16

Groups, Coding Theory, and Polya's Method of Enumeration

In our study of algebraic structures we examine properties shared by particular mathematical systems. Then we generalize our findings in order to study the underlying structure common to these particular examples.

In Chapter 14 we did this with the ring structure, which depended on two closed binary operations. Now we turn to a structure involving one closed binary operation. This structure is called a *group*.

Our study of groups will examine many ideas comparable to those for rings. However, here we shall dwell primarily on those aspects of the structure that are needed for applications in cryptology, coding theory, and a counting method developed by George Polya.

16.1 Definition, Examples, and Elementary Properties

Definition 16.1 If G is a nonempty set and $\circ$ is a binary operation on G, then $(G, \circ)$ is called a *group* if the following conditions are satisfied.

1) For all $a, b \in G$, $a \circ b \in G$. (Closure of G under $\circ$)
2) For all $a, b, c \in G$, $a \circ (b \circ c) = (a \circ b) \circ c$. (The Associative Property)
3) There exists $e \in G$ with $a \circ e = e \circ a = a$, for all $a \in G$. (The Existence of an Identity)
4) For each $a \in G$ there is an element $b \in G$ such that $a \circ b = b \circ a = e$. (Existence of Inverses)

Furthermore, if $a \circ b = b \circ a$ for all $a, b \in G$, then G is called a *commutative*, or *abelian*, group. The adjective *abelian* honors the Norwegian mathematician Niels Henrik Abel (1802–1829).

We realize that the first condition in Definition 16.1 could have been omitted if we simply required the binary operation for G to be a *closed* binary operation.

Following Definition 14.1 (for a ring) we mentioned how the associative laws for the closed binary operations of $+$ (ring addition) and $\cdot$ (ring multiplication) could be extended by mathematical induction. The same type of situation arises for groups. If $(G, \circ)$ is any group, and $r, n \in \mathbf{Z}^+$ with $n \geq 3$ and $1 \leq r < n$, then

$$(a_1 \circ a_2 \circ \cdots \circ a_r) \circ (a_{r+1} \circ \cdots \circ a_n) = a_1 \circ a_2 \circ \cdots \circ a_r \circ a_{r+1} \circ \cdots \circ a_n,$$

where $a_1, a_2, \ldots, a_r, a_{r+1}, \ldots, a_n$ are all elements from G.

EXAMPLE 16.1

Under ordinary addition, each of $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$ is an abelian group. None of these is a group under multiplication because 0 has no multiplicative inverse. However, $\mathbf{Q}^*$, $\mathbf{R}^*$, and $\mathbf{C}^*$ (the nonzero elements of $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{C}$, respectively) are abelian groups under ordinary multiplication.

If $(R, +, \cdot)$ is a ring, then $(R, +)$ is an abelian group; the nonzero elements of a *field* $(F, +, \cdot)$ form the abelian group $(F^*, \cdot)$.

EXAMPLE 16.2

For $n \in \mathbf{Z}^+$, $n > 1$, we find that $(\mathbf{Z}_n, +)$ is an abelian group. When p is a prime, $(\mathbf{Z}_p^*, \cdot)$ is an abelian group. Tables 16.1 and 16.2 demonstrate this for $n = 6$ and $p = 7$, respectively. (Recall that in $\mathbf{Z}_n$ we often write a for $[a] = \{a + kn | k \in \mathbf{Z}\}$. The same notation is used in $\mathbf{Z}_p^*$.)

Table 16.1

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

Table 16.2

·	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	1	3	5
3	3	6	2	5	1	4
4	4	1	5	2	6	3
5	5	3	1	6	4	2
6	6	5	4	3	2	1

Definition 16.2

For every group G the number of elements in G is called the *order* of G and this is denoted by $|G|$. When the number of elements in a group is not finite we say that G has infinite order.

EXAMPLE 16.3

For all $n \in \mathbf{Z}^+$, $|(\mathbf{Z}_n, +)| = n$, while $|(\mathbf{Z}_p^*, \cdot)| = p - 1$ for each prime p.

EXAMPLE 16.4

Let us start with the ring $(\mathbf{Z}_9, +, \cdot)$ and consider the subset $U_9 = \{a \in \mathbf{Z}_9 | a \text{ is a unit in } \mathbf{Z}_9\} = \{a \in \mathbf{Z}_9 | a^{-1} \text{exists}\} = \{1, 2, 4, 5, 7, 8\} = \{a \in \mathbf{Z}^+ | 1 \leq a \leq 8 \text{ and } \gcd(a, 9) = 1\}$. The results in Table 16.3 show us that U_9 is closed under the multiplication for the ring $(\mathbf{Z}_9, +, \cdot)$ — namely, multiplication modulo 9. Furthermore, we also see that 1 is the identity element and that each element has an inverse (in U_9). For instance, 5 is the inverse for 2, and 7 is the inverse for 4. Finally, since every ring is associative under the operation

of (ring) multiplication, it follows that $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in U_9$. Consequently, $(U_9, \cdot)$ is a group of order 6 — in fact, it is an abelian group of order 6.

Table 16.3

$\cdot$	1	2	4	5	7	8
1	1	2	4	5	7	8
2	2	4	8	1	5	7
4	4	8	7	2	1	5
5	5	1	2	7	8	4
7	7	5	1	8	4	2
8	8	7	5	4	2	1

In general, for each $n \in \mathbf{Z}^+$, where $n > 1$, if $U_n = \{a \in (\mathbf{Z}_n, +, \cdot) | \ a \text{ is a unit}\} = \{a \in \mathbf{Z}^+ | 1 \leq a \leq n - 1 \text{ and } \gcd(a, n) = 1\}$, then $(U_n, \cdot)$ is an abelian group under the (closed) binary operation of multiplication modulo n. The group $(U_n, \cdot)$ is called the *group of units* for the ring $(\mathbf{Z}_n, +, \cdot)$ and it has order $\phi(n)$, where ϕ denotes the Euler phi function of Section 8.1.

From here on the group operation will be written multiplicatively, unless it is given otherwise. So $a \circ b$ now becomes ab.

The following theorem provides several properties shared by all groups.

THEOREM 16.1 For every group G,

a) the identity of G is unique.

b) the inverse of each element of G is unique.

c) if $a, b, c \in G$ and $ab = ac$, then $b = c$. (Left-cancellation property)

d) if $a, b, c \in G$ and $ba = ca$, then $b = c$. (Right-cancellation property)

Proof:

a) If e_1, e_2 are both identities in G, then $e_1 = e_1 e_2 = e_2$. (Justify each equality.)

b) Let $a \in G$ and suppose that b, c are both inverses of a. Then $b = be = b(ac) = (ba)c = ec = c$. (Justify each equality.)

The proofs of properties (c) and (d) are left for the reader. (It is because of these properties that we find each group element appearing exactly once in each row and each column of the table for a finite group.)

On the basis of the result in Theorem 16.1(b) the unique inverse of a will be designated by a^{-1}. When the group is written additively, $-a$ is used to denote the (additive) inverse of a.

As in the case of multiplication in a ring, we have powers of elements in a group. We define $a^0 = e$, $a^1 = a$, $a^2 = a \cdot a$, and in general $a^{n+1} = a^n \cdot a$, for all $n \in \mathbf{N}$. Since each group element has an inverse, for $n \in \mathbf{Z}^+$, we define $a^{-n} = (a^{-1})^n$. Then a^n is defined for all $n \in \mathbf{Z}$, and it can be shown that for all $m, n \in \mathbf{Z}$, $a^m \cdot a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$.

If the group operation is addition, then multiples replace powers and for all $m, n \in \mathbf{Z}$, and all $a \in G$, we find that

$$ma + na = (m + n)a \qquad m(na) = (mn)a.$$

In this case the identity is written as 0, rather than e. And here, for all $a \in G$, we have $0a = 0$, where the "0" in front of a is the integer 0 (in $\mathbf{Z}$) while the "0" on the right side of the equation is the identity 0 (in G). [So these two "0"'s are different.]

For an abelian group G we also find that for all $n \in \mathbf{Z}$ and all $a, b \in G$, (1) $(ab)^n = a^n b^n$, when G is written multiplicatively; and (2) $n(a + b) = na + nb$, when the additive notation is used for G.

We now take a look at a special subset of a group.

EXAMPLE 16.5

Let $G = (\mathbf{Z}_6, +)$. If $H = \{0, 2, 4\}$, then H is a nonempty subset of G. Table 16.4 shows that $(H, +)$ is also a group under the binary operation of G.

Table 16.4

+	0	2	4
0	0	2	4
2	2	4	0
4	4	0	2

This situation motivates the following definition.

Definition 16.3

Let G be a group and $\emptyset \neq H \subseteq G$. If H is a group under the binary operation of G, then we call H a *subgroup* of G.

EXAMPLE 16.6

a) Every group G has $\{e\}$ and G as subgroups. These are the *trivial* subgroups of G. All others are termed *nontrivial*, or *proper*.

b) In addition to $H = \{0, 2, 4\}$, the subset $K = \{0, 3\}$ is also a (proper) subgroup of $G = (\mathbf{Z}_6, +)$.

c) Each of the nonempty subsets $\{1, 8\}$ and $\{1, 4, 7\}$ is a subgroup of $(U_9, \cdot)$.

d) The group $(\mathbf{Z}, +)$ is a subgroup of $(\mathbf{Q}, +)$, which is a subgroup of $(\mathbf{R}, +)$. Yet $\mathbf{Z}^*$ under multiplication is not a subgroup of $(\mathbf{Q}^*, \cdot)$. (Why not?)

For a group G and $\emptyset \neq H \subseteq G$, the following tells us when H is a subgroup of G.

THEOREM 16.2

If H is a nonempty subset of a group G, then H is a subgroup of G if and only if (a) for all $a, b \in H$, $ab \in H$, and (b) for all $a \in H$, $a^{-1} \in H$.

Proof: If H is a subgroup of G, then by Definition 16.3 H is a group under the same binary operation. Hence it satisfies all the group conditions, including the two mentioned here. Conversely, let $\emptyset \neq H \subseteq G$ with H satisfying conditions (a) and (b). For all $a, b, c \in H$, $(ab)c = a(bc)$ in G, so $(ab)c = a(bc)$ in H. (We say that H "inherits" the associative

property from G.) Finally, as $H \neq \emptyset$, let $a \in H$. By condition (b), $a^{-1} \in H$ and by condition (a), $aa^{-1} = e \in H$, so H contains the identity element and is a group.

A finiteness condition modifies the situation.

THEOREM 16.3 If G is a group and $\emptyset \neq H \subseteq G$, with H *finite*, then H is a subgroup of G if and only if H is closed under the binary operation of G.

Proof: As in the proof of Theorem 16.2, if H is a subgroup of G, then H is closed under the binary operation of G. Conversely, let H be a finite nonempty subset of G that is so closed. If $a \in H$, then $aH = \{ah | h \in H\} \subseteq H$ because of the closure condition. By left-cancellation in G, $ah_1 = ah_2 \Rightarrow h_1 = h_2$, so $|aH| = |H|$. With $aH \subseteq H$ and $|aH| = |H|$, it follows from H being *finite* that $aH = H$. As $a \in H$, there exists $b \in H$ with $ab = a$. But (in G) $ab = a = ae$, so $b = e$ and H contains the identity. Since $e \in H = aH$, there is an element $c \in H$ such that $ac = e$. Then $(ca)^2 = (ca)(ca) = (c(ac))a = (ce)a = ca = (ca)e$, so $ca = e$, and $c = a^{-1} \in H$. Consequently, by Theorem 16.2, H is a subgroup of G.

The finiteness condition in Theorem 16.3 is crucial. Both $\mathbf{Z}^+$ and $\mathbf{N}$ are nonempty closed subsets of the group $(\mathbf{Z}, +)$, yet neither has the additive inverses needed for the group structure.

The next example provides a nonabelian group.

EXAMPLE 16.7 Consider the first equilateral triangle shown in Fig. 16.1(a). When we rotate this triangle counterclockwise (within its plane) through 120° about an axis perpendicular to its plane and passing through its center C, we obtain the second triangle shown in Fig. 16.1(a). As a result, the vertex originally labeled 1 in Fig. 16.1(a) is now in the position that was originally labeled 3. Likewise, 2 is now in the position originally occupied by 1, and 3 has moved to where 2 was. This can be described by the function π_1: $\{1, 2, 3\} \to \{1, 2, 3\}$, where $\pi_1(1) = 3$, $\pi_1(2) = 1$, $\pi_1(3) = 2$. A more compact notation, $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$, where we write $\pi_1(i)$ below i for each $1 \leq i \leq 3$, emphasizes that π_1 is a permutation of $\{1, 2, 3\}$. If π_2 denotes the counterclockwise rotation through 240°, then $\pi_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$. For the identity π_0 — that is, the rotation through $n(360°)$ for $n \in \mathbf{Z}$ — we write $\pi_0 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$. These rotations are called *rigid motions* of the triangle. They are two-dimensional motions that keep the center C fixed and preserve the shape of the triangle. Hence the triangle looks the same as when we started, except for a possible rearrangement of the labels on some of its vertices.

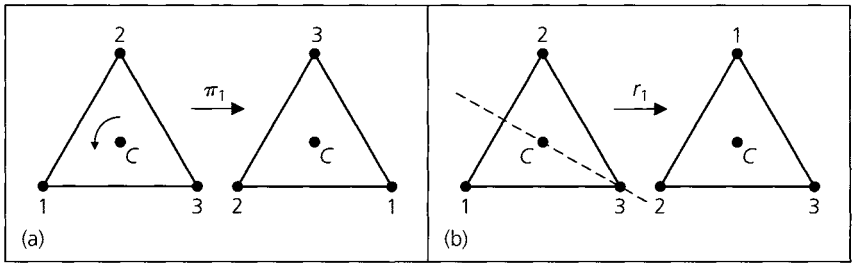


Figure 16.1

In addition to these rotations, the triangle can be reflected along an axis passing through a vertex and the midpoint of the opposite side. For the diagonal axis that bisects the base angle on the right, the reflection gives the result in Fig. 16.1(b). This we represent by $r_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$. A similar reflection about the axis bisecting the left base angle yields the permutation $r_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$. When the triangle is reflected about its vertical axis, we have $r_3 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. Each r_i, for $1 \leq i \leq 3$, is a three-dimensional rigid motion.

Let $G = \{\pi_0, \pi_1, \pi_2, r_1, r_2, r_3\}$, the set of rigid motions (in space) of the equilateral triangle. We make G into a group by defining the rigid motion $\alpha\beta$, for $\alpha, \beta \in G$, as that motion obtained by applying first α and then following up with β. Hence, for example, $\pi_1 r_1 = r_3$. We can see this geometrically, but it will be handy to consider the permutations as follows: $\pi_1 r_1 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$, where, for example, $\pi_1(1) = 3$ and $r_1(3) = 3$ and we write $1 \xrightarrow{\pi_1} 3 \xrightarrow{r_1} 3$. So $1 \xrightarrow{\pi_1 r_1} 3$ in the product $\pi_1 r_1$. (Note that the order in which we write the product $\pi_1 r_1$ here is the opposite of the order for their composite function as defined in Section 5.6. The notation of Section 5.6 occurs in analysis, whereas in algebra there is a tendency to employ this opposite order.) Also, since $2 \xrightarrow{\pi_1} 1 \xrightarrow{r_1} 2$ and $3 \xrightarrow{\pi_1} 2 \xrightarrow{r_1} 1$, it follows that $\pi_1 r_1 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = r_3$.

Table 16.5 verifies that under this binary operation G is closed, with identity π_0. Also $\pi_1^{-1} = \pi_2$, $\pi_2^{-1} = \pi_1$, and every other element is its own inverse. Since the elements of G are actually functions, the associative property follows from Theorem 5.6 (although in reverse order).

Table 16.5

$\cdot$	π_0	π_1	π_2	r_1	r_2	r_3
π_0	π_0	π_1	π_2	r_1	r_2	r_3
π_1	π_1	π_2	π_0	r_3	r_1	r_2
π_2	π_2	π_0	π_1	r_2	r_3	r_1
r_1	r_1	r_2	r_3	π_0	π_1	π_2
r_2	r_2	r_3	r_1	π_2	π_0	π_1
r_3	r_3	r_1	r_2	π_1	π_2	π_0

We computed $\pi_1 r_1$ as r_3, but from Table 16.5 we see that $r_1 \pi_1 = r_2$. With $\pi_1 r_1 = r_3 \neq r_2 = r_1 \pi_1$, it follows that G is nonabelian.

This group can also be obtained as the group of all permutations of the set $\{1, 2, 3\}$ under the binary operation of function composition. It is denoted by S_3 (the *symmetric* group on three symbols).

EXAMPLE 16.8

The symmetric group S_4 consists of the 24 permutations of $\{1, 2, 3, 4\}$. Here $\pi_0 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$ is the identity. If $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$, $\beta = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix}$, then $\alpha\beta = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 2 \end{pmatrix}$ but $\beta\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix}$, so S_4 is nonabelian. Also, $\beta^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 4 \end{pmatrix}$ and $\alpha^2 = \pi_0 = \beta^3$. Within S_4 there is a subgroup of order 8 that represents the group of rigid motions for a square.

We turn now to a construction for making larger groups out of smaller ones.

THEOREM 16.4 Let $(G, \circ)$ and $(H, *)$ be groups. Define the binary operation $\cdot$ on $G \times H$ by $(g_1, h_1) \cdot (g_2, h_2) = (g_1 \circ g_2, h_1 * h_2)$. Then $(G \times H, \cdot)$ is a group and is called the *direct product* of G and H.

Proof: The verification of the group properties for $(G \times H, \cdot)$ is left to the reader.

EXAMPLE 16.9 Consider the groups $(\mathbf{Z}_2, +)$, $(\mathbf{Z}_3, +)$. On $G = \mathbf{Z}_2 \times \mathbf{Z}_3$, define $(a_1, b_1) \cdot (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$. Then G is a group of order 6 where the identity is (0, 0), and the inverse, for example, of the element (1, 2) is (1, 1).

EXERCISES 16.1

1. For each of the following sets, determine whether or not the set is a group under the stated binary operation. If so, determine its identity and the inverse of each of its elements. If it is not a group, state the condition(s) of the definition that it violates.

a) $\{-1, 1\}$ under multiplication

b) $\{-1, 1\}$ under addition

c) $\{-1, 0, 1\}$ under addition

d) $\{10n | n \in \mathbf{Z}\}$ under addition

e) The set of all one-to-one functions $g: A \to A$, where $A = \{1, 2, 3, 4\}$, under function composition

f) $\{a/2^n | a, n \in \mathbf{Z}, n \geq 0\}$ under addition

2. Prove parts (c) and (d) of Theorem 16.1.

3. Why is the set $\mathbf{Z}$ *not* a group under subtraction?

4. Let $G = \{q \in \mathbf{Q} | q \neq -1\}$. Define the binary operation $\circ$ on G by $x \circ y = x + y + xy$. Prove that $(G, \circ)$ is an abelian group.

5. Define the binary operation $\circ$ on $\mathbf{Z}$ by $x \circ y = x + y + 1$. Verify that $(\mathbf{Z}, \circ)$ is an abelian group.

6. Let $S = \mathbf{R}^* \times \mathbf{R}$. Define the binary operation $\circ$ on S by $(u, v) \circ (x, y) = (ux, vx + y)$. Prove that $(S, \circ)$ is a non-abelian group.

7. Find the elements in the groups U_{20} and U_{24} — the groups of units for the rings $(\mathbf{Z}_{20}, +, \cdot)$ and $(\mathbf{Z}_{24}, +, \cdot)$, respectively.

8. For any group G prove that G is abelian if and only if $(ab)^2 = a^2b^2$ for all $a, b \in G$.

9. If G is a group, prove that for all $a, b \in G$,

a) $(a^{-1})^{-1} = a$ **b)** $(ab)^{-1} = b^{-1}a^{-1}$

10. Prove that a group G is abelian if and only if for all $a, b \in G$, $(ab)^{-1} = a^{-1}b^{-1}$.

11. Find all subgroups in each of the following groups.

a) $(\mathbf{Z}_{12}, +)$ **b)** $(\mathbf{Z}_{11}^*, \cdot)$ **c)** S_3

12. a) How many rigid motions (in two or three dimensions) are there for a square?

b) Make a group table for these rigid motions like the one in Table 16.5 for the equilateral triangle. What is the identity for this group? Describe the inverse of each element geometrically.

13. a) How many rigid motions (in two or three dimensions) are there for a regular pentagon? Describe them geometrically.

b) Answer part (a) for a regular n-gon, $n \geq 3$.

14. In the group S_5, let

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 4 & 5 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 5 & 3 & 4 \end{pmatrix}.$$

Determine $\alpha\beta$, $\beta\alpha$, α^3, β^4, α^{-1}, β^{-1}, $(\alpha\beta)^{-1}$, $(\beta\alpha)^{-1}$, and $\beta^{-1}\alpha^{-1}$.

15. If G is a group, let $H = \{a \in G | ag = ga \text{ for all } g \in G\}$. Prove that H is a subgroup of G. (The subgroup H is called the *center* of G.)

16. Let ω be the complex number $(1/\sqrt{2})(1 + i)$.

a) Show that $\omega^8 = 1$ but $\omega^n \neq 1$ for $n \in \mathbf{Z}^+$, $1 \leq n \leq 7$.

b) Verify that $\{\omega^n | n \in \mathbf{Z}^+, 1 \leq n \leq 8\}$ is an abelian group under multiplication.

17. a) Prove Theorem 16.4.

b) Extending the idea developed in Theorem 16.4 and Example 16.9 to the group $\mathbf{Z}_6 \times \mathbf{Z}_6 \times \mathbf{Z}_6 = \mathbf{Z}_6^3$, answer the following.

i) What is the order of this group?

ii) Find a subgroup of $\mathbf{Z}_6^3$ of order 6, one of order 12, and one of order 36.

iii) Determine the inverse of each of the elements (2, 3, 4), (4, 0, 2), (5, 1, 2).

18. a) If H, K are subgroups of a group G, prove that $H \cap K$ is also a subgroup of G.

b) Give an example of a group G with subgroups H, K such that $H \cup K$ is *not* a subgroup of G.

19. a) Find all x in $(\mathbf{Z}_5^*, \cdot)$ such that $x = x^{-1}$.

b) Find all x in $(\mathbf{Z}_{11}^*, \cdot)$ such that $x = x^{-1}$.

c) Let p be a prime. Find all x in $(\mathbf{Z}_p^*, \cdot)$ such that $x = x^{-1}$.

d) Prove that $(p-1)! \equiv -1 \pmod{p}$, for p a prime. [This result is known as Wilson's Theorem, although it was only conjectured by John Wilson (1741–1793). The first proof was given in 1770 by Joseph Louis Lagrange (1736–1813).]

20. a) Find x in $(U_8, \cdot)$ where $x \neq 1$, $x \neq 7$ but $x = x^{-1}$.

b) Find x in $(U_{16}, \cdot)$ where $x \neq 1$, $x \neq 15$ but $x = x^{-1}$.

c) Let $k \in \mathbf{Z}^+$, $k \geq 3$. Find x in $(U_{2^k}, \cdot)$ where $x \neq 1$, $x \neq 2^k - 1$ but $x = x^{-1}$.

16.2 Homomorphisms, Isomorphisms, and Cyclic Groups

We turn our attention once again to functions that preserve structure.

EXAMPLE 16.10

Let $G = (\mathbf{Z}, +)$ and $H = (\mathbf{Z}_4, +)$. Define $f: G \to H$ by

$$f(x) = [x] = \{x + 4k \mid k \in \mathbf{Z}\}.$$

For all $x, y \in G$,

$$f(x + y) = [x + y] = [x] + [y] = f(x) + f(y),$$

($f(x+y)$: the operation in G; $[x]+[y]$: the operation in H)

where the second equality follows from the way the addition of equivalence classes was developed in Section 14.3. Consequently, here f preserves the group operations and is an example of a special type of function that we shall now define.

Definition 16.4

If $(G, \circ)$ and $(H, *)$ are groups and $f: G \to H$, then f is called a *group homomorphism* if for all $a, b \in G$, $f(a \circ b) = f(a) * f(b)$.

When we know that the given structures are groups, the function f is simply called a homomorphism.

Some properties of homomorphisms are given in the following theorem.

THEOREM 16.5

Let $(G, \circ)$, $(H, *)$ be groups with respective identities e_G, e_H. If $f: G \to H$ is a homomorphism, then

a) $f(e_G) = e_H$.

b) $f(a^{-1}) = [f(a)]^{-1}$ for all $a \in G$.

c) $f(a^n) = [f(a)]^n$ for all $a \in G$ and all $n \in \mathbf{Z}$.

d) $f(S)$ is a subgroup of H for each subgroup S of G.

Proof:

a) $e_H * f(e_G) = f(e_G) = f(e_G \circ e_G) = f(e_G) * f(e_G)$, so by right-cancellation [Theorem 16.1(d)], it follows that $f(e_G) = e_H$.

b) & c) The proofs of these parts are left for the reader.

d) If S is a subgroup of G, then $S \neq \emptyset$, so $f(S) \neq \emptyset$. Let $x, y \in f(S)$. Then $x = f(a)$, $y = f(b)$, for some $a, b \in S$. Since S is a subgroup of G, it follows that $a \circ b \in S$, so $x * y = f(a) * f(b) = f(a \circ b) \in f(S)$. Finally, $x^{-1} = [f(a)]^{-1} = f(a^{-1}) \in f(S)$ because $a^{-1} \in S$ when $a \in S$. Consequently, by Theorem 16.2, $f(S)$ is a subgroup of H.

Definition 16.5

If $f: (G, \circ) \to (H, *)$ is a homomorphism, we call f an *isomorphism* if it is one-to-one and onto. In this case G, H are said to be *isomorphic groups*.

EXAMPLE 16.11

Let $f: (\mathbf{R}^+, \cdot) \to (\mathbf{R}, +)$ where $f(x) = \log_{10}(x)$. This function is both one-to-one and onto. (Verify these properties.) For all $a, b \in \mathbf{R}^+$, $f(ab) = \log_{10}(ab) = \log_{10} a + \log_{10} b = f(a) + f(b)$. Therefore, f is an isomorphism and the group of positive real numbers under multiplication is abstractly the same as the group of all real numbers under addition. Here the function f translates a problem in the multiplication of real numbers (a somewhat difficult problem without a calculator) into a problem dealing with the addition of real numbers (an easier arithmetic consideration). This was a major reason behind the use of logarithms before the advent of calculators.

EXAMPLE 16.12

Let G be the group of complex numbers $\{1, -1, i, -i\}$ under multiplication. Table 16.6 shows the multiplication table for this group. With $H = (\mathbf{Z}_4, +)$, consider $f: G \to H$ defined by

$$f(1) = [0] \qquad f(-1) = [2] \qquad f(i) = [1] \qquad f(-i) = [3].$$

Then $f((i)(-i)) = f(1) = [0] = [1] + [3] = f(i) + f(-i)$, and $f((-1)(-i)) = f(i) = [1] = [2] + [3] = f(-1) + f(-i)$.

Although we have not checked all possible cases, the function is an isomorphism. Note that the image under f of the subgroup $\{1, -1\}$ of G is $\{[0], [2]\}$, a subgroup of H.

Table 16.6

$\cdot$	1	-1	i	$-i$
1	1	-1	i	$-i$
-1	-1	1	$-i$	i
i	i	$-i$	-1	1
$-i$	$-i$	i	1	-1

Let us take a closer look at this group G. Here $i^1 = i$, $i^2 = -1$, $i^3 = -i$, and $i^4 = 1$, so every element of G is a power of i, and we say that i *generates* G. This is denoted by $G = \langle i \rangle$. (It is also true that $G = \langle -i \rangle$. Verify this.)

The last part of the preceding example leads us to the following definition.

Definition 16.6

A group G is called *cyclic* if there is an element $x \in G$ such that for each $a \in G$, $a = x^n$ for some $n \in \mathbf{Z}$.

EXAMPLE 16.13

a) The group $H = (\mathbf{Z}_4, +)$ is cyclic. Here the operation is addition, so we have multiples instead of powers. We find that both [1] and [3] generate H. For the case of [3], we have $1 \cdot [3] = [3]$, $2 \cdot [3]$ $(= [3] + [3]) = [2]$, $3 \cdot [3] = [1]$, and $4 \cdot [3] = [0]$. Hence $H = \langle [3] \rangle = \langle [1] \rangle$.

b) Consider the multiplicative group $U_9 = \{1, 2, 4, 5, 7, 8\}$ that we examined in Example 16.4. Here we find that $2^1 = 2$, $2^2 = 4$, $2^3 = 8$, $2^4 = 7$, $2^5 = 5$, $2^6 = 1$, so U_9 is a cyclic group of order 6 and $U_9 = \langle 2 \rangle$. It is also true that $U_9 = \langle 5 \rangle$ because $5^1 = 5$, $5^2 = 7$, $5^3 = 8$, $5^4 = 4$, $5^5 = 2$, $5^6 = 1$.

The concept of a cyclic group leads to a related idea. Given a group G, if $a \in G$ consider the set $S = \{a^k | k \in \mathbf{Z}\}$. From Theorem 16.2 it follows that S is a subgroup of G. This subgroup is called the *subgroup generated by a* and is designated by $\langle a \rangle$. In Example 16.12 $\langle i \rangle = \langle -i \rangle = G$; also, $\langle -1 \rangle = \{-1, 1\}$ and $\langle 1 \rangle = \{1\}$. For part (a) of Example 16.13 we consider multiples instead of powers and find that $H = \langle [1] \rangle = \langle [3] \rangle$, $\langle [2] \rangle = \{[0], [2]\}$, and $\langle [0] \rangle = \{[0]\}$. When we examine the group U_9 in part (b) of that example we see that $U_9 = \langle 2 \rangle$ (or $\langle [2] \rangle$) $= \langle 5 \rangle$, $\langle 4 \rangle = \{1, 4, 7\} = \langle 7 \rangle$, $\langle 8 \rangle = \{1, 8\}$, and $\langle 1 \rangle = \{1\}$.

Definition 16.7 If G is a group and $a \in G$, the *order of a*, denoted $\mathcal{O}(a)$, is $|\langle a \rangle|$. (If $|\langle a \rangle|$ is infinite, we say that a has infinite order.)

In Example 16.12, $\mathcal{O}(1) = 1$, $\mathcal{O}(-1) = 2$, whereas both i and $-i$ have order 4.

Let us take a second look at the idea of order for the case where $|\langle a \rangle|$ is finite. When $|\langle a \rangle| = 1$ then $a = e$ because $a = a^1 \in \langle a \rangle$ and $e = a^0 \in \langle a \rangle$. If $|\langle a \rangle|$ is finite but $a \neq e$, then $\langle a \rangle = \{a^m | m \in \mathbf{Z}\}$ is finite, so $\{a, a^2, a^3, \ldots\} = \{a^m | m \in \mathbf{Z}^+\}$ is also finite. Consequently, there exist $s, t \in \mathbf{Z}^+$, where $1 \leq s < t$ and $a^s = a^t$ —from which it follows that $a^{t-s} = e$, with $t - s \in \mathbf{Z}^+$. Since $e \in \{a^m | m \in \mathbf{Z}^+\}$, let n be the smallest positive integer such that $a^n = e$. We claim that $\langle a \rangle = \{a, a^2, a^3, \ldots, a^{n-1}, a^n\ (= e)\}$.

First we observe that $|\{a, a^2, a^3, \ldots, a^{n-1}, a^n\ (= e)\}| = n$. Otherwise, we have $a^u = a^v$ for positive integers u, v where $1 \leq u < v \leq n$, and then $a^{v-u} = e$ with $0 < v - u < n$. This, however, contradicts the minimality of n. So now we know that $|\langle a \rangle| \geq n$. But for each $k \in \mathbf{Z}$, it follows from the division algorithm that $k = qn + r$, where $0 \leq r < n$, and so $a^k = a^{qn+r} = (a^n)^q(a^r) = (e^q)(a^r) = a^r \in \{a, a^2, a^3, \ldots, a^{n-1}, a^n\ (= e = a^0)\}$. Therefore, $\langle a \rangle = \{a, a^2, a^3, \ldots, a^{n-1}, a^n\ (= e)\}$ and we can also define $\mathcal{O}(a)$ as the *smallest positive integer n* for which $a^n = e$. This alternative definition for the order of a group element (of finite order) proves to be of value in the following theorem.

THEOREM 16.6 Let $a \in G$ with $\mathcal{O}(a) = n$. If $k \in \mathbf{Z}$ and $a^k = e$, then $n|k$.

Proof: By the division algorithm (again), we have $k = qn + r$, for $0 \leq r < n$, and so it follows that $e = a^k = a^{qn+r} = (a^n)^q(a^r) = (e^q)(a^r) = a^r$. If $0 < r < n$, we contradict the definition of n as $\mathcal{O}(a)$. Hence $r = 0$ and $k = qn$.

We now examine some further results on cyclic groups. The next example helps us to motivate part (b) of Theorem 16.7.

EXAMPLE 16.14 It is known from part (b) of Example 16.13 that $U_9 = \{1, 2, 4, 5, 7, 8\} = \langle 2 \rangle$. We use this fact to define the function $f: U_9 \to (\mathbf{Z}_6, +)$ as follows:

$$f(1) = [0] \qquad f(2) = [1] \qquad f(4) = [2]$$
$$f(5) = f(2^5) = [5] \qquad f(7) = f(2^4) = [4] \qquad f(8) = f(2^3) = [3].$$

So, in general, for each $a \in U_9$ we write $a = 2^k$, for some $0 \leq k \leq 5$, and have $f(a) = f(2^k) = [k]$. This function f is one-to-one and onto and we find, for example, that $f(2 \cdot 5) = f(1) = [0] = [1] + [5] = f(2) + f(5)$, and $f(7 \cdot 8) = f(2) = [1] = [4] + [3] = f(7) + f(8)$.

In general, for a, b in U_9 we may write $a = 2^m$ and $b = 2^n$, where $0 \leq m \leq 5$ and $0 \leq n \leq 5$. It then follows that

$$f(a \cdot b) = f(2^m \cdot 2^n) = f(2^{m+n}) = [m + n] = [m] + [n] = f(a) + f(b).$$

Consequently, the function f is an isomorphism and the groups U_9 and $(\mathbf{Z}_6, +)$ are isomorphic.

[Note how the function f links the generators of the two cyclic groups. Also note that the function $g: U_9 \to (\mathbf{Z}_6, +)$ where

$$g(1) = [0] \qquad g(5) = [1] \qquad g(7) = g(5^2) = [2]$$
$$g(8) = g(5^3) = [3] \qquad g(4) = g(5^4) = [4] \qquad g(2) = g(5^5) = [5]$$

is another isomorphism between these two cyclic groups.]

THEOREM 16.7

Let G be a cyclic group.

a) If $|G|$ is infinite, then G is isomorphic to $(\mathbf{Z}, +)$.

b) If $|G| = n$, where $n > 1$, then G is isomorphic to $(\mathbf{Z}_n, +)$.

Proof:

a) For $G = \langle a \rangle = \{a^k | k \in \mathbf{Z}\}$, let $f: G \to \mathbf{Z}$ be defined by $f(a^k) = k$. (Could we have $a^k = a^t$ with $k \neq t$? If so, f would not be a function.) For a^m, $a^n \in G$, $f(a^m \cdot a^n) = f(a^{m+n}) = m + n = f(a^m) + f(a^n)$, so f is a homomorphism. We leave to the reader the verification that f is one-to-one and onto.

b) If $G = \langle a \rangle = \{a, a^2, \ldots, a^{n-1}, a^n = e\}$, then the function $f: G \to \mathbf{Z}_n$ defined by $f(a^k) = [k]$ is an isomorphism. (Verify this.)

EXAMPLE 16.15

If $G = \langle g \rangle$, G is abelian because $g^m \cdot g^n = g^{m+n} = g^{n+m} = g^n \cdot g^m$ for all $m, n \in \mathbf{Z}$. The converse, however, is false. The group H of Table 16.7 is abelian, and $\mathcal{O}(e) = 1$, $\mathcal{O}(a) = \mathcal{O}(b) = \mathcal{O}(c) = 2$. Since no element of H has order 4, H cannot be cyclic. (The group H is the smallest noncyclic group and is known as the *Klein Four* group.)

Table 16.7

$\cdot$	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Our last result concerns the structure of subgroups in a cyclic group.

THEOREM 16.8 Every subgroup of a cyclic group is cyclic.

Proof: Let $G = \langle a \rangle$. If H is a subgroup of G, each element of H has the form a^k, for some $k \in \mathbf{Z}$. For $H \neq \{e\}$, let t be the smallest positive integer such that $a^t \in H$. (How do we know such an integer t exists?) We claim that $H = \langle a^t \rangle$. Since $a^t \in H$, by the closure property for the subgroup H, $\langle a^t \rangle \subseteq H$. For the opposite inclusion, let $b \in H$, with $b = a^s$, for some $s \in \mathbf{Z}$. By the division algorithm, $s = qt + r$, where $q, r \in \mathbf{Z}$ and $0 \leq r < t$. Consequently, $a^s = a^{qt+r}$ and so $a^r = a^{-qt}a^s = (a^t)^{-q}b$. H is a subgroup of G, so $a^t \in H \Rightarrow (a^t)^{-q} \in H$. Then with $(a^t)^{-q}, b \in H$, it follows that $a^r = (a^t)^{-q}b \in H$. But if $a^r \in H$ with $r > 0$, then we contradict the minimality of t. Hence $r = 0$ and $b = a^{qt} = (a^t)^q \in \langle a^t \rangle$, so $H = \langle a^t \rangle$, a cyclic group.

EXERCISES 16.2

1. Prove parts (b) and (c) of Theorem 16.5.

2. Let $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

a) Determine A^2, A^3, and A^4.

b) Verify that $\{A, A^2, A^3, A^4\}$ is an abelian group under ordinary matrix multiplication.

c) Prove that the group in part (b) is isomorphic to the group shown in Table 16.6.

3. If $G = (\mathbf{Z}_6, +)$, $H = (\mathbf{Z}_3, +)$, and $K = (\mathbf{Z}_2, +)$, find an isomorphism for the groups $H \times K$ and G.

4. Let $f: G \to H$ be a group homomorphism onto H. If G is abelian, prove that H is abelian.

5. Let $(\mathbf{Z} \times \mathbf{Z}, \oplus)$ be the abelian group where $(a, b) \oplus (c, d) = (a + c, b + d)$—here $a + c$ and $b + d$ are computed using ordinary addition in $\mathbf{Z}$—and let $(G, +)$ be an additive group. If $f: \mathbf{Z} \times \mathbf{Z} \to G$ is a group homomorphism where $f(1, 3) = g_1$ and $f(3, 7) = g_2$, express $f(4, 6)$ in terms of g_1 and g_2.

6. Let $f: (\mathbf{Z} \times \mathbf{Z}, \oplus) \to (\mathbf{Z}, +)$ be the function defined by $f(x, y) = x - y$. [Here $(\mathbf{Z} \times \mathbf{Z}, \oplus)$ is the same group as in Exercise 5, and $(\mathbf{Z}, +)$ is the group of integers under ordinary addition.]

a) Prove that f is a homomorphism onto $\mathbf{Z}$.

b) Determine all $(a, b) \in \mathbf{Z} \times \mathbf{Z}$ with $f(a, b) = 0$.

c) Find $f^{-1}(7)$.

d) If $E = \{2n | n \in \mathbf{Z}\}$, what is $f^{-1}(E)$?

7. Find the order of each element in the group of rigid motions of (a) the equilateral triangle; and (b) the square.

8. In S_5 find an element of order n, for all $2 \leq n \leq 5$. Also determine the (cyclic) subgroup of S_5 that each of these elements generates.

9. a) Find all the elements of order 10 in $(\mathbf{Z}_{40}, +)$.

b) Let $G = \langle a \rangle$ be a cyclic group of order 40. Which elements of G have order 10?

10. a) Determine U_{14}, the group of units for the ring $(\mathbf{Z}_{14}, +, \cdot)$.

b) Show that U_{14} is cyclic and find all of its generators.

11. Verify that $(\mathbf{Z}_p^*, \cdot)$ is cyclic for the primes 5, 7, and 11.

12. For a group G, prove that the function $f: G \to G$ defined by $f(a) = a^{-1}$ is an isomorphism if and only if G is abelian.

13. If $f: G \to H$, $g: H \to K$ are homomorphisms, prove that the composite function $g \circ f: G \to K$, where $(g \circ f)(x) = g(f(x))$, is a homomorphism.

14. For $\omega = (1/\sqrt{2})(1 + i)$, let G be the multiplicative group $\{\omega^n | n \in \mathbf{Z}^+, 1 \leq n \leq 8\}$.

a) Show that G is cyclic and find each element $x \in G$ such that $\langle x \rangle = G$.

b) Prove that G is isomorphic to the group $(\mathbf{Z}_8, +)$.

15. a) Find all generators of the cyclic groups $(\mathbf{Z}_{12}, +)$, $(\mathbf{Z}_{16}, +)$, and $(\mathbf{Z}_{24}, +)$.

b) Let $G = \langle a \rangle$ with $\mathbb{o}(a) = n$. Prove that a^k, $k \in \mathbf{Z}^+$, generates G if and only if k and n are relatively prime.

c) If G is a cyclic group of order n, how many distinct generators does it have?

16. Let $f: G \to H$ be a group homomorphism. If $a \in G$ with $\mathbb{o}(a) = n$, and $\mathbb{o}(f(a)) = k$ (in H), prove that $k | n$.

16.3 Cosets and Lagrange's Theorem

In the last two sections, for all finite groups G and subgroups H of G, we had $|H|$ dividing $|G|$. In this section we'll see that this was not mere chance but is true in general. To prove this we need one new idea.

Definition 16.8 If H is a subgroup of G, then for each $a \in G$, the set $aH = \{ah | h \in H\}$ is called a *left coset* of H in G. The set $Ha = \{ha | h \in H\}$ is a *right coset* of H in G.

If the operation in G is addition, we write $a + H$ in place of aH, where $a + H = \{a + h | h \in H\}$.

When the term *coset* is used in this chapter, it will refer to a left coset. For abelian groups there is no need to distinguish between left and right cosets. However, at the end of the next example we'll see that this is not the case for nonabelian groups.

EXAMPLE 16.16 If G is the group of Example 16.7 and $H = \{\pi_0, \pi_1, \pi_2\}$, the coset $r_1H = \{r_1\pi_0, r_1\pi_1, r_1\pi_2\} = \{r_1, r_2, r_3\}$. Likewise we have $r_2H = r_3H = \{r_1, r_2, r_3\}$, whereas $\pi_0H = \pi_1H = \pi_2H = H$.

We see that $|\alpha H| = |H|$ for each $\alpha \in G$ and that $G = H \cup r_1H$ is a partition of G.

For the subgroup $K = \{\pi_0, r_1\}$, we find $r_2K = \{r_2, \pi_2\}$ and $r_3K = \{r_3, \pi_1\}$. Again a partition of G arises: $G = K \cup r_2K \cup r_3K$. (*Note*: $Kr_2 = \{\pi_0r_2, r_1r_2\} = \{r_2, \pi_1\} \neq r_2K$.)

EXAMPLE 16.17 For $G = (\mathbf{Z}_{12}, +)$ and $H = \{[0], [4], [8]\}$, we find that

$$[0] + H = \{[0], [4], [8]\} = [4] + H = [8] + H = H$$
$$[1] + H = \{[1], [5], [9]\} = [5] + H = [9] + H$$
$$[2] + H = \{[2], [6], [10]\} = [6] + H = [10] + H$$
$$[3] + H = \{[3], [7], [11]\} = [7] + H = [11] + H,$$

and $H \cup ([1] + H) \cup ([2] + H) \cup ([3] + H)$ is a partition of G.

These examples now prepare us for the following results.

LEMMA 16.1 If H is a subgroup of the finite group G, then for all $a, b \in G$, (a) $|aH| = |H|$; and (b) either $aH = bH$ or $aH \cap bH = \emptyset$.

Proof:

a) Since $aH = \{ah | h \in H\}$, it follows that $|aH| \leq |H|$. If $|aH| < |H|$, we have $ah_i = ah_j$ with h_i, h_j distinct elements of H. By left-cancellation in G we then get the contradiction $h_i = h_j$, so $|aH| = |H|$.

b) If $aH \cap bH \neq \emptyset$, let $c = ah_1 = bh_2$, for some $h_1, h_2 \in H$. If $x \in aH$, then $x = ah$ for some $h \in H$, and so $x = (bh_2h_1^{-1})h = b(h_2h_1^{-1}h) \in bH$, and $aH \subseteq bH$. Similarly, $y \in bH \Rightarrow y = bh_3$, for some $h_3 \in H \Rightarrow y = (ah_1h_2^{-1})h_3 = a(h_1h_2^{-1}h_3) \in aH$, so $bH \subseteq aH$. Therefore aH and bH are either disjoint or identical.

We observe that if $g \in G$, then $g \in gH$ because $e \in H$. Also, by part (b) of Lemma 16.1, G can be partitioned into mutually disjoint cosets.

At this point we are ready to prove the main result of this section.

THEOREM 16.9 *Lagrange's Theorem.* If G is a finite group of order n with H a subgroup of order m, then m divides n.

Proof: If $H = G$ the result follows. Otherwise $m < n$ and there exists an element $a \in G - H$. Since $a \notin H$, it follows that $aH \neq H$, so $aH \cap H = \emptyset$. If $G = aH \cup H$, then $|G| = |aH| + |H| = 2|H|$ and the theorem follows. If not, there is an element $b \in G - (H \cup aH)$, with $bH \cap H = \emptyset = bH \cap aH$ and $|bH| = |H|$. If $G = bH \cup aH \cup H$, we have $|G| = 3|H|$. Otherwise we're back to an element $c \in G$ with $c \notin bH \cup aH \cup H$. The group G is finite, so this process terminates and we find that $G = a_1H \cup a_2H \cup \cdots \cup a_kH$. Therefore, $|G| = k|H|$ and m divides n.

An alternative method for proving this theorem is given in Exercise 12 for this section.

We close with the statements of two corollaries. Their proofs are requested in the Section Exercises.

COROLLARY 16.1 If G is a finite group and $a \in G$, then $\mathcal{O}(a)$ divides $|G|$.

COROLLARY 16.2 Every group of prime order is cyclic.

EXERCISES 16.3

1. Let $G = S_4$. (a) For $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$, find the subgroup $H = \langle \alpha \rangle$. (b) Determine the left cosets of H in G.

2. Answer Exercise 1 for the case where α is replaced by $\beta = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 4 \end{pmatrix}$.

3. If $\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \in S_4$, how many cosets does $\langle \gamma \rangle$ determine?

4. For $G = (\mathbf{Z}_{24}, +)$, find the cosets determined by the subgroup $H = \langle [3] \rangle$. Do likewise for the subgroup $K = \langle [4] \rangle$.

5. Let G be a group with subgroups H and K. If $|G| = 660$, $|K| = 66$, and $K \subset H \subset G$, what are the possible values for $|H|$?

6. Let R be a ring with unity u. Prove that the units of R form a group under the multiplication of the ring.

7. Let $G = S_4$, the symmetric group on four symbols, and let H be the subset of G where

$$H = \left\{ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \right\}.$$

a) Construct a table to show that H is an abelian subgroup of G.

b) How many left cosets of H are there in G?

c) Consider the group $(\mathbf{Z}_2 \times \mathbf{Z}_2, \oplus)$ where $(a, b) \oplus (c, d) = (a + c, b + d)$ — and the sums $a + c$, $b + d$ are computed using addition modulo 2. Prove that H is isomorphic to this group.

8. If G is a group of order n and $a \in G$, prove that $a^n = e$.

9. Let p be a prime. (a) If G has order $2p$, prove that every proper subgroup of G is cyclic. (b) If G has order p^2, prove that G has a subgroup of order p.

10. Prove Corollaries 16.1 and 16.2.

11. Let H and K be subgroups of a group G, where e is the identity of G.

a) Prove that if $|H| = 10$ and $|K| = 21$, then $H \cap K = \{e\}$.

b) If $|H| = m$ and $|K| = n$, with $\gcd(m, n) = 1$, prove that $H \cap K = \{e\}$.

12. The following provides an alternative way to establish Lagrange's Theorem. Let G be a group of order n, and let H be a subgroup of G of order m.

a) Define the relation $\mathcal{R}$ on G as follows: If $a, b \in G$, then $a \mathcal{R} b$ if $a^{-1}b \in H$. Prove that $\mathcal{R}$ is an equivalence relation on G.

b) For $a, b \in G$, prove that $a \mathcal{R} b$ if and only if $aH = bH$.

c) If $a \in G$, prove that $[a]$, the equivalence class of a under $\mathcal{R}$, satisfies $[a] = aH$.

d) For each $a \in G$, prove that $|aH| = |H|$.

e) Now establish the conclusion of Lagrange's Theorem, namely that $|H|$ divides $|G|$.

13. a) *Fermat's Theorem.* If p is a prime, prove that $a^p \equiv a \pmod{p}$ for each $a \in \mathbf{Z}$. [How is this related to Exercise 22(a) of Section 14.3?]

b) *Euler's Theorem.* For each $n \in \mathbf{Z}^+$, $n > 1$, and each $a \in \mathbf{Z}$, prove that if $\gcd(a, n) = 1$, then $a^{\phi(n)} \equiv 1 \pmod{n}$.

c) How are the theorems in parts (a) and (b) related?

d) Is there any connection between these two theorems and the results in Exercises 6 and 8?

16.4 The RSA Cryptosystem (Optional)

This section provides us with an opportunity to use some of the theoretical ideas we encountered in Sections 14.3 and 16.3 in a more contemporary application.

In Example 14.15 of Section 14.3 we introduced two private-key cryptosystems: the cipher shift and the affine cipher. For an alphabet of m characters, the encryption function $E: \mathbf{Z}_m \to \mathbf{Z}_m$, for the cipher-shift cryptosystem, is given by $E(\theta) = (\theta + \kappa)$ **mod** m, where $\theta, \kappa \in \mathbf{Z}_m$, for κ ($\neq 0$) fixed. (Using $\kappa = 0$ would not alter any of the characters in a message.) Consequently, there are $m - 1$ possibilities to examine in an attempt to discover the value of the key κ. Further, once we know the value of κ, we also know the decryption function $D: \mathbf{Z}_m \to \mathbf{Z}_m$, for $D(\theta) = (\theta - \kappa)$ **mod** m. In the case of the affine-cipher cryptosystem (also with an alphabet of m characters) the encryption function $E: \mathbf{Z}_m \to \mathbf{Z}_m$ is now given by $E(\theta) = (\alpha\theta + \kappa)$ **mod** m, where $\theta, \alpha, \kappa \in \mathbf{Z}_m$, for fixed α, κ, with α invertible in $\mathbf{Z}_m$ [or, equivalently, with $\gcd(\alpha, m) = 1$]. Here the decryption function $D: \mathbf{Z}_m \to \mathbf{Z}_m$ is given by $D(\theta) = [\alpha^{-1}(\theta - \kappa)]$ **mod** m. Without prior knowledge of the key (α, κ), now one would have to check $m\phi(m)$ possibilities to discover the appropriate values of α and κ for this private-key cryptosystem.

The security of either of the above cryptosystems depends on having the key [be it κ or (α, κ)] known only to the sender and the recipient of the messages.

The RSA cryptosystem is an example of a *public-key* cryptosystem. This cryptosystem was developed in the 1970s (and patented in 1983) by Ronald Rivest (1948–), Adi Shamir (1952–), and Leonard Adleman (1945–). (Taking the first letter from the surname of each of these three men provides the adjective RSA.)

We shall describe how this cryptosystem works and provide an example for encryption and decryption. In so doing, we shall find ourselves using some of the results from Sections 14.3 and 16.3.

EXAMPLE 16.18

As with the two private-key cryptosystems, once again we have an alphabet of m characters. We start with two distinct primes p, q. In practice, these should be large primes—each with 100 or more digits. (However, for our example we shall use much smaller primes.) After selecting the primes p, q, we then consider the integers $n = pq$ and $r = (p - 1) \cdot (q - 1) = \phi(p)\phi(q) = \phi(pq) = \phi(n)$, and, at this point, we choose an invertible element e in $\mathbf{Z}_r = (\mathbf{Z}_{\phi(n)})$.

[Here, if the element e is chosen at random, then the only time we fail to obtain an invertible element is when the element chosen is a multiple of p (there are q possibilities) or a multiple of q (there are p possibilities). In this count of $p + q$ elements we have accounted for pq twice, so there are only $p + q - 1$ possibilities for failure. Hence, the probability for

failure is $(p + q - 1)/(pq) = (1/q) + (1/p) - (1/(pq))$, a very small number if p and q each have 100 or more digits.]

For instance, consider $p = 61$, $q = 127$, with $n = (61)(127) = 7747$ and $r = \phi(61) \cdot \phi(127) = (60)(126) = 7560$. Now suppose we select e as 17.

Consider the following message that we wish to encrypt.

INVEST IN BONDS

Using the same plaintext assignments as in part (b) of Example 14.15, here we would replace the letter "I" by 08 (not merely 8). Then we replace "N" by 13. This provides us with the first block of four digits — namely, 0813 — for the first two letters "IN". The assignment for the complete message is as follows [where we have appended the letter "X" to the right end, in order for the final block to have two letters (or, four digits)]:

I	N	V	E	S	T	I	N	B	O	N	D	S	X
08	13	21	04	18	19	08	13	01	14	13	03	18	23

We now encrypt each block B of four digits by the encryption function E, where $E(B) = B^e$ **mod** n. (This modular exponentiation can be carried out efficiently by using the procedure in Example 14.16.) So here the domain of E is the concatenation of $\mathbf{Z}_{26}$ with itself, and we find that

0813^{17} **mod** $7747 = 2169$ $\quad$ 2104^{17} **mod** $7747 = 0628$ $\quad$ 1819^{17} **mod** $7747 = 5540$

0813^{17} **mod** $7747 = 2169$ $\quad$ 0114^{17} **mod** $7747 = 6560$ $\quad$ 1303^{17} **mod** $7747 = 6401$

1823^{17} **mod** $7747 = 4829$.

Consequently, the recipient of the encrypted assignment (for the given plaintext message) receives the ciphertext

2169 0628 5540 2169 6560 6401 4829.

Now the question is: "How does the recipient decrypt the ciphertext received?"

Since e is a unit in $\mathbf{Z}_r$ $(= \mathbf{Z}_{\phi(n)})$, we can use the Euclidean algorithm (as in Example 14.13) to compute $e^{-1} = d$. Then we define the decryption function D, where $D(C) = C^d$ **mod** n, for a block C of four digits. Since $e^{-1} = d$, it follows that $ed \equiv 1 \bmod \phi(n)$ — that is, ed **mod** $\phi(n) = 1$. Therefore, $ed = k\phi(n) + 1$, for some $k \in \mathbf{Z}$. Now recall the argument given earlier for the probability that a randomly selected element e from $\mathbf{Z}_n$ is invertible (or a unit in $\mathbf{Z}_n$). For any block B of four digits, we consider B as an element of $\mathbf{Z}_n$ — in fact, we consider B as a unit in $\mathbf{Z}_n$. Since the units in the ring $(\mathbf{Z}_n, +, \cdot)$ form a group of order $\phi(n)$ under multiplication, it follows from the result in Exercise 8 of Section 16.3 that $B^{ed} = B^{k\phi(n)+1} = (B^{\phi(n)})^k B^1 \equiv B \pmod{n}$, or B^{ed} **mod** $n = B$. [This is also a consequence of Euler's Theorem, as stated in part (b) of Exercise 13 in Section 16.3.]

Applying the result from the previous paragraph in our example we have $p = 61$, $q = 127$, $n = pq = 7747$, $r = \phi(n) = (p - 1)(q - 1) = (60)(126) = 7560$, and $e = 17$. From the Euclidean algorithm we calculate $d = e^{-1} = 3113$. Now we find, for instance, that 2169^{3113} **mod** $7747 = 0813$ and that 0628^{3113} **mod** $7747 = 2104$. Continuing, the recipient determines the numeric assignment for the original plaintext and then the plaintext.

Now what makes the RSA cryptosystem more secure than the private-key cryptosystems we studied? First, we should relate that the RSA cryptosystem is *not* a private-key cryptosystem. This system is an example of a *public-key* cryptosystem, where the key (n, e) is made public. So it seems that all one needs to do to decrypt the encrypted assignment is

to determine $d = e^{-1}$ in $\mathbf{Z}_r$ $(= \mathbf{Z}_{\phi(n)})$. Now it is time to realize that by knowing n we do not immediately know r. For to be able to determine $r = (p-1)(q-1)$, we need to know p, q, the prime factors of n. And this is what makes this system so much more secure than the other cryptosystems we mentioned. Determining the primes p, q, when they are 100 or more digits long, is not a feasible problem. However, as computer power continues to improve, to keep the RSA cryptosystem secure, one may need to redefine the key using primes with more and more digits.

In closing, we show how the problem of factoring the modulus n as pq is related to the problem of determining $r = (p-1)(q-1)$. We start by observing that

$$p + q = pq - (p-1)(q-1) + 1 = n - \phi(n) + 1 = n - r + 1,$$

while

$$\begin{aligned} p - q &= \sqrt{(p-q)^2} = \sqrt{(p-q)^2 + 4pq - 4pq} = \sqrt{(p+q)^2 - 4pq} \\ &= \sqrt{(p+q)^2 - 4n} = \sqrt{(n-r+1)^2 - 4n}. \end{aligned}$$

Then, from these two equations, we learn that

$$p = (1/2)[(p+q) + (p-q)] = (1/2)[(n-r+1) + \sqrt{(n-r+1)^2 - 4n}]$$

and

$$q = (1/2)[(p+q) - (p-q)] = (1/2)[(n-r+1) - \sqrt{(n-r+1)^2 - 4n}].$$

Consequently, when we know n and r, then we can readily determine the primes p, q such that $n = pq$.

EXERCISES 16.4

The use of a computer algebra system is strongly recommended for the first four exercises.

1. Determine the ciphertext for the plaintext INVEST IN STOCKS, when using RSA encryption with $e = 7$ and $n = 2573$.

2. Determine the ciphertext for the plaintext ORDER A PIZZA, when using RSA encryption with $e = 5$ and $n = 1459$.

3. Determine the plaintext for the RSA ciphertext 1418 1436 2370 1102 1805 0250, if $e = 11$ and $n = 2501$.

4. Determine the plaintext for the RSA ciphertext 0986 3029 1134 1105 1232 2281 2967 0272 1818 2398 1153, if $e = 17$ and $n = 3053$.

5. Find the primes p, q if $n = pq = 121{,}361$ and $\phi(n) = 120{,}432$.

6. Find the primes p, q if $n = pq = 5{,}446{,}367$ and $\phi(n) = 5{,}441{,}640$.

16.5 Elements of Coding Theory

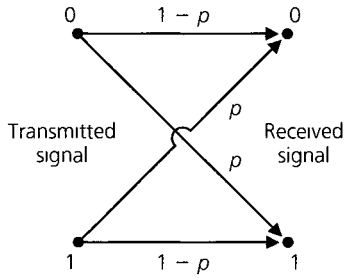
In this and the next four sections we introduce an area of applied mathematics called *algebraic coding theory*. This theory was inspired by the fundamental paper of Claude Shannon (1948) along with results by Marcel Golay (1949) and Richard Hamming (1950). Since that time it has become an area of great interest where algebraic structures, probability, and combinatorics all play a role.

Our coverage will be held to an introductory level as we seek to model the transmission of information represented by strings of the signals 0 and 1.

In digital communications, when information is transmitted in the form of strings of 0's and 1's, certain problems arise. As a result of "noise" in the channel, when a certain signal is transmitted a different signal may be received, thus causing the receiver to make a wrong

decision. Hence we want to develop techniques to help us detect, and perhaps even correct, transmission errors. However, we can only improve the chances of correct transmission; there are no guarantees.

Our model uses a *binary symmetric channel*, as shown in Fig. 16.2. The adjective *binary* appears because an individual signal is represented by one of the bits 0 or 1. When a transmitter sends the signal 0 or 1 in such a channel, associated with either signal is a (constant) probability p for incorrect transmission. When that probability p is the same for both signals, the channel is called *symmetric*. Here, for example, we have probability p of sending 0 and having 1 received. The probability of sending signal 0 and having it received correctly is then $1 - p$. All possibilities are illustrated in Fig. 16.2.



The Binary Symmetric Channel

Figure 16.2

EXAMPLE 16.19

Consider the string $c = 10110$. We regard c as an element of the group $\mathbf{Z}_2^5$, formed from the direct product of five copies of $(\mathbf{Z}_2, +)$. To shorten notation we write 10110 instead of $(1, 0, 1, 1, 0)$. When sending each bit (individual signal) of c through the binary symmetric channel, we assume that the probability of incorrect transmission is $p = 0.05$, so that the probability of transmitting c with no errors is $(0.95)^5 \doteq 0.77$.

Here, and throughout our discussion of coding theory, we assume that the transmission of each signal does not depend in any way on the transmissions of prior signals. Consequently, the probability of the occurrence of all of these *independent* events (in their prescribed order) is given by the product of their individual probabilities.

What is the probability that the party receiving the five-bit message receives the string $r = 00110$ — that is, the original message with an error in the first position? The probability of incorrect transmission for the first bit is 0.05, so with the assumption of independent events, $(0.05)(0.95)^4 \doteq 0.041$ is the probability of sending $c = 10110$ and receiving $r = 00110$. With $e = 10000$, we can write $c + e = r$ and interpret r as the result of the sum of the original message c and the particular *error pattern* $e = 10000$. Since $c, r, e \in \mathbf{Z}_2^5$ and $-1 = 1$ in $\mathbf{Z}_2$, we also have $c + r = e$ and $r + e = c$.

In transmitting $c = 10110$, the probability of receiving $r = 00100$ is

$$(0.05)(0.95)^2(0.05)(0.95) \doteq 0.002,$$

so this multiple error is not very likely to occur.

Finally if we transmit $c = 10110$, what is the probability that r differs from c in exactly two places? To answer this we sum the probabilities for each error pattern consisting of two 1's and three 0's. Each such pattern has probability 0.002. There are $\binom{5}{2}$ such patterns, so

the probability of two errors in transmission is given by

$$\binom{5}{2}(0.05)^2(0.95)^3 \doteq 0.021.$$

These results lead us to the following theorem.

THEOREM 16.10 Let $c \in \mathbf{Z}_2^n$. For the transmission of c through a binary symmetric channel with probability p of incorrect transmission,

a) the probability of receiving $r = c + e$, where e is a *particular* error pattern consisting of k 1's and $(n - k)$ 0's, is $p^k(1 - p)^{n-k}$.

b) the probability that (exactly) k errors are made in the transmission is

$$\binom{n}{k}p^k(1 - p)^{n-k}.^{\dagger}$$

In Example 16.19, the probability of making at most one error in the transmission of $c = 10110$ is $(0.95)^5 + \binom{5}{1}(0.05)(0.95)^4 \doteq 0.977$. Thus the chance for multiple errors in transmission will be considered negligible throughout the discussion in this chapter. Such an assumption is valid when p is small. In actuality, a binary symmetric channel is considered "good" when $p < 10^{-5}$. However, no matter what else we stipulate, we always want $p < 1/2$.

To improve the accuracy of transmission in a binary symmetric channel, certain types of coding schemes can be used where extra bits are provided.

For $m, n \in \mathbf{Z}^+$, let $n > m$. Consider $\emptyset \neq W \subseteq \mathbf{Z}_2^m$. The set W consists of the *messages* to be transmitted. To each $w \in W$ are appended $n - m$ extra bits to form the *code word* c, where $c \in \mathbf{Z}_2^n$. This process is called *encoding* and is represented by the function $E: W \to \mathbf{Z}_2^n$. Then $E(w) = c$ and $E(W) = C \subseteq \mathbf{Z}_2^n$. Since the function E simply appends extra bits to the (distinct) messages, the encoding process is one-to-one. Upon transmission, c is received as $T(c)$, where $T(c) \in \mathbf{Z}_2^n$. Unfortunately, T is not a function because $T(c)$ may be different at different transmission times (for the noise in the channel changes with time). (See Fig. 16.3.)

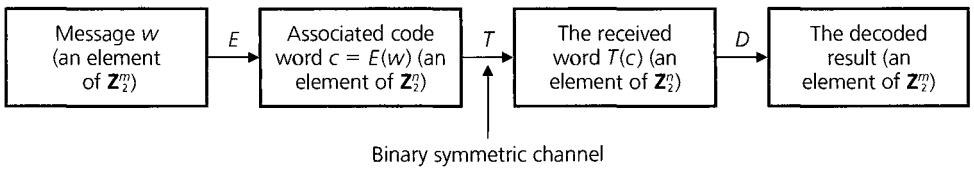


Figure 16.3

Upon receiving $T(c)$, we want to apply a decoding function $D: \mathbf{Z}_2^n \to \mathbf{Z}_2^m$ to remove the extra bits and, we hope, obtain the original message w. Ideally $D \circ T \circ E$ should be the identity function on W, with $D: C \to W$. Since this cannot be expected, we seek functions E and D such that there is a high probability of correctly decoding the received word $T(c)$ and recapturing the original message w. In addition, we want the ratio m/n to be as large as possible so that an excessive number of bits are not appended to w in getting the code

†This is the binomial probability distribution that was developed in (optional) Sections 3.5 and 3.7.

word $c = E(w)$. This ratio m/n measures the *efficiency* of our scheme and is called the *rate* of the code. Finally, the functions E and D should be more than theoretical results; they must be practical in the sense that they can be implemented electronically.

In such a scheme, the functions E and D are called the *encoding* and *decoding* functions, respectively, of an (n, m) *block code*.

We illustrate these ideas in the following two examples.

EXAMPLE 16.20

Consider the $(m + 1, m)$ block code for $m = 8$. Let $W = \mathbf{Z}_2^8$. For each $w = w_1 w_2 \cdots w_8 \in W$, define $E: \mathbf{Z}_2^8 \to \mathbf{Z}_2^9$ by $E(w) = w_1 w_2 \cdots w_8 w_9$, where $w_9 = \sum_{i=1}^{8} w_i$, with the addition performed modulo 2. For example, $E(11001101) = 110011011$, and $E(00110011) = 001100110$.

For all $w \in \mathbf{Z}_2^8$, $E(w)$ contains an even number of 1's. So for $w = 11010110$ and $E(w) = 110101101$, if we receive $T(c) = T(E(w))$ as 100101101, from the odd number of 1's in $T(c)$ we know that a mistake has occurred in transmission. Hence we are able to *detect* single errors in transmission. But we seem to have no way to correct such errors.

The probability of sending the code word 110101101 and making at most one error in transmission is

$$\underbrace{(1-p)^9}_{\substack{\text{All nine bits are}\\ \text{correctly transmitted.}}} + \underbrace{\tbinom{9}{1}p(1-p)^8}_{\substack{\text{One bit is changed in}\\ \text{transmission and an error is detected.}}}.$$

For $p = 0.001$ this gives $(0.999)^9 + \binom{9}{1}(0.001)(0.999)^8 \doteq 0.99996417$.

If we detect an error and we are able to relay a signal back to the transmitter to repeat the transmission of the code word, and continue this process until the received word has an even number of 1's, then the probability of sending the code word 110101101 and receiving the correct transmission is approximately 0.99996393.†

Should an even positive number of errors occur in transmission, $T(c)$ is unfortunately accepted as the correct code word and we interpret its first eight components as the original message. This scheme is called the $(m + 1, m)$ *parity-check code* and is appropriate only when multiple errors are not likely to occur.

If we send the message 11010110 through the channel, we have probability $(0.999)^8 = 0.99202794$ of correct transmission. By using this parity-check code, we increase our chances of getting the correct message to (approximately) 0.99996393. However, an extra signal is sent (and perhaps additional transmissions are needed) and the rate of the code has decreased from 1 to 8/9.

But suppose that instead of sending eight bits we sent 160 bits, in successive strings of length 8. The chances of receiving the correct message without any coding scheme would be

---

†For $p = 0.001$ the probability that an odd number of errors occurs in the transmission of the code word 110101101 is

$$p_{\text{odd}} = \tbinom{9}{1}(0.999)^8(0.001) + \tbinom{9}{3}(0.999)^6(0.001)^3 + \tbinom{9}{5}(0.999)^4(0.001)^5 + \tbinom{9}{7}(0.999)^2(0.001)^7 + \tbinom{9}{9}(0.001)^9$$
$$\doteq 0.008928251 + 0.000000083 + 0.000000000 + 0.000000000 + 0.000000000 = 0.008928334.$$

With q = the probability of the correct transmission of $110101101 = (0.999)^9$, the probability that this code word is transmitted and correctly received under these conditions (of retransmission) is then given by

$$q + p_{\text{odd}} \cdot q + (p_{\text{odd}})^2 q + (p_{\text{odd}})^3 q + \cdots = q/(1 - p_{\text{odd}}) \doteq 0.99996393 \text{ (to eight decimal places).}$$

$(0.999)^{160} \doteq 0.85207557$. With the parity-check method we send 180 bits, but the chances for correct transmission now increase to $(0.999964)^{20} \doteq 0.99928025$.

EXAMPLE 16.21

The $(3m, m)$ *triple repetition code* is one where we can both *detect* and *correct* single errors in transmission. With $m = 8$ and $W = \mathbf{Z}_2^8$, we define $E: \mathbf{Z}_2^8 \to \mathbf{Z}_2^{24}$ by $E(w_1 w_2 \cdots w_7 w_8) = w_1 w_2 \cdots w_8 w_1 w_2 \cdots w_8 w_1 w_2 \cdots w_8$.

Hence if $w = 10110111$, then $c = E(w) = 101101111011011110110111$.

The decoding function $D: \mathbf{Z}_2^{24} \to \mathbf{Z}_2^8$ is carried out by the majority rule. For example, if $T(c) = 101001110011011110110110$, then we have three errors occurring in positions 4, 9, and 24. We decode $T(c)$, by examining the first, ninth, and seventeenth positions to see which signal appears more times. Here it is 1 (which occurs twice), so we decode the first entry in the decoded message as 1. Continuing with the entries in the second, tenth, and eighteenth positions, the result for the second entry of the decoded message is 0 (which occurs all three times). As we proceed, we recapture the correct message, 10110111.

Although we have more than one transmission error here, all is well unless two (or more) errors occur with the second error eight or sixteen spaces after the first — that is, if two (or more) incorrect transmissions occur for the same bit of the original message.

Now how does this scheme compare with the other methods we have? With $p = 0.001$, the probability of correctly decoding a single bit is $(0.999)^3 + \binom{3}{1}(0.001)(0.999)^2 \doteq 0.99999700$. So the probability of receiving and correctly decoding the eight-bit message is $(0.99999700)^8 = 0.99997600$, just slightly better than the result from the parity-check method (where we may have to retransmit, thus increasing the overall transmission time). Here we transmit 24 signals for this message, so our rate is now 1/3. For this increased accuracy and the ability to detect and now *correct* single errors (which we could not do in any previous schemes), we may pay with an increase in transmission time. But we do not waste time with retransmissions.

EXERCISES 16.5

1. Let C be a set of code words, where $C \subseteq \mathbf{Z}_2^7$. In each of the following, two of e (error pattern), r (received word) and c (code word) are given, with $r = c + e$. Determine the third term.

a) $c = 1010110$, $r = 1011111$

b) $c = 1010110$, $e = 0101101$

c) $e = 0101111$, $r = 0000111$

2. A binary symmetric channel has probability $p = 0.05$ of incorrect transmission. If the code word $c = 011011101$ is transmitted, what is the probability that (a) we receive $r = 011111101$? (b) we receive $r = 111011100$? (c) a single error occurs? (d) a double error occurs? (e) a triple error occurs? (f) three errors occur, no two of them consecutive?

3. Let $E: \mathbf{Z}_2^3 \to \mathbf{Z}_2^9$ be the encoding function for the (9, 3) triple repetition code.

a) If $D: \mathbf{Z}_2^9 \to \mathbf{Z}_2^3$ is the corresponding decoding function, apply D to decode the received words (i) 111101100; (ii) 000100011; (iii) 010011111.

b) Find three different received words r for which $D(r) = 000$.

c) For each $w \in \mathbf{Z}_2^3$, what is $|D^{-1}(w)|$?

4. The $(5m, m)$ five-times repetition code has encoding function $E: \mathbf{Z}_2^m \to \mathbf{Z}_2^{5m}$, where $E(w) = wwwww$. Decoding with $D: \mathbf{Z}_2^{5m} \to \mathbf{Z}_2^m$ is accomplished by the majority rule. (Here we are able to correct single and double errors made in transmission.)

a) With $p = 0.05$, what is the probability for the transmission and correct decoding of the signal 0?

b) Answer part (a) for the message 110 in place of the signal 0.

c) For $m = 2$, decode the received word

$$r = 0111001001.$$

d) If $m = 2$, find three received words r where $D(r) = 00$.

e) For $m = 2$ and $D: \mathbf{Z}_2^{10} \to \mathbf{Z}_2^2$, what is $|D^{-1}(w)|$ for each $w \in \mathbf{Z}_2^2$?

16.6 The Hamming Metric

In this section we develop the general principles for discussing the error-detecting and error-correcting capabilities of a coding scheme. These ideas were developed by Richard Wesley Hamming (1915–1998).

We start by considering a code $C \subseteq \mathbf{Z}_2^4$, where $c_1 = 0111$, $c_2 = 1111 \in C$. Now both the transmitter and the receiver know the elements of C. So if the transmitter sends c_1 but the person receiving the code word receives $T(c_1)$ as 1111, then he or she feels that c_2 was transmitted and makes whatever decision (a wrong one) c_2 implies. Consequently, although only one transmission error was made, the results could be unpleasant. Why is this? Unfortunately we have two code words that are almost the same. They are rather *close* to each other, for they differ in only one component.

We describe this notion of closeness more precisely as follows.

Definition 16.9 For each element $x = x_1x_2 \cdots x_n \in \mathbf{Z}_2^n$, where $n \in \mathbf{Z}^+$, the *weight of* x, denoted $\text{wt}(x)$, is the number of components x_i of x, for $1 \leq i \leq n$, where $x_i = 1$. If $y \in \mathbf{Z}_2^n$, the *distance between* x *and* y, denoted $d(x, y)$, is the number of components where $x_i \neq y_i$, for $1 \leq i \leq n$.

EXAMPLE 16.22 For $n = 5$, let $x = 01001$ and $y = 11101$. Then $\text{wt}(x) = 2$, $\text{wt}(y) = 4$, and $d(x, y) = 2$. In addition, $x + y = 10100$, so $\text{wt}(x + y) = 2$. Is it just by chance that $d(x, y) = \text{wt}(x + y)$? For each $1 \leq i \leq 5$, $x_i + y_i$ contributes a count of 1 to $\text{wt}(x + y) \iff x_i \neq y_i \iff x_i, y_i$ contribute a count of 1 to $d(x, y)$. [This is actually true for all $n \in \mathbf{Z}^+$, so $\text{wt}(x + y) = d(x, y)$ for all $x, y \in \mathbf{Z}_2^n$.]

When $x, y \in \mathbf{Z}_2^n$, we write $d(x, y) = \sum_{i=1}^{n} d(x_i, y_i)$ where,

$$\text{for each } 1 \leq i \leq n, \qquad d(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i. \end{cases}$$

LEMMA 16.2 For all $x, y \in \mathbf{Z}_2^n$, $\text{wt}(x + y) \leq \text{wt}(x) + \text{wt}(y)$.

Proof: We prove this lemma by examining, for each $1 \leq i \leq n$, the components $x_i, y_i, x_i + y_i$, of $x, y, x + y$, respectively. Only one situation would cause this inequality to be false: if $x_i + y_i = 1$ while $x_i = 0$ and $y_i = 0$, for some $1 \leq i \leq n$. But this never occurs because $x_i + y_i = 1$ implies that exactly one of x_i and y_i is 1.

In Example 16.22 we found that

$$\text{wt}(x + y) = \text{wt}(10100) = 2 \leq 2 + 4 = \text{wt}(01001) + \text{wt}(11101) = \text{wt}(x) + \text{wt}(y).$$

THEOREM 16.11 The distance function d defined on $\mathbf{Z}_2^n \times \mathbf{Z}_2^n$ satisfies the following for all $x, y, z \in \mathbf{Z}_2^n$.

a) $d(x, y) \geq 0$

b) $d(x, y) = 0 \iff x = y$

c) $d(x, y) = d(y, x)$

d) $d(x, z) \leq d(x, y) + d(y, z)$

Proof: We leave the first three parts for the reader and prove part (d).

In $\mathbf{Z}_2^n$, $y + y = 0$, so $d(x, z) = \text{wt}(x + z) = \text{wt}(x + (y + y) + z) = \text{wt}((x + y) + (y + z)) \leq \text{wt}(x + y) + \text{wt}(y + z)$, by Lemma 16.2. With $\text{wt}(x + y) = d(x, y)$ and $\text{wt}(y + z) = d(y, z)$, the result follows. (This property is generally called the *Triangle Inequality*.)

When a function satisfies the four properties listed in Theorem 16.11, it is called a *distance function* or *metric*, and we call $(\mathbf{Z}_2^n, d)$ a *metric space*. Hence d (as given above) is often referred to as the *Hamming metric*. This metric is used in the following.

Definition 16.10 For $n, k \in \mathbf{Z}^+$ and $x \in \mathbf{Z}_2^n$, the *sphere* of radius k centered at x is defined as $S(x, k) = \{y \in \mathbf{Z}_2^n | \, d(x, y) \leq k\}$.

EXAMPLE 16.23 For $n = 3$ and $x = 110 \in \mathbf{Z}_2^3$, $S(x, 1) = \{110, 010, 100, 111\}$ and $S(x, 2) = \{110, 010, 100, 111, 000, 101, 011\}$.

With these preliminaries in hand we turn now to the two major results of this section.

THEOREM 16.12 Let $E: W \to C$ be an encoding function with the set of messages $W \subseteq \mathbf{Z}_2^m$ and the set of code words $E(W) = C \subseteq \mathbf{Z}_2^n$, where $m < n$. If our objective is error detection, then for $k \in \mathbf{Z}^+$, we can detect all transmission errors of weight $\leq k$ if and only if the minimum distance between code words is at least $k + 1$.

Proof: The set C is known to both the transmitter and the receiver, so if $w \in W$ is the message and $c = E(w)$ is transmitted, let $c \neq T(c) = r$. If the minimum distance between code words is at least $k + 1$, then the transmission of c can result in as many as k errors and r will not be listed in C. Hence we can detect all errors e where $\text{wt}(e) \leq k$. Conversely, let c_1, c_2 be code words with $d(c_1, c_2) < k + 1$. Then $c_2 = c_1 + e$ where $\text{wt}(e) \leq k$. If we send c_1 and $T(c_1) = c_2$, then we would feel that c_2 had been sent, thus failing to detect an error of weight $\leq k$.

What can we say about error-correcting capability?

THEOREM 16.13 Let E, W, and C be as in Theorem 16.12. If our objective is error correction, then for $k \in \mathbf{Z}^+$, we can construct a decoding function $D: \mathbf{Z}_2^n \to W$ that corrects all transmission errors of weight $\leq k$ if and only if the minimum distance between code words is at least $2k + 1$.

Proof: For $c \in C$, consider $S(c, k) = \{x \in \mathbf{Z}_2^n | d(c, x) \leq k\}$. Define $D: \mathbf{Z}_2^n \to W$ as follows. If $r \in \mathbf{Z}_2^n$ and $r \in S(c, k)$ for some code word c, then $D(r) = w$ where $E(w) = c$. [Here c is the (unique) code word *nearest* to r.] If $r \notin S(c, k)$ for any $c \in C$, then we define $D(r) = w_0$, where w_0 is some arbitrary message that remains fixed once it is chosen. The only problem we could face here is that D might not be a function. This will happen if there is an element r in $\mathbf{Z}_2^n$ with r in both $S(c_1, k)$ and $S(c_2, k)$ for distinct code words c_1, c_2. But $r \in S(c_1, k) \Rightarrow d(c_1, r) \leq k$, and $r \in S(c_2, k) \Rightarrow d(c_2, r) \leq k$, so $d(c_1, c_2) \leq d(c_1, r) + d(r, c_2) \leq k + k < 2k + 1$. Consequently, if the minimum distance between code words is at least $2k + 1$, then D is a function, and it will decode all possible

received words, correcting any transmission error of weight $\leq k$. Conversely, if $c_1, c_2 \in C$ and $d(c_1, c_2) \leq 2k$, then c_2 can be obtained from c_1 by making at most $2k$ changes. Starting at code word c_1 we make approximately half (exactly, $\lfloor d(c_1, c_2)/2 \rfloor$) of these changes. This brings us to $r = c_1 + e_1$ with $\text{wt}(e_1) \leq k$. Continuing from r, we make the remaining changes to get to c_2 and find $r + e_2 = c_2$ with $\text{wt}(e_2) \leq k$. But then $r = c_2 + e_2$. Now with $c_1 + e_1 = r = c_2 + e_2$ and $\text{wt}(e_1)$, $\text{wt}(e_2) \leq k$, how can one decide on the code word from which r arises? This ambiguity results in a possible error of weight $\leq k$ that cannot be corrected.

EXAMPLE 16.24

With $W = \mathbf{Z}_2^2$ let $E: W \to \mathbf{Z}_2^6$ be given by

$$E(00) = 000000 \qquad E(10) = 101010 \qquad E(01) = 010101 \qquad E(11) = 111111.$$

Then the minimum distance between code words is 3, so we can correct all single errors.

With

$$\begin{aligned} S(000000, 1) &= \{x \in \mathbf{Z}_2^6 | d(000000, x) \leq 1\} \\ &= \{000000, 100000, 010000, 001000, 000100, 000010, 000001\}, \end{aligned}$$

the decoding function $D: \mathbf{Z}_2^6 \to W$ gives $D(x) = 00$ for all $x \in S(000000, 1)$.

Similarly,

$$\begin{aligned} S(010101, 1) &= \{x \in \mathbf{Z}_2^6 | d(010101, x) \leq 1\} \\ &= \{010101, 110101, 000101, 011101, 010001, 010111, 010100\}, \end{aligned}$$

and here $D(x) = 01$ for each $x \in S(010101, 1)$. At this point our definition of D accounts for 14 of the elements in $\mathbf{Z}_2^6$. Continuing to define D for the 14 elements in $S(101010, 1)$ and $S(111111, 1)$ there remain 36 other elements to account for. We define $D(x) = 00$ (or any other message) for these 36 other elements and have a decoding function that will correct single errors.

Beware! There is a subtle point that needs to be made about Theorems 16.12 and 16.13. For example, if the minimum distance between code words is $2k + 1$ one may feel that we can detect all errors of weight $\leq 2k$ *and* correct all errors of weight $\leq k$. This is not necessarily true. That is, error detection and error correction need not take place at the same time and at the maximum levels. To see this, reconsider the (6, 2)-triple repetition code of Example 16.24. Here the encoding function $E: W(= \mathbf{Z}_2^2) \to \mathbf{Z}_2^6$ is given by $E(w_1 w_2) = w_1 w_2 w_1 w_2 w_1 w_2$ and the code comprises the four elements of $\mathbf{Z}_2^6$ in the range of E. Since the minimum distance between any two elements of $\mathbf{Z}_2^2$ is 1, it follows that the minimum distance between code words is 3 (as observed earlier in Example 16.24).

Now suppose that our major objective is error correction and that $r = 100000\ [\notin E(W)]$ is received. We see that $d(000000, r) = 1$, $d(101010, r) = 2$, $d(010101, r) = 4$, and $d(111111, r) = 5$. Consequently, we should choose to decode r as 000000, the unique code word nearest to r. Unfortunately, suppose that the actual message were 10 (with corresponding code word 101010), but we received $r = 100000$. Upon correcting r as 000000, we should then decode 000000 to get the incorrect message 00. And, in so doing, we have failed to detect an error of weight 2.

In this type of situation one can develop a scheme where a mixed strategy is used. Here both error correction *and* error detection may be carried out at some levels.

For $t \in \mathbf{N}$, if the received word is r and there is a unique code word c_1 such that $d(c_1, r) \leq t$, then we decode r as c_1. (*Note*: The case where $r = c_1$ is covered when $t = 0$.) If there exists a second code word c_2 such that $d(c_2, r) = d(c_1, r)$, or if $d(c, r) > t$ for all code words c, then an error is declared (and retransmission is generally requested). Using this scheme, if the minimum distance between code words is at least $2t + s + 1$, for $s \in \mathbf{N}$, then we can correct all errors of weight $\leq t$ and detect all errors with weights between $t + 1$ and $t + s$, inclusive.

When using this scheme for the (6, 2)-triple repetition code, our options include:

1) $t = 0$; $s = 2$: Here we can detect all errors of weight ≤ 2 but we have no error-correction capability.
2) $t = 1$; $s = 0$: Single errors are corrected here but there is no error-detecting capability.

If we use the (10, 2)-five-times repetition code, then the minimum distance is 5. Applying the above scheme in this case, our options now include:

1) $t = 0$; $s = 4$: Here we can detect all errors of weight ≤ 4 but we have no error-correction capability.
2) $t = 1$; $s = 2$: Now single errors are corrected and we can also detect all errors e, where $2 \leq \text{wt}(e) \leq 3$.
3) $t = 2$; $s = 0$: All errors of weight ≤ 2 are corrected but there is no error-detecting capability.

[For more on this, the interested reader should examine Chapter 4 of the text by S. Roman [24].]

16.7 The Parity-Check and Generator Matrices

In this section we introduce an example where the encoding and decoding functions are given by matrices over $\mathbf{Z}_2$. One of these matrices will help us to locate the *nearest* code word for a given received word. This will be especially helpful as the set C of code words grows larger.

EXAMPLE 16.25

Let

$$G = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

be a 3×6 matrix over $\mathbf{Z}_2$. The first three columns of G form the 3×3 identity matrix I_3. Letting A denote the matrix formed from the last three columns of G, we write $G = [I_3|A]$ to denote its structure. The (partitioned) matrix G is called a *generator matrix*.

We use G to define an encoding function $E: \mathbf{Z}_2^3 \to \mathbf{Z}_2^6$ as follows. For $w \in \mathbf{Z}_2^3$, $E(w) = wG$ is the element in $\mathbf{Z}_2^6$ obtained by multiplying w, considered as a three-dimensional row vector, by the matrix G on its right. Unlike the results on matrix multiplication in Chapter 7, in the calculations here we have $1 + 1 = 0$, not $1 + 1 = 1$.

(Even if the set W of messages is not all of $\mathbf{Z}_2^3$, we'll assume that all of $\mathbf{Z}_2^3$ is encoded and that the transmitter and receiver will both know the real messages of importance and their corresponding code words.)