that the set Σ^* of all *finite* words over an alphabet Σ is countable and the collection of all languages over Σ, i.e.: subsets of Σ^*, is therefore uncountable. On the other hand it is easy to see that the number of possible finite state machines is countable. Hence there are in fact uncountably many languages which are not regular.

Exercises Section 5.3

1. In the machine described by the diagram in Figure 5.2 suppose we changed the accepting state to consist of only s_3. Describe the language accepted by this machine. Can you find a machine with fewer than four states which will accept this language? Explain your answer.

2. Let $\Sigma = \{0, 1\}$, $S = \{s_0, \ldots, s_4\}$ and let the transition function for the machine $M = (S, m, s_0, \{s_4\})$ be given by the table

m	0	1
s_0	s_2	s_1
s_1	s_3	s_2
s_2	s_4	s_2
s_3	s_3	s_3
s_4	s_0	s_0

Draw the state diagram for M and give a simple description of $L(M)$.

3. Let $\Sigma = \{0, 1\}$. For each of the following languages L over Σ find a machine M defining L:

a) the set of all words beginning and ending with a 1;
b) the set of all words with the property that each occurrence of a 1 is followed by precisely two 0's,
c) the set of all words of length divisible by 3.

4. If L is a regular set prove that the complement $\Sigma^* - L$ is also regular.

5. Why is the set consisting of the empty word regular?

6. Let L be a nonempty regular language and let $Init(L)$ be the set of all nonempty words x such that for some word y, the word $xy \in L$. Prove that $Init(L)$ is regular.

7. Let $\Sigma = \{0, 1, 2\}$ and let

$$L = \{x \in \Sigma^* : \text{ both the first and last characters in } x \text{ are either 0 or 1}\}.$$

Show that L is regular.

8. Prove that for all $x, y \in \Sigma^*$ and $s \in S$, the extension of m to $S \times \Sigma^* \to S$ has the property $m(s, xy) = m(m(s, x), y)$.

5.4 The theorems of Myhill and Nerode

Our first characterization of the class of regular languages is due to John Myhill and was published in 1957.

Theorem 5.4.1 *(Myhill's Theorem) For a set $L \subset \Sigma^*$ of words the following are equivalent:*
1) $L = L(M)$ for some finite state machine M.
2) For some congruence relation θ on the monoid $(\Sigma^*, \cdot)$, having finite index, L is the union of some θ-congruence classes.
3) The explicit congruence relation ϕ on $(\Sigma^*, \cdot)$ defined by

$$x \phi y \Leftrightarrow (\forall z, w \in \Sigma^*)(zxw \in L \Leftrightarrow zyw \in L)$$

has finite index

Before giving the proof of the theorem let us first observe that independent of anything else, the relation ϕ described in statement 3) is a congruence relation, so that the significant content of 3) is that this ϕ has finite index. To verify this the reader should use the definition of ϕ to check the reflexive, symmetric, and transitive properties for ϕ and then verify that ϕ is both left and right stable so that Lemma 5.2.1 shows that it is a congruence relation.

Next, before giving the proof, let us understand the theorem better by considering a simple application. In particular let us redo Exercise 7 of the last section by showing that the set $L \subset \{0,1,2\}^*$ consisting of all words which both begin and end with either 0 or 1 satisfies both of statements 2) and 3) of the theorem. For 2) we can, for example define θ on $\{0,1,2\}^*$ by $x\theta y$ iff x and y both begin with the same characters and end with the same characters. Its easy to see that θ is a congruence relation of index 9; each congruence class consists of all words with the same initial character 0,1, or 2 and the same final character 0,1, or 2. L then consists of the union of 4 of these classes which shows that 2) is satisfied.

To verify that 3) is satisfied observe that the congruence relation ϕ partitions $\{0,1,2\}^*$ into just 2 classes: L and its complement $\{0,1,2\}^* - L$.

PROOF of Theorem 5.4.1: We begin by proving the implication 1) $\rightarrow$ 2). Thus we are given $M = (S, m, s_0, F)$ defining $L \subset \Sigma^*$. We define θ on Σ^* by

$$x\theta y \Leftrightarrow m(s,x) = m(s,y) \quad \text{for all } s \in S.$$

Since equality is an equivalence relation it easily follows that θ is also. To verify right stability observe that $x\theta y \Rightarrow m(s,x) = m(s,y)$ for all $s \in S \Rightarrow m(s,xz) = m(m(s,x),z) = m(s,y),z) = m(s,yz)$ for all $s \in S \Rightarrow xz\theta yz$. For left stability, observe that $x\theta y \Rightarrow m(s,x) = m(s,y)$ for all $s \in S \Rightarrow m(s,zx) = m(m(s,z),x) = m(m(s,z),y) = m(s,zy)$ for all $s \in S \Rightarrow zx\theta zy$. Hence θ is a congruence relation.

The definition of θ implies that for $x \in \Sigma^*$, the various functions $\phi_x : S \rightarrow S$ each defined for $s \in S$ by $\phi_x(s) = m(s,x)$ are the same over the elements of x/θ and different congruence classes yield different functions. Hence the number of such functions = index θ; but also the total number of functions from S to S is $|S|^{|S|}$, so the index of θ is $\leq |S|^{|S|}$ which is finite since S is finite.

Also $x \in L(M) \Leftrightarrow m(s_0,x) \in F$. Hence if $x\theta y$, then $m(s_0,y) = m(s_0,x) \in F$ so $y \in L(M)$. Thus $L = L(M)$ is the union of θ congruence classes.

Next we prove that 2) $\rightarrow$ 3). Suppose θ is a congruence relation on Σ^* having finite index and L is the union of some θ congruence classes. Suppose $x\theta y$. Then, since θ has the substitution property, for all $z, w \in \Sigma^*$,

$$zxw\theta zyw.$$

Hence $zyw \in L \Leftrightarrow zxw \in L$. Thus $x\theta y$ implies $x\phi y$ for the ϕ defined in statement 3). Hence $\theta \leq \phi$, i.e.: θ is a refinement of ϕ.Therefore index $\phi \leq$ index θ which is finite. This proves statement 3).

To prove 3) $\rightarrow$ 1), assume that the congruence relation ϕ defined in 3) has finite index. We construct a finite state machine M over Σ as follows:

$$M = (S, m, s_0, F)$$

where

$S = \{x/\phi : x \in \Sigma^*\}$; $|S| < \infty$ since ϕ has finite index.

$m : S \times \Sigma \to S$ is defined by $m(x/\phi, \sigma) = (x\sigma)/\phi$.
Note that m is well defined, for $x\phi y \Rightarrow x\sigma\phi y\sigma$ by right stability.

$s_0 = \epsilon/\phi$.

$F = \{x/\phi : x \in L\} \subset S$.

Certainly M, so defined, is a finite state machine; also notice that

i) if $x \in L$ and $x\phi y$ then $y \in L$ by the definition of ϕ;

ii) the extension of m to a function in $S \times \Sigma^* \to S$ yields $m(x/\phi, y) = (xy)/\phi$.

To establish 1) we need to show that $x \in L \Leftrightarrow m(s_0, x) \in F$, and in fact we have

$$\begin{aligned} m(s_0, x) &\in F \\ &\Leftrightarrow m(\epsilon/\phi, x) \in \{x/\phi : x \in L\} \\ &\Leftrightarrow (\epsilon x)/\phi \in \{x/\phi : x \in L\} \\ &\Leftrightarrow x \in L. \end{aligned}$$

This completes the proof of Myhill's theorem.

The problem with Myhill's theorem is that the number of states in the machine constructed in the proof of 3) $\to$ 1) equals the index of ϕ and this could potentially be quite large, and in fact the upper bound observed in the proof 1) $\to$ 2) is

$$\text{index } \phi \leq \text{index } \theta \leq |S|^{|S|}$$

where S is the set of states in the machine we started with in 1). The great advantage of the following theorem of Nerode (1958) is that it will in fact produce a machine which has a *minimum* number of states.

Theorem 5.4.2 *(Nerode's theorem) For a set $L \subset \Sigma^*$ of words the following are equivalent:*

1) $L = L(M)$ for some finite state machine M.

2) For some right stable equivalence relation θ' on the monoid $(\Sigma^*, \cdot)$, having finite index, L is the union of some θ' equivalence classes.

3) The explicit right stable equivalence relation π defined on $(\Sigma^*, \cdot)$ by

$$x\pi y \Leftrightarrow (\forall z \in \Sigma^*)(xz \in L \Leftrightarrow yz \in L)$$

has finite index.

Before we give the proof notice, as in Myhill's theorem, that π is a right stable equivalence relation independent of the truth of statement 3). Also notice that $\phi \leq \pi$ where ϕ is Myhill's congruence, so that index $\pi \leq$ index ϕ.

Proof of Theorem 5.4.2 1) $\to$ 2). Define θ' on Σ^* by

$$x\theta' y \Leftrightarrow m(s_0, x) = m(s_0, y).$$

Clearly θ' is an equivalence relation and, by the same proof as in Myhill's theorem, is right stable (but not left stable).

We can easily see that θ' has finite index in either of two ways:

a) Obviously $\theta \leq \theta'$ so index $\theta' \leq$ index θ which is finite ($\leq |S|^{|S|}$ by the proof given in Myhill's theorem.

b) A better way is to observe that if x/θ' and y/θ' are distinct θ' classes, then it is not true that $x\theta' y$ so $m(s_0, x)$ and $m(s_0, y)$ are distinct states; hence index $\theta' \leq |S|$, which is finite by hypothesis. Notice that this is a much lower upper bound than $|S|^{|S|}$ which is all we could get by proof a).

L is the union of θ' classes as in Myhill's theorem.

To prove 2) $\rightarrow$ 3) suppose L is the union of θ' classes for some right stable θ' of finite index. Suppose $x\theta' y$ and that $xw \in L$ for some w. Then $xw\theta' yw$ by right stability. But L is the union of θ' classes. Hence $yw \in L$. Therefore $x\pi y$ and thus $\theta' \leq \pi$ so that index $\pi \leq$ index θ', which is finite by assumption.

For 3) $\rightarrow$ 1) we construct a finite state machine M_π as in Myhill's theorem except that we use π classes (instead of ϕ classes for the states:

$S_\pi = \{x/\pi : x \in \Sigma^*\}$, $|S_\pi| < \infty$ by assumption;

$m_\pi(x/\pi, \sigma) = (x\sigma)/\pi$. m is well defined since π is right stable;

$s_{0\pi} = \epsilon/\pi$;

$F_\pi = \{x/\pi : x \in L\}$.

The proof of 3) $\rightarrow$ 1) is the same as in Myhill's theorem, taking

$$M_\pi = (S_\pi, m_\pi, s_{0\pi}, F_\pi)$$

as defined above. Specifically we have

$$\begin{aligned} m_\pi(s_{0\pi}, x) &\in F_\pi \\ &\Leftrightarrow m_\pi(\epsilon/\pi, x) \in \{x/\pi : x \in L\} \\ &\Leftrightarrow (\epsilon x)/\pi = x/\pi \in \{x/\pi : x \in L\} \\ &\Leftrightarrow x \in L. \end{aligned}$$

Example. Let $\Sigma = \{0, 1\}$. Then $L = \{0^n 1 0^n : n \in \mathbb{N}\} \subset \Sigma^*$ is not definable, for if L were the union of θ' classes, (θ' right stable of finite index), then by the pigeonhole principle $0^m \theta' 0^n$ for some $m \neq n$. Then by right stability $0^m 1 0^m \theta' 0^n 1 0^m$ so that $0^n 1 0^m \in L$, a contradiction.

Nerode's theorem has an extra important dividend: For any given regular language L, by the well ordering principle, there is a machine M with a minimum number of states which defines L. Now we shall prove that this minimal machine is unique *up to isomorphism* and is in fact the machine M_π constructed in the proof of Nerode's theorem and in this sense is canonical. This is particularly significant when we contrast finite state machines, as we have defined them, with other, more general models of computation. For example the most general mathematical model of a computer is a so called Turing machine. In contrast to the present context, an important theorem of

the theory of Turing machines asserts that for a given language accepted by a Turing machine there is no longer a unique minimal machine defining the language!

The construction of the unique minimal machine proved by Nerode's theorem also solves, in principle at least, the classic "state minimization problem" of electrical engineering.

In order to establish the uniqueness of the minimal machine for a regular language we obviously need a precise definition of isomorphism of two finite state machines over the same alphabet Σ. As a first approximation to a formal definition we can agree that we want there to be a 1-1 correspondence between the sets of states of the two machines with the property that if we rename the states of one by means of this correspondence then it becomes the same machine as the other. This is easy to restate formally. In particular for a machine $M = (S, m, s_0, F)$ to be isomorphic to the machine $M_\pi = (S_\pi, m_\pi, s_{0\pi}, F_\pi)$ we should have a 1-1 correspondence between the set S of states of M and the set S_π of states of M_π, with the properties:

a) $s_0 \mapsto s_{0\pi}$,

b) The states in F correspond to the states in F_π, and

c) for all $s \in S$ if $s \mapsto s_\pi$, then for any $\sigma \in \Sigma$, $m(s, \sigma) \mapsto m_\pi(s_\pi, \sigma)$.

Notice that from the way we extend the transition function from $S \times \Sigma \to S$ to $S \times \Sigma^* \to S$ (and likewise for S_π) it follows easily that c) can be replaced by

c′) for all $s \in S$ if $s \mapsto s_\pi$, then for any $x \in \Sigma^*$, $m(s, x) \mapsto m_\pi(s_\pi, x)$.

With this definition we can now establish the following important result:

Theorem 5.4.3 *A minimum state finite state machine defining a given regular language is unique up to isomorphism and is canonically described by the machine M_π of Nerode's theorem.*

Proof First observe that the proof of 2) $\to$ 3) in Nerode's theorem shows that for *any* finite state machine M defining L, there is a θ' on Σ^* with $\theta \leq \pi$. Hence index $\pi \leq$ index θ'. Also in the proof of 1) $\to$ 2) we showed (method b)) that index $\theta' \leq |S|$, the number of states of M. But the number of states of M_π is index π. Therefore the number of states of M_π is $\leq$ the number of states of any M defining L, i.e.: M_π is a minimal state machine defining L.

Now let M be any other minimum state machine defining L. Then we know that M and M_π have the same number of states. We must establish a 1-1 correspondence between the two sets of states which will yield an isomorphism. To do this observe that for any state s of M there must be an $x \in \Sigma^*$ such that $m(s_0, x) = s$, for otherwise we could delete this "inaccessible" state s, which would contradict the minimality of M. If $m(s_0, x) = m(s_0, y) = s$, then $x\theta' y$ so $x\pi y$ and hence $m(s_{0\pi}, x) = m(s_{0\pi}, y)$. This means that if we start with any state s of M and choose *any* x for which $m(s_0, x) = s$, then

$$s = m(s_0, x) \mapsto m_\pi(s_{0\pi}, x) = s_\pi$$

is a well defined 1-1 function of the states of M onto the states of M_π. Under this correspondence if y is any word, then

$$\begin{aligned} m(s,y) = m(m(s_0,x),y) = m(s_0,xy) &\mapsto \\ m_\pi(s_{0\pi},x) &= m_\pi(m_\pi(s_{0\pi},x),y) = m_\pi(s_\pi,y) \end{aligned}$$

which establishes property c′ in the definition of isomorphism. Finally notice that since $m(s_0,\epsilon) = s_0$ it follows that this correspondence entails $s_0 \mapsto s_{0\pi}$ (property a)), and since both machines define L, $m(s_0,x) \in F \Leftrightarrow m_\pi(s_{0\pi},x) \in F_\pi$ so property b) is satisfied. This proves the theorem.

Appendix: Further closure properties. In Exercise 5 below you are asked to prove that the collection of all regular languages over a given alphabet Σ is a field of sets, which means that the regular languages contain both the empty set and Σ^*, and are closed under the operations of union, intersection and complementation. Now we briefly discuss several other important closure properties of regular sets.

Let L and M be two sets of words from the same alphabet Σ. We define the *product* LM of the two sets by

$$LM = \{xy : x \in L \text{ and } y \in M\},$$

i.e.: a word is in LM if it is the concatenation of a word in L with a word in M. It is clear that this product is an associative operation on sets of words: if L, M, N are sets of words over Σ, then

$$(LM)N = L(MN).$$

This leads to the introduction of finite exponents, where we define

$$L^n = L \cdots L \quad \text{(the product of } n \text{ } L\text{'s,)}$$

with the convention that $L^0 = \{\epsilon\}$.

Finally, if L is a set of words we can define the *closure* L^* of L to be the least set which contains L, has the empty word ϵ as an element, and has the property that if x and y are in L^* then xy is in L^*. Thus if Σ is a finite alphabet, then Σ^* is, as before, the set of all words (including ϵ) over Σ. An equivalent way to define the closure is by the equation

$$L^* = L^0 \cup L^1 \cup L^2 \cup \cdots$$

where this union extends over all finite exponents. For example, if $\Sigma = \{0,1\}$ and $L = \{01,11\}$, then $L^* = \{\epsilon, 01, 11, 0101, 0111, 1101, 1111, 010101, \ldots\}$. In particular notice that we have $\emptyset^* = \{\epsilon\}$.

While the proof would require a little further development of the theory of machines, it is not difficult to show that if L and M are regular sets, then so is their product LM and also if L is regular then so is the closure L^* of L. The importance of these operations on regular sets was discovered by the logician S. C. Kleene who first proved the following charactrization of regular sets.

Theorem 5.4.4 *The class of regular sets over a fixed alphabet is the least class which contains all of the finite sets and which is closed under the operations of union, product and closure.*

For example, suppose $\Sigma = \{0, 1\}$ and

$$L = \{x \in \Sigma^* : x \text{ contains two or three occurances of 1,} \\ \text{the first and second of which are not consecutive}\}.$$

Using union, product, and closure,

$$L = \{0\}^*\{1\}\{0\}^*\{0\}\{1\}\{0\}^*((\{1\}\{0\}^*) \cup \emptyset^*),$$

and it follows from Theorem 5.4.4 the L is a regular set. In describing regular sets by such "regular" expressions it is customary to drop the braces, thus letting a character $a \in \Sigma$ represent $\{a\}$. In the example above we would write

$$L = 0^*10^*010^*(10^* \cup \emptyset^*).$$

For a final example, if $\Sigma = \{a, b, c\}$, then the regular expression

$$c^*(a \cup (bc^*))^*$$

denotes the set of words in Σ^* which contain no occurances of the substring ac.

Exercises Section 5.4

1. Using the equivalence of 1) and 3) in Nerode's theorem, give two proofs that for any single word x, the set $\{x\}$ is regular.

2. If $x = \sigma_1 \cdots \sigma_n$ is any word let $x^R = \sigma_n \cdots \sigma_1$, that is x^R is the *reverse* of x. Notice that $x^{RR} = x$ and for words x, z, $(xz)^R = z^R x^R$. For any set W of words let $W^R = \{x^R : x \in W\}$. Obviously $W^{RR} = W$. With reference to the relation π in 3) of Nerode's theorem, define π^R by

$$x\pi^R y \Leftrightarrow (\forall z \in \Sigma^*)(zx \in L^R \Leftrightarrow zy \in L^R).$$

Show that $x\pi^R y$ iff $x^R \pi y^R$ for all words $x, y \in \Sigma^*$. Conclude that L is regular iff L^R is regular. (Notice that for finite state machine, since there is only one initial state and possibly many accepting states, it is not clear how to directly describe the "reverse" of a machine.)

3. If $A_i = (S_i, m_i, s_{0,i}, F_i)$, $i = 1, 2$, are two finite state machines over the same alphabet, their *direct product* is the machine

$$A_1 \times A_2 = (S_1 \times S_2, m_1 \times m_2, (s_{0,1}, s_{0,2}), F_1 \times F_2)$$

where $\times$ denotes the cartesian product of sets, and for any word x, and ordered pair of states $(s_1, s_2) \in S_1 \times S_2$,

$$m_1 \times m_2((s_1, s_2), x) = (m_1(s_1, x), m_2(s_2, x)).$$

Show that $L(A_1 \times A_2) = L(A_1) \cap L(A_2)$ and conclude that the intersection of two regular sets is regular.

4. Use either Myhill's or Nerode's theorem to give another proof that the intersection of two regular sets is regular. Hint: for equivalence relations θ_1 and θ_2 on the same set $\theta_1 \cap \theta_2$ is the equivalence relation whose equivalence classes are the various nonempty intersections of θ_1 classes with θ_2 classes. Formally,

$$x(\theta_1 \cap \theta_2)y \Leftrightarrow x\theta_1 y \text{ and } x\theta_2 y.$$

5. Show that the set of all regular languages in Σ^* is a field of sets.

6. Show that any finite set of words in Σ^* is regular.

7. Suppose someone gave you a finite state machine and told you only that it had n states. Describe an algorithm for deciding in a finite number of steps whether or not $L(A) = \emptyset$.

Chapter 6

Appendix: Induction

One of the fundamental defining properties of the positive integers $\mathbb{Z}^+$ is the *well-ordering* principle:

Proposition 6.0.1 *(Well-ordering principle for $\mathbb{Z}^+$) If S is any nonempty subset of $\mathbb{Z}^+$, then S has a least element; stated symbolically this can be expressed as*

$$S \subset \mathbb{Z}^+ \text{ and } S \neq \emptyset \Longrightarrow (\exists a \in S)(\forall x \in S)(a \leq x).$$

In particular, let p be a propositional function on $\mathbb{Z}^+$, i.e.: for each $n \in \mathbb{Z}^+$, $p(n)$ is a proposition about n which is either true or false. Suppose we wish to prove that $p(n)$ is true for all positive integers. Then if $T = \{n \in \mathbb{Z}^+ : p(n) \text{ is true}\}$, to prove that $p(n)$ is true for all positive integers is equivalent to proving that the subset $S = \mathbb{Z}^+ - T$ of *failure integers* is empty. Hence we argue that if S is nonempty then by the well-ordering principle S has a least element, i.e.: there is a *least* failure integer n. One then tries to obtain a contradiction—usually by constructing a smaller failure integer—to complete the proof. This idea was recognized in the 17th century by Fermat and was called the *principle of infinite descent*. The name was apparently intended to suggest that if $p(n)$ was false for some integer then one could imagine descending through all of the integers to a least failure case.

The well-ordering principle for $\mathbb{Z}^+$ is more commonly expressed in one of the forms of the principle of mathematical induction (also called the principle of finite induction). For subsets of $\mathbb{Z}^+$ this takes the following form:

Theorem 6.0.5 *If a set T of positive integers has the properties*
a) $1 \in T$, and
b) $(\forall n \in \mathbb{Z}^+)(n \in T \Rightarrow n + 1 \in T)$,
then $T = \mathbb{Z}^+$.

PROOF As observed above it is enough to show that $S = \mathbb{Z}^+ - T$ is empty; if not S has a least element n. By a) $n \neq 1$. Hence $n > 1$ so $n - 1 \in \mathbb{Z}^+$ and $n - 1 < n$ so

$n-1 \in T$ by the definition of n. By b) $n = (n-1)+1 \in T$, a contradiction. Hence $S = \emptyset$ so $T = \mathbb{Z}^+$.

The most commonly cited version of the principle of mathematical induction is called the principle of *weak induction*

Theorem 6.0.6 *(Principle of weak induction) Let p be a propositional function on $\mathbb{Z}^+$ with the properties*

a) $p(1)$ is true, and

b) $(\forall n \in \mathbb{Z}^+)[p(n) \Rightarrow p(n+1)]$ is true.

Then $p(n)$ is true for all $n \in \mathbb{Z}^+$.

PROOF Apply the preceding theorem to $T = \{n \in \mathbb{Z}^+ : p(n) \text{ is true}\}$.

The principle of weak induction permits us to *assume* the truth of $p(n)$ *gratis* in proving the truth of $p(n+1)$. It is called weak induction because the hypothesis $p(n)$ of the implication $p(n) \Rightarrow p(n+1)$ occurring in b), is weak in the sense that we assume the truth of $p(n)$ only for this single n preceding $n+1$. In an induction proof, establishing the truth of $p(1)$ is often called the base step. Establishment of *b)* is called the *induction step* and the assumption of the truth of $p(n)$ in the induction step is called the *induction hypothesis*.

The following principle of *strong induction* is so called since in the corresponding implication we strengthen the induction hypothesis to include the assumption of the truth of $p(k)$ for *all* preceding k. The name is not intended to convey the idea that it is a stronger principle—indeed since the antecedent of the implication is stronger it appears to be formally weaker than weak induction. In fact, all of these principles are equivalent in the presence of the other usual properties of the positive integers.

Theorem 6.0.7 *(Principle of strong induction) Let p be a propositional function on $\mathbb{Z}^+$ with the property*

> *for each $m \in \mathbb{Z}^+$, the assumption that $p(k)$ is true for all $1 \leq k < m$ implies the conclusion that $p(m)$ is itself true.*

Then $p(n)$ is true for all $n \in \mathbb{Z}^+$.

Notice that we are in effect deleting statement a) in the principle of weak induction and replacing b) by the assertion that the formal statement

b′) $(\forall m \in \mathbb{Z}^+)[(\forall k \in \mathbb{Z}^+)(k < m \Rightarrow p(k)) \Rightarrow p(m)]$

is true. But suppose that $m = 1$ in b′). Then there are no $k < m$ in $\mathbb{Z}^+$. Hence for each k, the statement $k < m$ is false so that the implication $(\forall k \in \mathbb{Z}^+)(k < m \Rightarrow p(k))$ is vacuously true. Hence to establish b′) we must explicitly prove that $p(1)$ is true, i.e.: b′) actually includes a).

To prove the principle of strong induction, as before, let S be the set of integers n for which $p(n)$ is false. Unless $S = \emptyset$ it will have a least element m. By the choice of m, $p(k)$ will be true for all $k < m$; hence by b′) $p(m)$ is true, giving a contradiction.

The only way out is to admit that S is empty.

In a proof by strong induction we refer to the assumption of the truth of $p(k)$ for all $k < m$ as the induction hypothesis.

EXAMPLES

1) Prove the formula

$$1 + 2 + \cdots + n = \frac{(n+1)n}{2}.$$

We define the proposition p for each n by

$$p(n): \quad 1 + 2 + \cdots + n = \frac{(n+1)n}{2}.$$

Then $p(1)$ is the proposition $1 = \frac{(1+1)1}{2}$ which is true. This establishes *a)* of the Principle of weak induction.

To establish *b)* we *assume*, for any positive integer n, that $p(n)$ is true, that is we assume

$$1 + \cdots + n = \frac{(n+1)n}{2}.$$

From this it follows that

$$1 + \cdots + n + (n+1) = \frac{(n+1)n}{2} + (n+1) = \frac{(n+2)(n+1)}{2}$$

is true, i.e.: $p(n+1)$ is true.

This example illustrates what might be considered a weakness of the principle of mathematical induction: although we have doubtless established the truth of the formula in question, induction gave us no clue as to how to discover the formula in the first place! In fact the reader is probably familiar with a much simpler way to both discover and establish the truth of this formula: just add the two sums $S = 1+2+\cdots+n$ and $S = n + \cdots + 2 + 1$ term by term to obtain the value $2S = n(n+1)$. (See page 104 of the text.) For this reason, although induction is a fundamental principle about the integers, its use in establishing propositions about the integers is often the last resort when all else fails.

2) In contrast to the preceding example, we consider a situation where induction is absolutely essential: namely where we wish to prove an assertion about entities which are actually *defined* inductively; in this case there is obviously no escape from induction.

Define the following sets of symbols as indicated:

$V = \{x_1, x_2, \ldots\}$, (variables)

$O = \{-, +, \times, /\}$, (operations).

The alphabet $A = V \cup O$ and the set of all *words* (character strings) over A is denoted by A^*. If $u, v \in A^*$ then uv denotes their concatenation. The set F of *prefix formulas* of arithmetic is defined as follows:

$w \in A^*$ is a member of F iff

a) $w \in V$, or

b) w is $-u$ where $u \in F$, or

c) w is one of $+uv$, $\times uv$, or $/uv$ for some $u, v \in F$.

The set F of (prefix) formulas is inductively (or recursively) defined since the definition asserts precisely what are the formulas containing just one character (namely the variables x_i), and then in general defines formulas of $n > 1$ characters in terms of formulas of fewer than n characters. For example, if $\times uv$ has n characters, then it is a formula if u and v are already known to be formulas and the sum of the number of characters in u and v is $n - 1$. To show that such u and v are formulas involves demonstrating that they each satisfy one of the clauses a)-c) of the definition, etc.

For example the usual *infix* formula $(x_1 + x_2) + x_3$ becomes $++x_1x_2x_3$ in prefix form while $x_1 + (x_2 + x_3)$ becomes $+x_1 + x_2x_3$.

Next define the integer valued function f on the nonempty words of A^* by
$f(x_i) = -1$ for each $x_i \in V$,
$f(-) = 0$,
$f(+) = f(\times) = f(/) = +1$
and if $w = s_1 \dots s_k \in A^*$ where each $s_i \in A$, $f(w) = f(s_1) + \cdots + f(s_k)$.

Proposition *Let $w = s_1 \cdots s_n \in A^*$ where each $s_i \in A$. Then $w \in F$ iff a) $f(w) = -1$ and b) for each k in the range $1 \leq k < n$, $f(s_1 \cdots s_k) \geq 0$.*

Briefly, we refer to $f(w)$ as the *weight* of w. The proposition says that a word is a formula iff it has weight -1 and the weight of each of its *proper initial segments* is nonnegative. The proposition provides an efficient algorithm for determining if a given character string is a properly constructed prefix formula. For example the word $\times x_1 + x_2x_3$ (which in infix becomes $x_1(x_2 + x_3)$) has weights of its initial segments $= 1, 0, +1, 0, -1$ and hence is a prefix formula.

We use induction to prove this proposition. To do this we let $p(n)$ be the statement of the proposition. Since we will need to assume that a formula of n characters is constructed from formulas of fewer than n characters, but not necessarily $n - 1$ characters, it is appropriate to use strong induction. We will prove the "only if " direction of the proposition here and leave the "if" direction for the exercises. Hence we begin by supposing that $w = s_1 \cdots s_n$ is a prefix formula; we must show that f has the properties *a)* and *b)*.

If $n = 1$ then since w is a formula $w = x_i$ for some variable. It follows that *a)* $f(w) = -1$ and since w in this case has no proper initial segments, *b)* is vacuously true.

Now for the induction step. Suppose that w is a formula of n characters where we may suppose $n > 1$. We must consider each of the conditions of the definition of prefix formulas. Case a) does not apply since $n > 1$. For case b), $w = -u$, where u is a formula. By the induction hypothesis, $f(u) = -1$ and hence $f(w) = f(-u) = f(-) + f(u) = 0 + (-1) = -1$. Thus w has weight -1. Further, since $f(-) = 0$ and by the induction hypothesis each proper initial segment of u has nonnegative weight, it follows that each proper initial segment of $w = -u$ has nonnegative weight.

For case c) suppose u and v are formulas and $w = buv$ where b is any of $+$, $\times$, or $/$; the argument is the same for each. Since each of u and v has fewer than n characters, by the induction hypothesis each has weight -1. Hence $f(buv) = f(b)+f(u)+f(v) = +1 - 1 - 1 = -1$ so w has weight -1. Finally, consider a proper initial segment s of $w = buv$:

If s consists of just b then its weight is $+1$.

If s consists of b followed by a proper initial segment of u then the weight of s is $+1$ + the weight of the proper initial segment of u, so this is positive.

If s consists of bu then $f(bu) = +1 - 1 = 0$.

If s consists of bu followed by a proper initial segment of v then the weight of s equals the weight of the proper initial segment of v, which is nonnegative.

Hence in all cases, $f(s)$ is nonnegative; this establishes the "only if " direction of the Proposition.

Exercises Induction

1. Use induction to prove $1 + 8 + \cdots + n^3 = [n(n + 1)/2]^2$.

2. Use induction to prove that for every integer $n \geq 5$, $2^n > n^2$. (This exercise illustrates the fact that the principle of mathematical induction is easily seen to apply not just to $\mathbb{Z}^+$, but to any nonempty subset of $\mathbb{Z}$ having a least element.)

3. Prove the well-ordering principle for $\mathbb{Z}^+$ from the principle of weak induction.

4. What is wrong with the following proof that every group of students is either all male or all female? The proof is by weak induction on the number n of students in the group.

$n = 1$: In this case a group of just one student is obviously all male or all female.

Induction step: Suppose that no group of n students has mixed sexes, and suppose $s_1, s_2, \ldots, s_{n+1}$ is a group of $n + 1$ students. Form the two groups of n students:

$$s_1, s_2, \ldots, s_n$$

$$s_2, \ldots, s_n, s_{n+1}.$$

Then by the induction hypothesis each of these two groups is single sexed. Since s_n is in both groups, it follows that both groups have the same sex; hence all of $s_1, \ldots, s_{n+1}$ have the same sex.

5. Prove the "if" direction of the Proposition of Example 2 above.

6. If $n \geq 0$ straight lines are drawn in the plane let L_n be the maximum number of regions defined. If we think of the plane as a giant pizza, L_n is the maximum number of slices of pizza which can be obtained by making n straight cuts with a pizza knife. Thus $L_0 = 1$, $L_1 = 2$, $L_2 = 4$, $L_3 = 7$ (Draw pictures to see these.) Find a formula for L_n. Hint: Begin by establishing the recursive formula $L_n = L_{n-1} + n$. Then find a closed, i.e.: nonrecursive, formula for L_n.

7. In Dotville everyone has a dot on his forehead which is either red or blue. The law in Dotville is that if you learn the color of the dot on your forehead some day, then you must commit suicide at midnight of the same day. If you learn your color at midnight, you must commit suicide at midnight 24 hours later. There are no mirrors in Dotville, everyone in Dotville can see everyone else in Dotville, and the residents of Dotville do not speak to each other. One day a stranger comes to town, looks around, and says, "Someone has a blue dot on his head." Then he leaves. What happens? (We assume that the stranger was telling the truth.)

Hint: First, of course, try a few simple cases to help formulate a conjecture. Then formulate a proposition $p(n)$ where $n \in \mathbb{Z}^+$ is the number of blue dot people and which describes what happens if the stranger comes to town on day 0.

Dotville Problem: Solution

Let $p(n)$ be:
If the stranger comes to town on day 0 and there are n blue dot people, then all blue dot people commit suicide at midnight of day $n-1$ and all red dot people commit suicide at midnight of day n.

We prove that $p(n)$ is true for all $n \in \mathbb{Z}^+$, by induction on n.

PROOF $n = 1$. In this case the only blue dot person will realize that he has a blue dot because the stranger said that there was a blue dot person and he sees that everyone else has a red dot. So he kills himself at midnight of day $n-1 = 1-1 = 0$. On the other hand, each red dot person thinks: if I had a blue dot, then there would have been two of us and the blue dot who just committed suicide could not have known that he was the only blue dot so he would *not* have committed suicide. But he did, therefore I have a red dot. Hence each red dot commits suicide at midnight of day $n = 1$. This proves $p(1)$.

Induction step. Suppose $p(n)$ is true and that Dotville has $n+1$ blue dot residents. The stranger, on day 0, says that someone has a blue dot. Each blue dot person thinks: I see n blue dots. Since $p(n)$ is true, if I have a red dot then there are exactly n blue dots in town so by the induction hypothesis, all of the blue dots will commit suicide at midnight of day $n-1$, so he waits (patiently) until midnight of day $n-1$ and nothing happens. Thus he concludes that he must have a blue dot so he commits suicide at midnight of the next day, namely day $(n = (n+1) - 1$. Each of the red dot people thinks: I saw n blue dot people and they all committed suicide on day n. If I also had a blue dot, then (by the reasoning just described for the blue dots) the other n could not have known that they had blue dots on day n, therefore I have a red dot. Hence all red dots commit suicide on day $n+1$. Hence $p(n+1)$ is true.