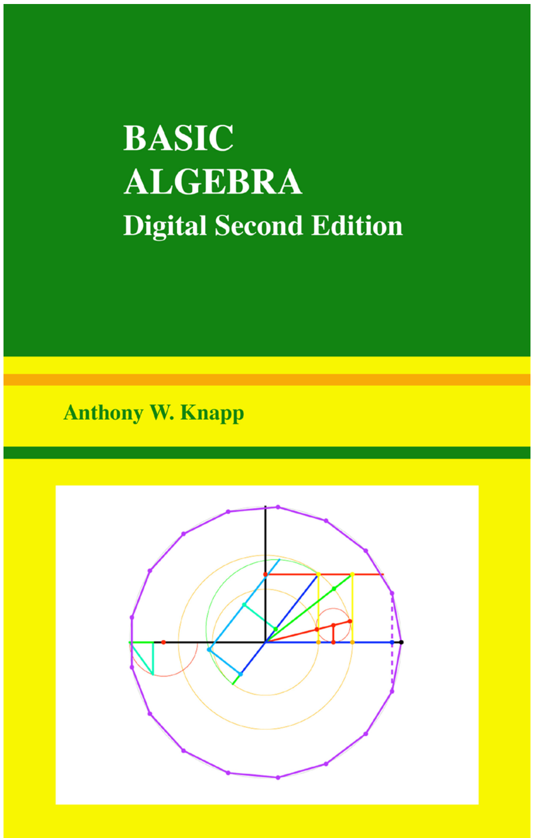
BASIC
ALGEBRA
Digital Second Edition
Anthony W. Knapp

Anthony W. Knapp
81 Upper Sheep Pasture Road
East Setauket, N.Y. 11733–1729, U.S.A.
Email to: `aknapp@math.stonybrook.edu`
Homepage: `www.math.stonybrook.edu/∼aknapp`

Title: Basic Algebra
Cover: Construction of a regular heptadecagon, the steps shown in color sequence; see page 505.

Mathematics Subject Classification (2010): 15–01, 20–01, 13–01, 12–01, 16–01, 08–01, 18A05, 68P30.

First Edition, ISBN-13 978-0-8176-3248-9
c 2006 Anthony W. Knapp
Published by Birkhäuser Boston

Digital Second Edition, not to be sold, no ISBN
c 2016 Anthony W. Knapp
Published by the Author

CHAPTER I

Preliminaries about the Integers, Polynomials, and Matrices

Abstract. This chapter is mostly a review, discussing unique factorization of positive integers, unique factorization of polynomials whose coefficients are rational or real or complex, signs of permutations, and matrix algebra.

Sections 1–2 concern unique factorization of positive integers. Section 1 proves the division and Euclidean algorithms, used to compute greatest common divisors. Section 2 establishes unique factorization as a consequence and gives several number-theoretic consequences, including the Chinese Remainder Theorem and the evaluation of the Euler φ function.

Section 3 develops unique factorization of rational and real and complex polynomials in one indeterminate completely analogously, and it derives the complete factorization of complex polynomials from the Fundamental Theorem of Algebra. The proof of the fundamental theorem is postponed to Chapter IX.

Section 4 discusses permutations of a finite set, establishing the decomposition of each permutation as a disjoint product of cycles. The sign of a permutation is introduced, and it is proved that the sign of a product is the product of the signs.

Sections 5–6 concern matrix algebra. Section 5 reviews row reduction and its role in the solution of simultaneous linear equations. Section 6 defines the arithmetic operations of addition, scalar multiplication, and multiplication of matrices. The process of matrix inversion is related to the method of row reduction, and it is shown that a square matrix with a one-sided inverse automatically has a two-sided inverse that is computable via row reduction.

1. Division and Euclidean Algorithms

The first three sections give a careful proof of unique factorization for integers and for polynomials with rational or real or complex coefficients, and they give an indication of some first consequences of this factorization. For the moment let us restrict attention to the set $\mathbb{Z}$ of integers. We take addition, subtraction, and multiplication within $\mathbb{Z}$ as established, as well as the properties of the usual ordering in $\mathbb{Z}$.

A **factor** of an integer n is a nonzero integer k such that $n = kl$ for some integer l. In this case we say also that k **divides** n, that k is a **divisor** of n, and that n is a **multiple** of k. We write $k \mid n$ for this relationship. If n is nonzero, any product formula $n = kl_1 \cdots l_r$ is a **factorization** of n. A **unit** in $\mathbb{Z}$ is a divisor of 1, hence is either $+1$ or -1. The factorization $n = kl$ of $n \neq 0$ is called **nontrivial** if neither k nor l is a unit. An integer $p > 1$ is said to be **prime** if it has no nontrivial factorization $p = kl$.

The statement of unique factorization for positive integers, which will be given precisely in Section 2, says roughly that each positive integer is the product of primes and that this decomposition is unique apart from the order of the factors.[1] Existence will follow by an easy induction. The difficulty is in the uniqueness. We shall prove uniqueness by a sequence of steps based on the "Euclidean algorithm," which we discuss in a moment. In turn, the Euclidean algorithm relies on the following.

Proposition 1.1 (division algorithm). If a and b are integers with $b \neq 0$, then there exist unique integers q and r such that $a = bq + r$ and $0 \leq r < |b|$.

PROOF. Possibly replacing q by $-q$, we may assume that $b > 0$. The integers n with $bn \leq a$ are bounded above by $|a|$, and there exists such an n, namely $n = -|a|$. Therefore there is a largest such integer, say $n = q$. Set $r = a - bq$. Then $0 \leq r$ and $a = bq + r$. If $r \geq b$, then $r - b \geq 0$ says that $a = b(q+1) + (r - b) \geq b(q+1)$. The inequality $q + 1 > q$ contradicts the maximality of q, and we conclude that $r < b$. This proves existence.

For uniqueness when $b > 0$, suppose $a = bq_1 + r_1 = bq_2 + r_2$. Subtracting, we obtain $b(q_1 - q_2) = r_2 - r_1$ with $|r_2 - r_1| < b$, and this is a contradiction unless $r_2 - r_1 = 0$. □

Let a and b be integers not both 0. The **greatest common divisor** of a and b is the largest integer $d > 0$ such that $d \mid a$ and $d \mid b$. Let us see existence. The integer 1 divides a and b. If b, for example, is nonzero, then any such d has $|d| \leq |b|$, and hence the greatest common divisor indeed exists. We write $d = \mathrm{GCD}(a, b)$.

Let us suppose that $b \neq 0$. The **Euclidean algorithm** consists of iterated application of the division algorithm (Proposition 1.1) to a and b until the remainder term r disappears:

$$\begin{aligned}
a &= bq_1 + r_1, & & 0 \leq r_1 < b,\\
b &= r_1 q_2 + r_2, & & 0 \leq r_2 < r_1,\\
r_1 &= r_2 q_3 + r_3, & & 0 \leq r_3 < r_2,\\
&\ \vdots\\
r_{n-2} &= r_{n-1} q_n + r_n, & & 0 \leq r_n < r_{n-1} \quad (\text{with } r_n \neq 0,\ \text{say}),\\
r_{n-1} &= r_n q_{n+1}.
\end{aligned}$$

[1] It is to be understood that the prime factorization of 1 is as the empty product.

The process must stop with some remainder term r_{n+1} equal to 0 in this way since $b > r_1 > r_2 > \cdots \geq 0$. The last nonzero remainder term, namely r_n above, will be of interest to us.

EXAMPLE. For $a = 13$ and $b = 5$, the steps read

$$\begin{aligned} 13 &= 5 \cdot 2 + 3, \\ 5 &= 3 \cdot 1 + 2, \\ 3 &= 2 \cdot 1 + \boxed{1}, \\ 2 &= 1 \cdot 2. \end{aligned}$$

The last nonzero remainder term is written with a box around it.

Proposition 1.2. Let a and b be integers with $b \neq 0$, and let $d = \mathrm{GCD}(a, b)$. Then

(a) the number r_n in the Euclidean algorithm is exactly d,
(b) any divisor d' of both a and b necessarily divides d,
(c) there exist integers x and y such that $ax + by = d$.

REMARK. Proposition 1.2c is sometimes called **Bezout's identity.**

EXAMPLE, CONTINUED. We rewrite the steps of the Euclidean algorithm, as applied in the above example with $a = 13$ and $b = 5$, so as to yield successive substitutions:

$$\begin{aligned} 13 &= 5 \cdot 2 + 3, & 3 &= 13 - 5 \cdot 2, \\ 5 &= 3 \cdot 1 + 2, & 2 &= 5 - 3 \cdot 1 = 5 - (13 - 5 \cdot 2) \cdot 1 = 5 \cdot 3 - 13 \cdot 1, \\ 3 &= 2 \cdot 1 + \boxed{1}, & 1 &= 3 - 2 \cdot 1 = (13 - 5 \cdot 2) - (5 \cdot 3 - 13 \cdot 1) \cdot 1 \\ & & &= 13 \cdot 2 - 5 \cdot 5. \end{aligned}$$

Thus we see that $1 = 13x + 5y$ with $x = 2$ and $y = -5$. This shows for the example that the number r_n works in place of d in Proposition 1.2c, and the rest of the proof of the proposition for this example is quite easy. Let us now adjust this computation to obtain a complete proof of the proposition in general.

PROOF OF PROPOSITION 1.2. Put $r_0 = b$ and $r_{-1} = a$, so that

$$r_{k-2} = r_{k-1} q_k + r_k \qquad \text{for } 1 \leq k \leq n. \tag{$*$}$$

The argument proceeds in three steps.

Step 1. We show that r_n is a divisor of both a and b. In fact, from $r_{n-1} = r_n q_{n+1}$, we have $r_n \mid r_{n-1}$. Let $k \le n$, and assume inductively that r_n divides $r_{k-1}, \dots, r_{n-1}, r_n$. Then $(*)$ shows that r_n divides r_{k-2}. Induction allows us to conclude that r_n divides $r_{-1}, r_0, \dots, r_{n-1}$. In particular, r_n divides a and b.

Step 2. We prove that $ax + by = r_n$ for suitable integers x and y. In fact, we show by induction on k for $k \le n$ that there exist integers x and y with $ax + by = r_k$. For $k = -1$ and $k = 0$, this conclusion is trivial. If $k \ge 1$ is given and if the result is known for $k-2$ and $k-1$, then we have

$$\begin{aligned} ax_2 + by_2 &= r_{k-2}, \\ ax_1 + by_1 &= r_{k-1} \end{aligned} \tag{$**$}$$

for suitable integers x_2, y_2, x_1, y_1. We multiply the second of the equalities of $(**)$ by q_k, subtract, and substitute into $(*)$. The result is

$$r_k = r_{k-2} - r_{k-1} q_k = a(x_2 - q_k x_1) + b(y_2 - q_k y_1),$$

and the induction is complete. Thus $ax + by = r_n$ for suitable x and y.

Step 3. Finally we deduce (a), (b), and (c). Step 1 shows that r_n divides a and b. If $d' > 0$ divides both a and b, the result of Step 2 shows that $d' \mid r_n$. Thus $d' \le r_n$, and r_n is the greatest common divisor. This is the conclusion of (a); (b) follows from (a) since $d' \mid r_n$, and (c) follows from (a) and Step 2. □

Corollary 1.3. Within $\mathbb{Z}$, if c is a nonzero integer that divides a product mn and if $\mathrm{GCD}(c, m) = 1$, then c divides n.

PROOF. Proposition 1.2c produces integers x and y with $cx + my = 1$. Multiplying by n, we obtain $cnx + mny = n$. Since c divides mn and divides itself, c divides both terms on the left side. Therefore it divides the right side, which is n. □

Corollary 1.4. Within $\mathbb{Z}$, if a and b are nonzero integers with $\mathrm{GCD}(a, b) = 1$ and if both of them divide the integer m, then ab divides m.

PROOF. Proposition 1.2c produces integers x and y with $ax + by = 1$. Multiplying by m, we obtain $amx + bmy = m$, which we rewrite in integers as $ab(m/b)x + ab(m/a)y = m$. Since ab divides each term on the left side, it divides the right side, which is m. □

2. Unique Factorization of Integers

We come now to the theorem asserting unique factorization for the integers. The precise statement is as follows.

Theorem 1.5 (Fundamental Theorem of Arithmetic). Each positive integer n can be written as a product of primes, $n = p_1p_2\cdots p_r$, with the integer 1 being written as an empty product. This factorization is unique in the following sense: if $n = q_1q_2\cdots q_s$ is another such factorization, then $r = s$ and, after some reordering of the factors, $q_j = p_j$ for $1 \le j \le r$.

The main step is the following lemma, which relies on Corollary 1.3.

Lemma 1.6. Within $\mathbb{Z}$, if p is a prime and p divides a product ab, then p divides a or p divides b.

REMARK. Lemma 1.6 is sometimes known as **Euclid's Lemma.**

PROOF. Suppose that p does not divide a. Since p is prime, $\mathrm{GCD}(a, p) = 1$. Taking $m = a$, $n = b$, and $c = p$ in Corollary 1.3, we see that p divides b. □

PROOF OF EXISTENCE IN THEOREM 1.5. We induct on n, the case $n = 1$ being handled by an empty product expansion. If the result holds for $k = 1$ through $k = n - 1$, there are two cases: n is prime and n is not prime. If n is prime, then $n = n$ is the desired factorization. Otherwise we can write $n = ab$ nontrivially with $a > 1$ and $b > 1$. Then $a \le n - 1$ and $b \le n - 1$, so that a and b have factorizations into primes by the inductive hypothesis. Putting them together yields a factorization into primes for $n = ab$. □

PROOF OF UNIQUENESS IN THEOREM 1.5. Suppose that $n = p_1p_2\cdots p_r = q_1q_2\cdots q_s$ with all factors prime and with $r \le s$. We prove the uniqueness by induction on r, the case $r = 0$ being trivial and the case $r = 1$ following from the definition of "prime." Inductively from Lemma 1.6 we have $p_r \mid q_k$ for some k. Since q_k is prime, $p_r = q_k$. Thus we can cancel and obtain $p_1p_2\cdots p_{r-1} = q_1q_2\cdots \widehat{q_k}\cdots q_s$, the hat indicating an omitted factor. By induction the factors on the two sides here are the same except for order. Thus the same conclusion is valid when comparing the two sides of the equality $p_1p_2\cdots p_r = q_1q_2\cdots q_s$. The induction is complete, and the desired uniqueness follows. □

In the product expansion of Theorem 1.5, it is customary to group factors that are equal, thus writing the positive integer n as $n = p_1^{k_1}\cdots p_r^{k_r}$ with the primes p_j distinct and with the integers k_j all ≥ 0. This kind of decomposition is unique up to order if all factors $p_j^{k_j}$ with $k_j = 0$ are dropped, and we call it a **prime factorization** of n.

Corollary 1.7. If $n = p_1^{k_1}\cdots p_r^{k_r}$ is a prime factorization of a positive integer n, then the positive divisors d of n are exactly all products $d = p_1^{l_1}\cdots p_r^{l_r}$ with $0 \le l_j \le k_j$ for all j.

REMARK. A general divisor of n within $\mathbb{Z}$ is the product of a unit ± 1 and a positive divisor.

PROOF. Certainly any such product divides n. Conversely if d divides n, write $n = dx$ for some positive integer x. Apply Theorem 1.5 to d and to x, form the resulting prime factorizations, and multiply them together. Then we see from the uniqueness for the prime factorization of n that the only primes that can occur in the expansions of d and x are $p_1, \dots, p_r$ and that the sum of the exponents of p_j in the expansions of d and x is k_j. The result follows. □

If we want to compare prime factorizations for two positive integers, we can insert 0th powers of primes as necessary and thereby assume that the same primes appear in both expansions. Using this device, we obtain a formula for greatest common divisors.

Corollary 1.8. If two positive integers a and b have expansions as products of powers of r distinct primes given by $a = p_1^{k_1} \cdots p_r^{k_r}$ and $b = p_1^{l_1} \cdots p_r^{l_r}$, then

$$\mathrm{GCD}(a, b) = p_1^{\min(k_1, l_1)} \cdots p_r^{\min(k_r, l_r)}.$$

PROOF. Let d' be the right side of the displayed equation. It is plain that d' is positive and that d' divides a and b. On the other hand, two applications of Corollary 1.7 show that the greatest common divisor of a and b is a number d of the form $p_1^{m_1} \cdots p_r^{m_r}$ with the property that $m_j \leq k_j$ and $m_j \leq l_j$ for all j. Therefore $m_j \leq \min(k_j, l_j)$ for all j, and $d \leq d'$. Since any positive divisor of both a and b is $\leq d$, we have $d' \leq d$. Thus $d' = d$. □

In special cases Corollary 1.8 provides a useful way to compute $\mathrm{GCD}(a, b)$, but the Euclidean algorithm is usually a more efficient procedure. Nevertheless, Corollary 1.8 remains a handy tool for theoretical purposes. Here is an example: Two nonzero integers a and b are said to be **relatively prime** if $\mathrm{GCD}(a, b) = 1$. It is immediate from Corollary 1.8 that two nonzero integers a and b are relatively prime if and only if there is no prime p that divides both a and b.

Corollary 1.9 (Chinese Remainder Theorem). Let a and b be positive relatively prime integers. To each pair (r, s) of integers with $0 \leq r < a$ and $0 \leq s < b$ corresponds a unique integer n such that $0 \leq n < ab$, a divides $n - r$, and b divides $n - s$. Moreover, every integer n with $0 \leq n < ab$ arises from some such pair (r, s).

REMARK. In notation for congruences that we introduce formally in Chapter IV, the result says that if $\mathrm{GCD}(a, b) = 1$, then the congruences $n \equiv r \bmod a$ and $n \equiv s \bmod b$ have one and only one simultaneous solution n with $0 \leq n < ab$.

PROOF. Let us see that n exists as asserted. Since a and b are relatively prime, Proposition 1.2c produces integers x' and y' such that $ax' - by' = 1$. Multiplying by $s - r$, we obtain $ax - by = s - r$ for suitable integers x and y. Put $t = ax + r = by + s$, and write by the division algorithm (Proposition 1.1) $t = abq + n$ for some integer q and for some integer n with $0 \leq n < ab$. Then $n - r = t - abq - r = ax - abq$ is divisible by a, and similarly $n - s$ is divisible by b.

Suppose that n and n' both have the asserted properties. Then a divides $n - n' = (n - r) - (n' - r)$, and b divides $n - n' = (n - s) - (n' - s)$. Since a and b are relatively prime, Corollary 1.4 shows that ab divides $n - n'$. But $|n - n'| < ab$, and the only integer N with $|N| < ab$ that is divisible by ab is $N = 0$. Thus $n - n' = 0$ and $n = n'$. This proves uniqueness.

Finally the argument just given defines a one-one function from a set of ab pairs (r, s) to a set of ab elements n. Its image must therefore be all such integers n. This proves the corollary. □

If n is a positive integer, we define $\varphi(n)$ to be the number of integers k with $0 \leq k < n$ such that k and n are relatively prime. The function φ is called the **Euler φ function**.

Corollary 1.10. Let $N > 1$ be an integer, and let $N = p_1^{k_1} \cdots p_r^{k_r}$ be a prime factorization of N. Then

$$\varphi(N) = \prod_{j=1}^{r} p_j^{k_j - 1}(p_j - 1).$$

REMARK. The conclusion is valid also for $N = 1$ if we interpret the right side of the formula to be the empty product.

PROOF. For positive integers a and b, let us check that

$$\varphi(ab) = \varphi(a)\varphi(b) \qquad \text{if} \quad \mathrm{GCD}(a, b) = 1. \tag{$*$}$$

In view of Corollary 1.9, it is enough to prove that the mapping $(r, s) \mapsto n$ given in that corollary has the property that $\mathrm{GCD}(r, a) = \mathrm{GCD}(s, b) = 1$ if and only if $\mathrm{GCD}(n, ab) = 1$.

To see this property, suppose that n satisfies $0 \leq n < ab$ and $\mathrm{GCD}(n, ab) > 1$. Choose a prime p dividing both n and ab. By Lemma 1.6, p divides a or p divides b. By symmetry we may assume that p divides a. If (r, s) is the pair corresponding to n under Corollary 1.9, then the corollary says that a divides $n - r$. Since p divides a, p divides $n - r$. Since p divides n, p divides r. Thus $\mathrm{GCD}(r, a) > 1$.

Conversely suppose that (r, s) is a pair with $0 \leq r < a$ and $0 \leq s < b$ such that $\mathrm{GCD}(r, a) = \mathrm{GCD}(s, b) = 1$ is false. Without loss of generality, we may assume that $\mathrm{GCD}(r, a) > 1$. Choose a prime p dividing both r and a. If n is the integer with $0 \le n < ab$ that corresponds to (r, s) under Corollary 1.9, then the corollary says that a divides $n - r$. Since p divides a, p divides $n - r$. Since p divides r, p divides n. Thus $\mathrm{GCD}(n, ab) > 1$. This completes the proof of $(*)$.

For a power p^k of a prime p with $k > 0$, the integers n with $0 \le n < p^k$ such that $\mathrm{GCD}(n, p^k) > 1$ are the multiples of p, namely $0, p, 2p, \dots, p^k - p$. There are p^{k-1} of them. Thus the number of integers n with $0 \le n < p^k$ such that $\mathrm{GCD}(n, p^k) = 1$ is $p^k - p^{k-1} = p^{k-1}(p - 1)$. In other words,

$$\varphi(p^k) = p^{k-1}(p-1) \qquad \text{if } p \text{ is prime and } k \ge 1. \qquad (**)$$

To prove the corollary, we induct on r, the case $r = 1$ being handled by $(**)$. If the formula of the corollary is valid for $r - 1$, then $(*)$ allows us to combine that result with the formula for $\varphi(p^{k_r})$ given in $(**)$ to obtain the formula for $\varphi(N)$. □

We conclude this section by extending the notion of greatest common divisor to apply to more than two integers. If $a_1, \dots, a_t$ are integers not all 0, their **greatest common divisor** is the largest integer $d > 0$ that divides all of $a_1, \dots, a_t$. This exists, and we write $d = \mathrm{GCD}(a_1, \dots, a_t)$ for it. It is immediate that d equals the greatest common divisor of the nonzero members of the set $\{a_1, \dots, a_t\}$. Thus, in deriving properties of greatest common divisors, we may assume that all the integers are nonzero.

Corollary 1.11. Let $a_1, \dots, a_t$ be positive integers, and let d be their greatest common divisor. Then

(a) if for each j with $1 \le j \le t$, $a_j = p_1^{k_{1,j}} \cdots p_r^{k_{r,j}}$ is an expansion of a_j as a product of powers of r distinct primes $p_1, \dots, p_r$, it follows that

$$d = p_1^{\min_{1\le j\le t}\{k_{1,j}\}} \cdots p_r^{\min_{1\le j\le t}\{k_{r,j}\}},$$

(b) any divisor d' of all of $a_1, \dots, a_t$ necessarily divides d,

(c) $d = \mathrm{GCD}\big(\mathrm{GCD}(a_1, \dots, a_{t-1}), a_t\big)$ if $t > 1$,

(d) there exist integers $x_1, \dots, x_t$ such that $a_1x_1 + \cdots + a_tx_t = d$.

PROOF. Part (a) is proved in the same way as Corollary 1.8 except that Corollary 1.7 is to be applied r times rather than just twice. Further application of Corollary 1.7 shows that any positive divisor d' of $a_1, \dots, a_t$ is of the form $d' = p_1^{m_1} \cdots p_r^{m_r}$ with $m_1 \le k_{1,j}$ for all j, ... , and with $m_r \le k_{r,j}$ for all j. Therefore $m_1 \le \min_{1\le j\le r}\{k_{1,j}\}$, ... , and $m_r \le \min_{1\le j\le r}\{k_{r,j}\}$, and it follows that d' divides d. This proves (b). Conclusion (c) follows by using the formula in (a), and (d) follows by combining (c), Proposition 1.2c, and induction. □

3. Unique Factorization of Polynomials

This section establishes unique factorization for ordinary rational, real, and complex polynomials. We write $\mathbb{Q}$ for the set of rational numbers, $\mathbb{R}$ for the set of real numbers, and $\mathbb{C}$ for the set of complex numbers, each with its arithmetic operations. The rational numbers are constructed from the integers by a process reviewed in Section A3 of the appendix, the real numbers are defined from the rational numbers by a process reviewed in that same section, and the complex numbers are defined from the real numbers by a process reviewed in Section A4 of the appendix. Sections A3 and A4 of the appendix mention special properties of $\mathbb{R}$ and $\mathbb{C}$ beyond those of the arithmetic operations, but we shall not make serious use of these special properties here until nearly the end of the section—after unique factorization of polynomials has been established. Let $\mathbb{F}$ denote any of $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$. The members of $\mathbb{F}$ are called **scalars**.

We work with ordinary polynomials with coefficients in $\mathbb{F}$. Informally these are expressions $P(X) = a_nX^n + \cdots + a_1X + a_0$ with $a_n, \dots, a_1, a_0$ in $\mathbb{F}$. Although it is tempting to think of $P(X)$ as a function with independent variable X, it is better to identify P with the sequence $(a_0, a_1, \dots, a_n, 0, 0, \dots)$ of coefficients, using expressions $P(X) = a_nX^n + \cdots + a_1X + a_0$ only for conciseness and for motivation of the definitions of various operations.

The precise definition therefore is that a **polynomial** in **one indeterminate** with **coefficients** in $\mathbb{F}$ is an infinite sequence of members of $\mathbb{F}$ such that all terms of the sequence are 0 from some point on. The indexing of the sequence is to begin with 0. We may refer to a polynomial P as $P(X)$ if we want to emphasize that the indeterminate is called X. Addition, subtraction, and scalar multiplication are defined in coordinate-by-coordinate fashion:

$$\begin{aligned}
&(a_0, a_1, \dots, a_n, 0, 0, \dots) + (b_0, b_1, \dots, b_n, 0, 0, \dots)\\
&\qquad\qquad = (a_0 + b_0, a_1 + b_1, \dots, a_n + b_n, 0, 0, \dots),\\
&(a_0, a_1, \dots, a_n, 0, 0, \dots) - (b_0, b_1, \dots, b_n, 0, 0, \dots)\\
&\qquad\qquad = (a_0 - b_0, a_1 - b_1, \dots, a_n - b_n, 0, 0, \dots),\\
&\quad c(a_0, a_1, \dots, a_n, 0, 0, \dots) = (ca_0, ca_1, \dots, ca_n, 0, 0, \dots).
\end{aligned}$$

Polynomial multiplication is defined so as to match multiplication of expressions $a_nX^n + \cdots + a_1X + a_0$ if the product is expanded out, powers of X are added, and then terms containing like powers of X are collected:

$$(a_0, a_1, \dots, 0, 0, \dots)(b_0, b_1, \dots, 0, 0, \dots) = (c_0, c_1, \dots, 0, 0, \dots),$$

where $c_N = \sum_{k=0}^{N} a_k b_{N-k}$. We take it as known that the usual associative, commutative, and distributive laws are then valid. The set of all polynomials in the indeterminate X is denoted by $\mathbb{F}[X]$.

The polynomial with all entries 0 is denoted by 0 and is called the **zero polynomial**. For all polynomials $P = (a_0, \dots, a_n, 0, \dots)$ other than 0, the **degree** of P, denoted by $\deg P$, is defined to be the largest index n such that $a_n \neq 0$. The **constant polynomials** are by definition the zero polynomial and the polynomials of degree 0. If P and Q are nonzero polynomials, then

$$\begin{gathered} P + Q = 0 \qquad \text{or} \qquad \deg(P+Q) \le \max(\deg P, \deg Q), \\ \deg(cP) = \deg P, \\ \deg(PQ) = \deg P + \deg Q. \end{gathered}$$

In the formula for $\deg(P+Q)$, equality holds if $\deg P \neq \deg Q$. Implicit in the formula for $\deg(PQ)$ is the fact that PQ cannot be 0 unless $P = 0$ or $Q = 0$. A cancellation law for multiplication is an immediate consequence:

$$PR = QR \text{ with } R \neq 0 \qquad \text{implies} \qquad P = Q.$$

In fact, $PR = QR$ implies $(P - Q)R = 0$; since $R \neq 0$, $P - Q$ must be 0.

If $P = (a_0, \dots, a_n, 0, \dots)$ is a polynomial and r is in $\mathbb{F}$, we can **evaluate** P at r, obtaining as a result the number $P(r) = a_n r^n + \cdots + a_1 r + a_0$. Taking into account all values of r, we obtain a mapping $P \mapsto P(\,\cdot\,)$ of $\mathbb{F}[X]$ into the set of functions from $\mathbb{F}$ into $\mathbb{F}$. Because of the way that the arithmetic operations on polynomials have been defined, we have

$$\begin{aligned} (P+Q)(r) &= P(r) + Q(r), \\ (P-Q)(r) &= P(r) - Q(r), \\ (cP)(r) &= cP(r), \\ (PQ)(r) &= P(r)Q(r). \end{aligned}$$

In other words, the mapping $P \mapsto P(\,\cdot\,)$ respects the arithmetic operations. We say that r is a **root** of P if $P(r) = 0$.

Now we turn to the question of unique factorization. The definitions and the proof are completely analogous to those for the integers. A **factor** of a polynomial A is a nonzero polynomial B such that $A = BQ$ for some polynomial Q. In this case we say also that B **divides** A, that B is a **divisor** of A, and that A is a **multiple** of B. We write $B \mid A$ for this relationship. If A is nonzero, any product formula $A = BQ_1 \cdots Q_r$ is a **factorization** of A. A **unit** in $\mathbb{F}[X]$ is a divisor of 1, hence is any polynomial of degree 0; such a polynomial is a constant polynomial $A(X) = c$ with c equal to a nonzero scalar. The factorization $A = BQ$ of $A \neq 0$ is called **nontrivial** if neither B nor Q is a unit. A **prime** P in $\mathbb{F}[X]$ is a nonzero polynomial that is not a unit and has no nontrivial factorization $P = BQ$. Observe that the product of a prime and a unit is always a prime.

Proposition 1.12 (division algorithm). If A and B are polynomials in $\mathbb{F}[X]$ and if B not the 0 polynomial, then there exist unique polynomials Q and R in $\mathbb{F}[X]$ such that

(a) $A = BQ + R$ and
(b) either R is the 0 polynomial or $\deg R < \deg B$.

REMARK. This result codifies the usual method of dividing polynomials in high-school algebra. That method writes $A/B = Q + R/B$, and then one obtains the above result by multiplying by B. The polynomial Q is the quotient in the division, and R is the remainder.

PROOF OF UNIQUENESS. If $A = BQ + R = BQ_1 + R_1$, then $B(Q - Q_1) = R_1 - R$. Without loss of generality, $R_1 - R$ is not the 0 polynomial since otherwise $Q - Q_1 = 0$ also. Then

$$\deg B + \deg(Q - Q_1) = \deg(R_1 - R) \le \max(\deg R, \deg R_1) < \deg B,$$

and we have a contradiction. □

PROOF OF EXISTENCE. If $A = 0$ or $\deg A < \deg B$, we take $Q = 0$ and $R = A$, and we are done. Otherwise we induct on $\deg A$. Assume the result for degree $\le n - 1$, and let $\deg A = n$. Write $A = a_n X^n + A_1$ with $A_1 = 0$ or $\deg A_1 < \deg A$. Let $B = b_k X^k + B_1$ with $B_1 = 0$ or $\deg B_1 < \deg B$. Put $Q_1 = a_n b_k^{-1} X^{n-k}$. Then

$$A - BQ_1 = a_n X^n + A_1 - a_n X^n - a_n b_k^{-1} X^{n-k} B_1 = A_1 - a_n b_k^{-1} X^{n-k} B_1$$

with the right side equal to 0 or of degree $< \deg A$. Then the right side, by induction, is of the form $BQ_2 + R$, and $A = B(Q_1 + Q_2) + R$ is the required decomposition. □

Corollary 1.13 (Factor Theorem). If r is in $\mathbb{F}$ and if P is a polynomial in $\mathbb{F}[X]$, then $X - r$ divides P if and only if $P(r) = 0$.

PROOF. If $P = (X - r)Q$, then $P(r) = (r - r)Q(r) = 0$. Conversely let $P(r) = 0$. Taking $B(X) = X - r$ in the division algorithm (Proposition 1.12), we obtain $P = (X - r)Q + R$ with $R = 0$ or $\deg R < \deg(X - r) = 1$. Thus R is a constant polynomial, possibly 0. In any case we have $0 = P(r) = (r - r)Q(r) + R(r)$, and thus $R(r) = 0$. Since R is constant, we must have $R = 0$, and then $P = (X - r)Q$. □

Corollary 1.14. If P is a nonzero polynomial with coefficients in $\mathbb{F}$ and if $\deg P = n$, then P has at most n distinct roots.

REMARKS. Since there are infinitely many scalars in any of $\mathbb{Q}$ and $\mathbb{R}$ and $\mathbb{C}$, the corollary implies that the function from $\mathbb{F}$ to $\mathbb{F}$ associated to P, namely $r \mapsto P(r)$, cannot be identically 0 if $P \neq 0$. Starting in Chapter IV, we shall allow other $\mathbb{F}$'s besides $\mathbb{Q}$ and $\mathbb{R}$ and $\mathbb{C}$, and then this implication can fail. For example, when $\mathbb{F}$ is the two-element "field" $\mathbb{F} = \{0, 1\}$ with $1 + 1 = 0$ and with otherwise the expected addition and multiplication, then $P(X) = X^2 + X$ is not the zero polynomial but $P(r) = 0$ for $r = 0$ and $r = 1$. It is thus important to distinguish polynomials in one indeterminate from their associated functions of one variable.

PROOF. Let $r_1, \dots, r_{n+1}$ be distinct roots of $P(X)$. By the Factor Theorem (Corollary 1.13), $X - r_1$ is a factor of $P(X)$. We prove inductively on k that the product $(X - r_1)(X - r_2) \cdots (X - r_k)$ is a factor of $P(X)$. Assume that this assertion holds for k, so that $P(X) = (X - r_1) \cdots (X - r_k) Q(X)$ and

$$0 = P(r_{k+1}) = (r_{k+1} - r_1) \cdots (r_{k+1} - r_k) Q(r_{k+1}).$$

Since the r_j's are distinct, we must have $Q(r_{k+1}) = 0$. By the Factor Theorem, we can write $Q(X) = (X - r_{k+1}) R(X)$ for some polynomial $R(X)$. Substitution gives $P(X) = (X - r_1) \cdots (X - r_k)(X - r_{k+1}) R(X)$, and $(X - r_1) \cdots (X - r_{k+1})$ is exhibited as a factor of $P(X)$. This completes the induction. Consequently

$$P(X) = (X - r_1) \cdots (X - r_{n+1}) S(X)$$

for some polynomial $S(X)$. Comparing the degrees of the two sides, we find that $\deg S = -1$, and we have a contradiction. □

We can use the division algorithm in the same way as with the integers in Sections 1–2 to obtain unique factorization. Within the set of integers, we defined greatest common divisors so as to be positive, but their negatives would have worked equally well. That flexibility persists with polynomials; the essential feature of any greatest common divisor of polynomials is shared by any product of that polynomial by a unit. A **greatest common divisor** of polynomials A and B with $B \neq 0$ is any polynomial D of maximum degree such that D divides A and D divides B. We shall see that D is indeed unique up to multiplication by a nonzero scalar.[2]

[2] For some purposes it is helpful to isolate one particular greatest common divisor by taking the coefficient of the highest power of X to be 1.

The **Euclidean algorithm** is the iterative process that makes use of the division algorithm in the form

$$\begin{aligned}
A &= BQ_1 + R_1, & R_1 &= 0 \text{ or } \deg R_1 < \deg B,\\
B &= R_1 Q_2 + R_2, & R_2 &= 0 \text{ or } \deg R_2 < \deg R_1,\\
R_1 &= R_2 Q_3 + R_3, & R_3 &= 0 \text{ or } \deg R_3 < \deg R_2,\\
&\vdots\\
R_{n-2} &= R_{n-1} Q_n + R_n, & R_n &= 0 \text{ or } \deg R_n < \deg R_{n-1},\\
R_{n-1} &= R_n Q_{n+1}.
\end{aligned}$$

In the above computation the integer n is defined by the conditions that $R_n \neq 0$ and that $R_{n+1} = 0$. Such an n must exist since $\deg B > \deg R_1 > \cdots \geq 0$. We can now obtain an analog for $\mathbb{F}[X]$ of the result for $\mathbb{Z}$ given as Proposition 1.2.

Proposition 1.15. Let A and B be polynomials in $\mathbb{F}[X]$ with $B \neq 0$, and let $R_1, \ldots, R_n$ be the remainders generated by the Euclidean algorithm when applied to A and B. Then

(a) R_n is a greatest common divisor of A and B,
(b) any D_1 that divides both A and B necessarily divides R_n,
(c) the greatest common divisor of A and B is unique up to multiplication by a nonzero scalar,
(d) any greatest common divisor D has the property that there exist polynomials P and Q with $AP + BQ = D$.

PROOF. Conclusions (a) and (b) are proved in the same way that parts (a) and (b) of Proposition 1.2 are proved, and conclusion (d) is proved with $D = R_n$ in the same way that Proposition 1.2c is proved.

If D is a greatest common divisor of A and B, it follows from (a) and (b) that D divides R_n and that $\deg D = \deg R_n$. This proves (c). □

Using Proposition 1.15, we can prove analogs for $\mathbb{F}[X]$ of the two corollaries of Proposition 1.2. But let us instead skip directly to what is needed to obtain an analog for $\mathbb{F}[X]$ of unique factorization as in Theorem 1.5.

Lemma 1.16. If A and B are nonzero polynomials with coefficients in $\mathbb{F}$ and if P is a prime polynomial such that P divides AB, then P divides A or P divides B.

PROOF. If P does not divide A, then 1 is a greatest common divisor of A and P, and Proposition 1.15d produces polynomials S and T such that $AS + PT = 1$. Multiplication by B gives $ABS + PTB = B$. Then P divides ABS because it divides AB, and P divides PTB because it divides P. Hence P divides B. □

Theorem 1.17 (unique factorization). Every member of $\mathbb{F}[X]$ of degree ≥ 1 is a product of primes. This factorization is unique up to order and up to multiplication of each prime factor by a unit, i.e., by a nonzero scalar.

PROOF. The existence follows in the same way as the existence in Theorem 1.5; induction on the integers is to be replaced by induction on the degree. The uniqueness follows from Lemma 1.16 in the same way that the uniqueness in Theorem 1.5 follows from Lemma 1.6. □

We turn to a consideration of properties of polynomials that take into account special features of $\mathbb{R}$ and $\mathbb{C}$. If $\mathbb{F}$ is $\mathbb{R}$, then $X^2 + 1$ is prime. The reason is that a nontrivial factorization of $X^2 + 1$ would have to involve two first-degree real polynomials and then r^2+1 would have to be 0 for some real r, namely for r equal to the root of either of the first-degree polynomials. On the other hand, $X^2 + 1$ is not prime when $\mathbb{F} = \mathbb{C}$ since $X^2 + 1 = (X + i)(X - i)$. The Fundamental Theorem of Algebra, stated below, implies that every prime polynomial over $\mathbb{C}$ is of degree 1. It is possible to prove the Fundamental Theorem of Algebra within complex analysis as a consequence of Liouville's Theorem or within real analysis as a consequence of the Heine–Borel Theorem and other facts about compactness. This text gives a proof of the Fundamental Theorem of Algebra in Chapter IX using modern algebra, specifically Sylow theory as in Chapter IV and Galois theory as in Chapter IX. One further fact is needed; this fact uses elementary calculus and is proved below as Proposition 1.20.

Theorem 1.18 (Fundamental Theorem of Algebra). Any polynomial in $\mathbb{C}[X]$ with degree ≥ 1 has at least one root.

Corollary 1.19. Let P be a nonzero polynomial of degree n in $\mathbb{C}[X]$, and let $r_1, \dots, r_k$ be the distinct roots. Then there exist unique integers $m_j > 0$ for $1 \leq j \leq k$ such that $P(X)$ is a scalar multiple of $\prod_{j=1}^k (X - r_j)^{m_j}$. The numbers m_j have $\sum_{j=1}^k m_j = n$.

PROOF. We may assume that $\deg P > 0$. We apply unique factorization (Theorem 1.17) to $P(X)$. It follows from the Fundamental Theorem of Algebra (Theorem 1.18) and the Factor Theorem (Corollary 1.13) that each prime polynomial with coefficients in $\mathbb{C}$ has degree 1. Thus the unique factorization of $P(X)$ has to be of the form $c\prod_{l=1}^n (X - z_l)$ for some $c \neq 0$ and for some complex numbers z_l that are unique up to order. The z_l's are roots, and every root is a z_l by the Factor Theorem. Grouping like factors proves the desired factorization and its uniqueness. The numbers m_j have $\sum_{j=1}^k m_j = n$ by a count of degrees. □

The integers m_j in the corollary are called the **multiplicities** of the roots of the polynomial $P(X)$.

We conclude this section by proving the result from calculus that will enter the proof of the Fundamental Theorem of Algebra in Chapter IX.

Proposition 1.20. Any polynomial in $\mathbb{R}[X]$ with odd degree has at least one root.

PROOF. Without loss of generality, we may take the leading coefficient to be 1. Thus let the polynomial be $P(X) = X^{2n+1} + a_{2n}X^{2n} + \cdots + a_1X + a_0 = X^{2n+1} + R(X)$. Since $\lim_{x\to\pm\infty} P(x)/x^{2n+1} = 1$, there is some positive r_0 such that $P(-r_0) < 0$ and $P(r_0) > 0$. By the Intermediate Value Theorem, given in Section A3 of the appendix, $P(r) = 0$ for some r with $-r_0 \le r \le r_0$. □

4. Permutations and Their Signs

Let S be a finite nonempty set of n elements. A **permutation** of S is a one-one function from S onto S. The elements might be listed as $a_1, a_2, \ldots, a_n$, but it will simplify the notation to view them simply as $1, 2, \ldots, n$. We use ordinary function notation for describing the effect of permutations. Thus the value of a permutation σ at j is $\sigma(j)$, and the composition of τ followed by σ is $\sigma \circ \tau$ or simply $\sigma\tau$, with $(\sigma\tau)(j) = \sigma(\tau(j))$. Composition is automatically associative, i.e., $(\rho\sigma)\tau = \rho(\sigma\tau)$, because the effect of both sides on j, when we expand things out, is $\rho(\sigma(\tau(j)))$. The composition of two permutations is also called their **product**.

The identity permutation will be denoted by 1. Any permutation σ, being a one-one onto function, has a well-defined inverse permutation σ^{-1} with the property that $\sigma\sigma^{-1} = \sigma^{-1}\sigma = 1$. One way of describing concisely the effect of a permutation is to list its domain values and to put the corresponding range values beneath them. Thus $\sigma = \begin{pmatrix} 1\,2\,3\,4\,5 \\ 4\,3\,5\,1\,2 \end{pmatrix}$ is the permutation of $\{1, 2, 3, 4, 5\}$ with $\sigma(1) = 4$, $\sigma(2) = 3$, $\sigma(3) = 5$, $\sigma(4) = 1$, and $\sigma(5) = 2$. The inverse permutation is obtained by interchanging the two rows to obtain $\begin{pmatrix} 4\,3\,5\,1\,2 \\ 1\,2\,3\,4\,5 \end{pmatrix}$ and then adjusting the entries in the rows so that the first row is in the usual order: $\sigma^{-1} = \begin{pmatrix} 1\,2\,3\,4\,5 \\ 4\,5\,2\,1\,3 \end{pmatrix}$.

If $2 \le k \le n$, a k-**cycle** is a permutation σ that fixes each element in some subset of $n-k$ elements and moves the remaining elements $c_1, \ldots, c_k$ according to $\sigma(c_1) = c_2$, $\sigma(c_2) = c_3, \ldots, \sigma(c_{k-1}) = c_k$, $\sigma(c_k) = c_1$. Such a cycle may be denoted by $(c_1\ \ c_2\ \ \cdots\ \ c_{k-1}\ \ c_k)$ to stress its structure. For example take $n = 5$; then $\sigma = (2\ \ 3\ \ 5)$ is the 3-cycle given in our earlier notation by $\begin{pmatrix} 1\,2\,3\,4\,5 \\ 1\,3\,5\,4\,2 \end{pmatrix}$.

The cycle $(2\ \ 3\ \ 5)$ is the same as the cycle $(3\ \ 5\ \ 2)$ and the cycle $(5\ \ 2\ \ 3)$. It is sometimes helpful to speak of the identity permutation 1 as the unique 1-cycle.

A system of cycles is said to be **disjoint** if the sets that each of them moves are disjoint in pairs. Thus $(2\ \ 3\ \ 5)$ and $(1\ \ 4)$ are disjoint, but $(2\ \ 3\ \ 5)$ and $(1\ \ 3)$ are not. Any two disjoint cycles σ and τ commute in the sense that $\sigma\tau = \tau\sigma$.

Proposition 1.21. Any permutation σ of $\{1, 2, \dots, n\}$ is a product of disjoint cycles. The individual cycles in the decomposition are unique in the sense of being determined by σ.

EXAMPLE. $\begin{pmatrix} 1\,2\,3\,4\,5 \\ 4\,3\,5\,1\,2 \end{pmatrix} = (2\ \ 3\ \ 5)(1\ \ 4)$.

PROOF. Let us prove existence. Working with $\{1, 2, \dots, n\}$, we show that any σ is the disjoint product of cycles in such a way that no cycle moves an element j unless σ moves j. We do so for all σ simultaneously by induction downward on the number of elements fixed by σ. The starting case of the induction is that σ fixes all n elements. Then σ is the identity, and we are regarding the identity as a 1-cycle.

For the inductive step suppose σ fixes the elements in a subset T of r elements of $\{1, 2, \dots, n\}$ with $r < n$. Let j be an element not in T, so that $\sigma(j) \neq j$. Choose k as small as possible so that some element is repeated among $j, \sigma(j), \sigma^2(j), \dots, \sigma^k(j)$. This condition means that $\sigma^l(j) = \sigma^k(j)$ for some l with $0 \leq l < k$. Then $\sigma^{k-l}(j) = j$, and we obtain a contradiction to the minimality of k unless $k - l = k$, i.e., $l = 0$. In other words, we have $\sigma^k(j) = j$. We may thus form the k-cycle $\gamma = (j\ \ \sigma(j)\ \ \sigma^2(j)\ \ \sigma^{k-1}(j))$. The permutation $\gamma^{-1}\sigma$ then fixes the $r + k$ elements of $T \cup U$, where U is the set of elements $j, \sigma(j), \sigma^2(j), \dots, \sigma^{k-1}(j)$. By the inductive hypothesis, $\gamma^{-1}\sigma$ is the product $\tau_1 \cdots \tau_p$ of disjoint cycles that move only elements not in $T \cup U$. Since γ moves only the elements in U, γ is disjoint from each of $\tau_1, \dots, \tau_p$. Therefore $\sigma = \gamma\tau_1 \cdots \tau_p$ provides the required decomposition of σ.

For uniqueness we observe from the proof of existence that each element j generates a k-cycle C_j for some $k \geq 1$ depending on j. If we have two decompositions as in the proposition, then the cycle within each decomposition that contains j must be C_j. Hence the cycles in the two decompositions must match. □

A 2-cycle is often called a **transposition**. The proposition allows us to see quickly that any permutation is a product of transpositions.

Corollary 1.22. Any k-cycle σ permuting $\{1, 2, \dots, n\}$ is a product of $k - 1$ transpositions if $k > 1$. Therefore any permutation σ of $\{1, 2, \dots, n\}$ is a product of transpositions.

PROOF. For the first statement, we observe that $(c_1 \;\; c_2 \;\; \cdots \;\; c_{k-1} \;\; c_k) = (c_1 \;\; c_k)(c_1 \;\; c_{k-1}) \cdots (c_1 \;\; c_3)(c_1 \;\; c_2)$. The second statement follows by combining this fact with Proposition 1.21. $\square$

Our final tasks for this section are to attach a sign to each permutation and to examine the properties of these signs. We begin with the special case that our underlying set S is $\{1, \dots, n\}$. If σ is a permutation of $\{1, \dots, n\}$, consider the numerical products

$$\prod_{1 \le j < k \le n} |\sigma(k) - \sigma(j)| \qquad \text{and} \qquad \prod_{1 \le j < k \le n} (\sigma(k) - \sigma(j)).$$

If (r, s) is any pair of integers with $1 \le r < s \le n$, then the expression $s - r$ appears once and only once as a factor in the first product. Therefore the first product is independent of σ and equals $\prod_{1 \le j < k \le n} (k - j)$. Meanwhile, each factor of the second product is ± 1 times the corresponding factor of the first product. Therefore we have

$$\prod_{1 \le j < k \le n} (\sigma(k) - \sigma(j)) = (\operatorname{sgn} \sigma) \prod_{1 \le j < k \le n} (k - j),$$

where $\operatorname{sgn} \sigma$ is $+1$ or -1, depending on σ. This sign is called the **sign** of the permutation σ.

Lemma 1.23. Let σ be a permutation of $\{1, \dots, n\}$, let $(a \;\; b)$ be a transposition, and form the product $\sigma(a \;\; b)$. Then $\operatorname{sgn}\big(\sigma(a \;\; b)\big) = -\operatorname{sgn} \sigma$.

PROOF. For the pairs (j, k) with $j < k$, we are to compare $\sigma(k) - \sigma(j)$ with $\sigma(a \;\; b)(k) - \sigma(a \;\; b)(j)$. There are five cases. Without loss of generality, we may assume that $a < b$.

Case 1. If neither j nor k equals a or b, then $\sigma(a \;\; b)(k) - \sigma(a \;\; b)(j) = \sigma(k) - \sigma(j)$. Thus such pairs (j, k) make the same contribution to the product for $\sigma(a \;\; b)$ as to the product for σ, and they can be ignored.

Case 2. If one of j and k equals one of a and b while the other does not, there are three situations of interest. For each we compare the contributions of two such pairs together. The first situation is that of pairs (a, t) and (t, b) with $a < t < b$. These together contribute the factors $(\sigma(t) - \sigma(a))$ and $(\sigma(b) - \sigma(t))$ to the product for σ, and they contribute the factors $(\sigma(t) - \sigma(b))$ and $(\sigma(a) - \sigma(t))$ to the product for $\sigma(a \;\; b)$. Since

$$(\sigma(t) - \sigma(a))(\sigma(b) - \sigma(t)) = (\sigma(t) - \sigma(b))(\sigma(a) - \sigma(t)),$$

the pairs together make the same contribution to the product for $\sigma(a \;\; b)$ as to the product for σ, and they can be ignored.

Case 3. Continuing with matters as in Case 2, we next consider pairs (a, t) and (b, t) with $a < b < t$. These together contribute the factors $(\sigma(t) - \sigma(a))$ and $(\sigma(t) - \sigma(b))$ to the product for σ, and they contribute the factors $(\sigma(t) - \sigma(b))$ and $(\sigma(t) - \sigma(a))$ to the product for $\sigma(a\ \ b)$. Since

$$(\sigma(t) - \sigma(a))(\sigma(t) - \sigma(b)) = (\sigma(t) - \sigma(b))(\sigma(t) - \sigma(a)),$$

the pairs together make the same contribution to the product for $\sigma(a\ \ b)$ as to the product for σ, and they can be ignored.

Case 4. Still with matters as in Case 2, we consider pairs (t, a) and (t, b) with $t < a < b$. Arguing as in Case 3, we are led to an equality

$$(\sigma(a) - \sigma(t))(\sigma(b) - \sigma(t)) = (\sigma(b) - \sigma(t))(\sigma(a) - \sigma(t)),$$

and these pairs can be ignored.

Case 5. Finally we consider the pair (a, b) itself. It contributes $\sigma(b) - \sigma(a)$ to the product for σ, and it contributes $\sigma(a) - \sigma(b)$ to the product for $\sigma(a\ \ b)$. These are negatives of one another, and we get a net contribution of one minus sign in comparing our two product formulas. The lemma follows. □

Proposition 1.24. The signs of permutations of $\{1, 2, \ldots, n\}$ have the following properties:

(a) $\operatorname{sgn} 1 = +1$,
(b) $\operatorname{sgn} \sigma = (-1)^k$ if σ can be written as the product of k transpositions,
(c) $\operatorname{sgn}(\sigma\tau) = (\operatorname{sgn}\sigma)(\operatorname{sgn}\tau)$,
(d) $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}\sigma$.

PROOF. Conclusion (a) is immediate from the definition. For (b), let $\sigma = \tau_1 \cdots \tau_k$ with each τ_j equal to a transposition. We apply Lemma 1.23 recursively, using (a) at the end:

$$\begin{aligned} \operatorname{sgn}(\tau_1 \cdots \tau_k) &= (-1) \operatorname{sgn}(\tau_1 \cdots \tau_{k-1}) = (-1)^2 \operatorname{sgn}(\tau_1 \cdots \tau_{k-2}) \\ &= \cdots = (-1)^{k-1} \operatorname{sgn} \tau_1 = (-1)^k \operatorname{sgn} 1 = (-1)^k. \end{aligned}$$

For (c), Corollary 1.22 shows that any permutation is the product of transpositions. If σ is the product of k transpositions and τ is the product of l transpositions, then $\sigma\tau$ is manifestly the product of $k + l$ transpositions. Thus (c) follows from (b). Finally (d) follows from (c) and (a) by taking $\tau = \sigma^{-1}$. □

Our discussion of signs has so far attached signs only to permutations of $S = \{1, 2, \ldots, n\}$. If we are given some other set S' of n elements and we want to adapt our discussion of signs so that it applies to permutations of S', we need to identify S with S', say by a one-one onto function $\varphi : S \to S'$. If σ is a permutation of S', then $\varphi^{-1}\sigma\varphi$ is a permutation of S, and we can define $\operatorname{sgn}_\varphi(\sigma) = \operatorname{sgn}(\varphi^{-1}\sigma\varphi)$. The question is whether this definition is independent of φ.

Fortunately the answer is yes, and the proof is easy. Suppose that $\psi : S \to S'$ is a second one-one onto function, so that $\operatorname{sgn}_\psi(\sigma) = \operatorname{sgn}(\psi^{-1}\sigma\psi)$. Then $\varphi^{-1}\psi = \tau$ is a permutation of $\{1, 2, \dots, n\}$, and (c) and (d) in Proposition 1.24 give

$$\begin{aligned}\operatorname{sgn}_\psi(\sigma) &= \operatorname{sgn}(\psi^{-1}\sigma\psi) = \operatorname{sgn}(\psi^{-1}\varphi\varphi^{-1}\sigma\varphi\varphi^{-1}\psi) \\ &= \operatorname{sgn}(\tau^{-1})\operatorname{sgn}(\varphi^{-1}\sigma\varphi)\operatorname{sgn}(\tau) = \operatorname{sgn}(\tau)\operatorname{sgn}_\varphi(\sigma)\operatorname{sgn}(\tau) = \operatorname{sgn}_\varphi(\sigma).\end{aligned}$$

Consequently the definition of signs of permutations of $\{1, 2, \dots, n\}$ can be carried over to give a definition of signs of permutations of any finite nonempty set of n elements, and the resulting signs are independent of the way we enumerate the set. The conclusions of Proposition 1.24 are valid for this extended definition of signs of permutations.

5. Row Reduction

This section and the next review row reduction and matrix algebra for rational, real, and complex matrices. As in Section 3 let $\mathbb{F}$ denote $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$. The members of $\mathbb{F}$ are called **scalars**.

The term "row reduction" refers to the main part of the algorithm used for solving simultaneous systems of algebraic linear equations with coefficients in $\mathbb{F}$. Such a system is of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ &\ \,\vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n &= b_k, \end{aligned}$$

where the a_{ij} and b_i are known scalars and the x_j are the **unknowns**, or **variables**. The algorithm makes repeated use of three operations on the equations, each of which preserves the set of solutions $(x_1, \dots, x_n)$ because its inverse is an operation of the same kind:

(i) interchange two equations,
(ii) multiply an equation by a nonzero scalar,
(iii) replace an equation by the sum of it and a multiple of some other equation.

The repeated writing of the variables in carrying out these steps is tedious and unnecessary, since the steps affect only the known coefficients. Instead, we can simply work with an array of the form

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ & & \ddots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} & b_k \end{array}\right).$$

The individual scalars appearing in the array are called **entries**. The above operations on equations correspond exactly to operations on the rows[3] of the array, and they become

(i) interchange two rows,
(ii) multiply a row by a nonzero scalar,
(iii) replace a row by the sum of it and a multiple of some other row.

Any operation of these types is called an **elementary row operation**. The vertical line in the array is handy from one point of view in that it separates the left sides of the equations from the right sides; if we have more than one set of right sides, we can include all of them to the right of the vertical line and thereby solve all the systems at the same time. But from another point of view, the vertical line is unnecessary since it does not affect which operation we perform at a particular time. Let us therefore drop it, abbreviating the system as

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ & & \ddots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} & b_k \end{pmatrix}.$$

The main step in solving the system is to apply the three operations in succession to the array to reduce it to a particularly simple form. An array with k rows and m columns[4] is in **reduced row-echelon form** if it meets several conditions:

- Each member of the first l of the rows, for some l with $0 \le l \le k$, has at least one nonzero entry, and the other rows have all entries 0.
- Each of the nonzero rows has 1 as its first nonzero entry; let us say that the i^{th} nonzero row has this 1 in its $j(i)^{\text{th}}$ entry.
- The integers $j(i)$ are to be strictly increasing as a function of i, and the only entry in the $j(i)^{\text{th}}$ column that is nonzero is to be the one in the i^{th} row.

Proposition 1.25. Any array with k rows and m columns can be transformed into reduced row-echelon form by a succession of steps of types (i), (ii), (iii).

[3] "Rows" are understood to be horizontal, while "columns" are vertical.

[4] In the above displayed matrix, the array has $m = n + 1$ columns.

In fact, the transformation in the proposition is carried out by an algorithm known as the method of **row reduction** of the array. Let us begin with an example, indicating the particular operation at each stage by a label over an arrow $\mapsto$. To keep the example from being unwieldy, we consolidate steps of type (iii) into a single step when the "other row" is the same.

EXAMPLE. In this example, $k = m = 4$. Row reduction gives

$$\begin{pmatrix} 0 & 0 & 2 & 7 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & -4 & 5 \\ -2 & 2 & -5 & 4 \end{pmatrix} \overset{\text{(i)}}{\mapsto} \begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & 0 & 2 & 7 \\ -1 & 1 & -4 & 5 \\ -2 & 2 & -5 & 4 \end{pmatrix} \overset{\text{(iii)}}{\mapsto} \begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & 0 & 2 & 7 \\ 0 & 0 & -3 & 6 \\ 0 & 0 & -3 & 6 \end{pmatrix}$$

$$\overset{\text{(ii)}}{\mapsto} \begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & 0 & 1 & \frac{7}{2} \\ 0 & 0 & -3 & 6 \\ 0 & 0 & -3 & 6 \end{pmatrix} \overset{\text{(iii)}}{\mapsto} \begin{pmatrix} 1 & -1 & 0 & -\frac{5}{2} \\ 0 & 0 & 1 & \frac{7}{2} \\ 0 & 0 & 0 & \frac{33}{2} \\ 0 & 0 & 0 & \frac{33}{2} \end{pmatrix} \overset{\text{(ii)}}{\mapsto} \begin{pmatrix} 1 & -1 & 0 & -\frac{5}{2} \\ 0 & 0 & 1 & \frac{7}{2} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{33}{2} \end{pmatrix}$$

$$\overset{\text{(iii)}}{\mapsto} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The final matrix here is in reduced row-echelon form. In the notation of the definition, the number of nonzero rows in the reduced row-echelon form is $l = 3$, and the integers $j(i)$ are $j(1) = 1$, $j(2) = 3$, and $j(3) = 4$.

The example makes clear what the algorithm is that proves Proposition 1.25. We find the first nonzero column, apply an interchange (an operation of type (i)) if necessary to make the first entry in the column nonzero, multiply by a nonzero scalar to make the first entry 1 (an operation of type (ii)), and apply operations of type (iii) to eliminate the other nonzero entries in the column. Then we look for the next column with a nonzero entry in entries 2 and later, interchange to get the nonzero entry into entry 2 of the column, multiply to make the entry 1, and apply operations of type (iii) to eliminate the other entries in the column. Continuing in this way, we arrive at reduced row-echelon form.

In the general case, as soon as our array, which contains both sides of our system of equations, has been transformed into reduced row-echelon form, we can read off exactly what the solutions are. It will be handy to distinguish two kinds of variables among $x_1, \ldots, x_n$ without including any added variables $x_{n+1}, \ldots, x_m$ in either of the classes. The **corner variables** are those x_j's for which j is $\leq n$ and is some $j(i)$ in the definition of "reduced row-echelon form," and the other x_j's with $j \leq n$ will be called **independent variables**. Let us describe the last steps of the solution technique in the setting of an example. We restore the vertical line that separated the data on the two sides of the equations.

EXAMPLE. We consider what might happen to a certain system of 4 equations in 4 unknowns. Putting the data in place for the right side makes the array have 4 rows and 5 columns. We transform the array into reduced row-echelon form and suppose that it comes out to be

$$\left(\begin{array}{cccc|c} 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 \text{ or } 0 \end{array}\right).$$

If the lower right entry is 1, there are no solutions. In fact, the last row corresponds to an equation $0 = 1$, which announces a contradiction. More generally, if any row of 0's to the left of the vertical line is equal to something nonzero, there are no solutions. In other words, there are no solutions to a system if the reduced row-echelon form of the entire array has more nonzero rows than the reduced row-echelon form of the part of the array to the left of the vertical line.

On the other hand, if the lower right entry is 0, then there are solutions. To see this, we restore the reduced array to a system of equations:

$$\begin{aligned} x_1 - x_2 \qquad\qquad &= 1, \\ x_3 \qquad &= 2, \\ x_4 &= 3; \end{aligned}$$

we move the independent variables (namely x_2 here) to the right side to obtain

$$\begin{aligned} x_1 &= 1 + x_2, \\ x_3 &= 2, \\ x_4 &= 3; \end{aligned}$$

and we collect everything in a tidy fashion as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 3 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The independent variables are allowed to take on arbitrary values, and we have succeeded in giving a formula for the solution that corresponds to an arbitrary set of values for the independent variables.

The method in the above example works completely generally. We obtain solutions whenever each row of 0's to the left of the vertical line is matched by a 0 on the right side, and we obtain no solutions otherwise. In the case that we are solving several systems with the same left sides, solutions exist for each of the systems if the reduced row-echelon form of the entire array has the same number of nonzero rows as the reduced row-echelon form of the part of the array to the left of the vertical line.

Let us record some observations about the method for solving systems of linear equations and then some observations about the method of row reduction itself.

Proposition 1.26. In the solution process for a system of k linear equations in n variables with the vertical line in place,

(a) the sum of the number of corner variables and the number of independent variables is n,

(b) the number of corner variables equals the number of nonzero rows on the left side of the vertical line and hence is $\le k$,

(c) when solutions exist, they are of the form

$$\text{column} + \begin{matrix}\text{independent}\\ \text{variable}\end{matrix} \times \text{column} + \cdots + \begin{matrix}\text{independent}\\ \text{variable}\end{matrix} \times \text{column}$$

in such a way that each independent variable x_j is a free parameter in $\mathbb{F}$, the column multiplying x_j has a 1 in its j^{th} entry, and the other columns have a 0 in that entry,

(d) a **homogeneous system**, i.e., one with all right sides equal to 0, has a nonzero solution if the number k of equations is $<$ the number n of variables,

(e) the solutions of an **inhomogeneous system**, i.e., one in which the right sides are not necessarily all 0, are all given by the sum of any one particular solution and an arbitrary solution of the corresponding homogeneous system.

PROOF. Conclusions (a), (b), and (c) follow immediately by inspection of the solution method. For (d), we observe that no contradictory equation can arise when the right sides are 0 and, in addition, that there must be at least one independent variable by (a) since (b) shows that the number of corner variables is $\le k < n$. Conclusion (e) is apparent from (c), since the first column in the solution written in (c) is a column of 0's in the homogeneous case. □

Proposition 1.27. For an array with k rows and n columns in reduced row-echelon form,

(a) the sum of the number of corner variables and the number of independent variables is n,

(b) the number of corner variables equals the number of nonzero rows and hence is $\le k$,

(c) when $k = n$, either the array is of the form

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

or else it has a row of 0's.

PROOF. Conclusions (a) and (b) are immediate by inspection. In (c), failure of the reduced row-echelon form to be as indicated forces there to be some noncorner variable, so that the number of corner variables is $< n$. By (b), the number of nonzero rows is $< n$, and hence there is a row of 0's. □

One final comment: For the special case of n equations in n variables, some readers may be familiar with a formula known as "Cramer's rule" for using determinants to solve the system when the determinant of the array of coefficients on the left side of the vertical line is nonzero. Determinants, including their evaluation, and Cramer's rule will be discussed in Chapter II. The point to make for current purposes is that the use of Cramer's rule for computation is, for n large, normally a more lengthy process than the method of row reduction. In fact, Problem 13 at the end of this chapter shows that the number of steps for solving the system via row reduction is at most a certain multiple of n^3. On the other hand, the typical number of steps for solving the system by rote application of Cramer's rule is approximately a multiple of n^4.

6. Matrix Operations

A rectangular array of scalars (i.e., members of $\mathbb{F}$) with k rows and n columns is called a k-by-n matrix. More precisely a k-**by**-n **matrix** over $\mathbb{F}$ is a function from $\{1, \dots, k\} \times \{1, \dots, n\}$ to $\mathbb{F}$. The expression "k-by-n" is called the **size** of the matrix. The value of the function at the ordered pair (i, j) is often indicated with subscript notation, such as a_{ij}, rather than with the usual function notation $a(i, j)$. It is called the $(i, j)^{\text{th}}$ **entry**. Two matrices are **equal** if they are the same function on ordered pairs; this means that they have the same size and their corresponding entries are equal. A matrix is called **square** if its number of rows equals its number of columns. A square matrix with all entries 0 for $i \neq j$ is called **diagonal**, and the entries with $i = j$ are the **diagonal entries**.

As the reader likely already knows, it is customary to write matrices in rectangular patterns. By convention the first index always tells the number of the row and the second index tells the number of the column. Thus a typical 2-by-3 matrix is $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$. In the indication of the size of the matrix, here 2-by-3, the 2 refers to the number of rows and the 3 refers to the number of columns.

An n-dimensional **row vector** is a 1-by-n matrix, while a k-dimensional **column vector** is a k-by-1 matrix. The set of all k-dimensional column vectors is denoted by $\mathbb{F}^k$. The set $\mathbb{F}^k$ is to be regarded as the space of all ordinary garden-variety vectors. For economy of space, books often write such vectors horizontally with entries separated by commas, for example as (c_1, c_2, c_3), and it is extremely important to treat such vectors as *column* vectors, not as row vectors, in order to get matrix operations and the effect of linear transformations to correspond nicely.[5] Thus in this book, (c_1, c_2, c_3) *is to be regarded as a space-saving way of writing the column vector* $\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$.

If a matrix is denoted by some letter like A, its $(i, j)^{\text{th}}$ entry will typically be denoted by A_{ij}. In the reverse direction, sometimes a matrix is assembled from its individual entries, which may be expressions depending on i and j. If some such expression a_{ij} is given for each pair (i, j), then we denote the corresponding matrix by $[a_{ij}]_{\substack{i=1,\dots,k \\ j=1,\dots,n}}$, or simply by $[a_{ij}]$ if there is no possibility of confusion.

Various operations are defined on matrices. Specifically let $M_{kn}(\mathbb{F})$ be the set of k-by-n matrices with entries in $\mathbb{F}$, so that $M_{k1}(\mathbb{F})$ is the same thing as $\mathbb{F}^k$. **Addition** of matrices is defined whenever two matrices have the same size, and it is defined entry by entry; thus if A and B are in $M_{kn}(\mathbb{F})$, then $A+B$ is the member of $M_{kn}(\mathbb{F})$ with $(A+B)_{ij} = A_{ij} + B_{ij}$. **Scalar multiplication** on matrices is defined entry by entry as well; thus if A is in $M_{kn}(\mathbb{F})$ and c is in $\mathbb{F}$, then cA is the member of $M_{kn}(\mathbb{F})$ with $(cA)_{ij} = cA_{ij}$. The matrix $(-1)A$ is denoted by $-A$. The k-by-n matrix with 0 in each entry is called a **zero matrix**. Ordinarily it is denoted simply by 0; if some confusion is possible in a particular situation, more precise notation will be introduced at the time. With these operations the set $M_{kn}(\mathbb{F})$ has the following properties:

(i) the operation of addition satisfies

(a) $A + (B + C) = (A + B) + C$ for all A, B, C in $M_{kn}(\mathbb{F})$ (associative law),
(b) $A + 0 = 0 + A = A$ for all A in $M_{kn}(\mathbb{F})$,
(c) $A + (-A) = (-A) + A = 0$ for all A in $M_{kn}(\mathbb{F})$,
(d) $A + B = B + A$ for all A and B in $M_{kn}(\mathbb{F})$ (commutative law);

[5] The alternatives are unpleasant. Either one is forced to write certain functions in the unnatural notation $x \mapsto (x)f$, or the correspondence is forced to involve transpose operations on frequent occasions. Unhappily, books following either of these alternative conventions may be found.

(ii) the operation of scalar multiplication satisfies

(a) $(cd)A = c(dA)$ for all A in $M_{kn}(\mathbb{F})$ and all scalars c and d,
(b) $1A = A$ for all A in $M_{kn}(\mathbb{F})$ and for the scalar 1;

(iii) the two operations are related by the distributive laws

(a) $c(A+B) = cA + cB$ for all A and B in $M_{kn}(\mathbb{F})$ and for all scalars c,
(b) $(c+d)A = cA + dA$ for all A in $M_{kn}(\mathbb{F})$ and all scalars c and d.

Since addition and scalar multiplication are defined entry by entry, all of these identities follow from the corresponding identities for members of $\mathbb{F}$.

Multiplication of matrices is defined in such a way that the kind of system of linear equations discussed in the previous section can be written as a matrix equation in the form $AX = B$, where

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \\ a_{k1} & \cdots & a_{kn} \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

More precisely if A is a k-by-m matrix and B is an m-by-n matrix, then the product $C = AB$ is the k-by-n matrix defined by

$$C_{ij} = \sum_{l=1}^{m} A_{il} B_{lj}.$$

The $(i, j)^{\text{th}}$ entry of C is therefore the product of the i^{th} row of A and the j^{th} column of B.

Let us emphasize that the condition for a product AB to be defined is that the number of columns of A should equal the number of rows of B. With this definition the system of equations mentioned above is indeed of the form $AX = B$.

Proposition 1.28. Matrix multiplication has the properties that

(a) it is associative in the sense that $(AB)C = A(BC)$, provided that the sizes match correctly, i.e., A is in $M_{km}(\mathbb{F})$, B is in $M_{mn}(\mathbb{F})$, and C is in $M_{np}(\mathbb{F})$,
(b) it is distributive over addition in the sense that $A(B+C) = AB + AC$ and $(B+C)D = BD + CD$ if the sizes match correctly.

REMARK. Matrix multiplication is not necessarily commutative, even for square matrices. For example, $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, while $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

PROOF. For (a), we have

$$((AB)C)_{ij} = \textstyle\sum_{t=1}^{n} (AB)_{it}C_{tj} = \sum_{t=1}^{n} \sum_{s=1}^{m} A_{is}B_{st}C_{tj}$$

and

$$(A(BC))_{ij} = \textstyle\sum_{s=1}^{m} A_{is}(BC)_{sj} = \sum_{s=1}^{m} \sum_{t=1}^{n} A_{is}B_{st}C_{tj},$$

and these are equal. For the first identity in (b), we have

$$\begin{aligned} (A(B+C))_{ij} &= \textstyle\sum_l A_{il}(B+C)_{lj} = \sum_l A_{il}(B_{lj}+C_{lj}) \\ &= \textstyle\sum_l A_{il}B_{lj} + \sum_l A_{il}C_{lj} = (AB)_{ij} + (AC)_{ij}, \end{aligned}$$

and the second identity is proved similarly. □

We have already defined the zero matrix 0 of a given size to be the matrix having 0 in each entry. This matrix has the property that $0A = 0$ and $B0 = 0$ if the sizes match properly. The n-by-n **identity matrix**, denoted by I or sometimes 1, is defined to be the matrix with $I_{ij} = \delta_{ij}$, where δ_{ij} is the **Kronecker delta** defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

In other words, the identity matrix is the square matrix of the form

$$I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

It has the property that $IA = A$ and $BI = I$ whenever the sizes match properly.

Let A be an n-by-n matrix. We say that A is **invertible** and has the n-by-n matrix B as **inverse** if $AB = BA = I$. If B and C are n-by-n matrices with $AB = I$ and $CA = I$, then associativity of multiplication (Proposition 1.28a) implies that $B = IB = (CA)B = C(AB) = CI = C$. Hence an inverse for A is unique if it exists. We write A^{-1} for this inverse if it exists. Inverses of n-by-n matrices have the property that if A and D are invertible, then AD is invertible and $(AD)^{-1} = D^{-1}A^{-1}$; moreover, if A is invertible, then A^{-1} is invertible and its inverse is A.

The method of row reduction in the previous section suggests a way of computing the inverse of a matrix. Suppose that A is a square matrix to be inverted and we are seeking its inverse B. Then $AB = I$. Examining the definition of matrix multiplication, we see that this matrix equation means that the product of A and the first column of B equals the first column of I, the product of A and the second column of B equals the second column of I, and so on. We can thus think of a column of B as the unknowns in a system of linear equations, the known right sides being the entries of the column of the identity matrix. As the column index varies, the left sides of these equations do not change, since they are always given by A. So we can attempt to solve all of the systems (one for each column) simultaneously. For example, to attempt to invert $A = \left(\begin{smallmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{smallmatrix}\right)$, we set up

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 7 & 8 & 10 & 0 & 0 & 1 \end{array}\right).$$

Imagine doing the row reduction. We can hope that the result will be of the form

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \text{-} & \text{-} & \text{-} \\ 0 & 1 & 0 & \text{-} & \text{-} & \text{-} \\ 0 & 0 & 1 & \text{-} & \text{-} & \text{-} \end{array}\right),$$

with the identity matrix on the left side of the vertical line. If this is indeed the result, then the computation shows that the matrix on the right side of the vertical line is the only possibility for A^{-1}. But does A^{-1} in fact exist?

Actually, another question arises as well. According to Proposition 1.27c, the other possibility in applying row reduction is that the left side has a row of 0's. In this case, can we deduce that A^{-1} does not exist? Or, to put it another way, can we be sure that some row of the reduced row-echelon form has all 0's on the left side of the vertical line and something nonzero on the right side?

All of the answers to these questions are yes, and we prove them in a moment. First we need to see that elementary row operations are given by matrix multiplications.

Proposition 1.29. Each elementary row operation is given by left multiplication by an invertible matrix. The inverse matrix is the matrix of another elementary row operation.

REMARK. The square matrices giving these left multiplications are called **elementary matrices**.

PROOF. For the interchange of rows i and j, the part of the elementary matrix in the rows and columns with i or j as index is

$$\begin{array}{cc} & \begin{array}{cc} i & j \end{array} \\ \begin{array}{c} i \\ j \end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array},$$

and otherwise the matrix is the identity. This matrix is its own inverse.

For the multiplication of the i^{th} row by a nonzero scalar c, the matrix is diagonal with c in the i^{th} diagonal entry and with 1 in all other diagonal entries. The inverse matrix is of this form with c^{-1} in place of c.

For the replacement of the i^{th} row by the sum of the i^{th} row and the product of a times the j^{th} row, the part of the elementary matrix in the rows and columns with i or j as index is

$$\begin{array}{c c} & \begin{array}{cc} i & j \end{array} \\ \begin{array}{c} i \\ j \end{array} & \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, \end{array}$$

and otherwise the matrix is the identity. The inverse of this matrix is the same except that a is replaced by $-a$. □

Theorem 1.30. The following conditions on an n-by-n square matrix A are equivalent:

(a) the reduced row-echelon form of A is the identity,

(b) A is the product of elementary matrices,

(c) A has an inverse,

(d) the system of equations $AX = 0$ with $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ has only the solution $X = 0$.

PROOF. If (a) holds, choose a sequence of elementary row operations that reduce A to the identity, and let $E_1, \dots, E_r$ be the corresponding elementary matrices given by Proposition 1.29. Then we have $E_r \cdots E_1 A = I$, and hence $A = E_1^{-1} \cdots E_r^{-1}$. The proposition says that each E_j^{-1} is an elementary matrix, and thus (b) holds.

If (b) holds, then (c) holds because the elementary matrices are invertible and the product of invertible matrices is invertible.

If (c) holds and if $AX = 0$, then $X = IX = (A^{-1}A)X = A^{-1}(AX) = A^{-1}0 = 0$. Hence (d) holds.

If (d) holds, then the number of independent variables in the row reduction of A is 0. Proposition 1.26a shows that the number of corner variables is n, and parts (b) and (c) of Proposition 1.27 show that the reduced row-echelon form of A is I. Thus (a) holds. □

Corollary 1.31. If the solution procedure for finding the inverse of a square matrix A leads from $(A \mid I)$ to $(I \mid X)$, then A is invertible and its inverse is X. Conversely if the solution procedure leads to $(R \mid Y)$ and R has a row of 0's, then A is not invertible.

REMARK. Proposition 1.27c shows that this corollary addresses the only possible outcomes of the solution procedure.

PROOF. We apply the equivalence of (a) and (c) in Theorem 1.30 to settle the existence or nonexistence of A^{-1}. In the case that A^{-1} exists, we know that the solution procedure has to yield the inverse. □

Corollary 1.32. Let A be a square matrix. If B is a square matrix such that $BA = I$, then A is invertible and B is its inverse. If C is a square matrix such that $AC = I$, then A is invertible with inverse C.

PROOF. Suppose $BA = I$. Let X be a column vector with $AX = 0$. Then $X = IX = (BA)X = B(AX) = B0 = 0$. Since (d) implies (c) in Theorem 1.30, A is invertible.

Suppose $AC = I$. Applying the result of the previous paragraph to C, we conclude that C is invertible with inverse A. Therefore A is invertible with inverse C. □

7. Problems

1. What is the greatest common divisor of 9894 and 11058?

2. (a) Find integers x and y such that $11x + 7y = 1$.
   (b) How are all pairs (x, y) of integers satisfying $11x + 7y = 1$ related to the pair you found in (a)?

3. Let $\{a_n\}_{n\geq 1}$ be a sequence of positive integers, and let d be the largest integer dividing all a_n. Prove that d is the greatest common divisor of finitely many of the a_n.

4. Determine the integers n for which there exist integers x and y such that n divides $x + y - 2$ and $2x - 3y - 3$.

5. Let $P(X)$ and $Q(X)$ be the polynomials $P(X) = X^4 + X^3 + 2X^2 + X + 1$ and $Q(X) = X^5 + 2X^3 + X$ in $\mathbb{R}[X]$.
   (a) Find a greatest common divisor $D(X)$ of $P(X)$ and $Q(X)$.
   (b) Find polynomials A and B such that $AP + BQ = D$.

6. Let $P(X)$ and $Q(X)$ be polynomials in $\mathbb{R}[X]$. Prove that if $D(X)$ is a greatest common divisor of $P(X)$ and $Q(X)$ in $\mathbb{C}[X]$, then there exists a nonzero complex number c such that $cD(X)$ is in $\mathbb{R}[X]$.

7. (a) Let $P(X)$ be in $\mathbb{R}[X]$, and regard it as in $\mathbb{C}[X]$. Applying the Fundamental Theorem of Algebra and its corollary to P, prove that if z_j is a root of P, then so is $\bar{z}_j$, and z_j and $\bar{z}_j$ have the same multiplicity.
   (b) Deduce that any prime polynomial in $\mathbb{R}[X]$ has degree at most 2.

8. (a) Suppose that a polynomial $A(X)$ of degree > 0 in $\mathbb{Q}[X]$ has integer coefficients and leading coefficient 1. Show that if p/q is a root of $A(X)$ with p and q integers such that $\mathrm{GCD}(p, q) = 1$, then p/q is an integer n and n divides the constant term of $A(X)$.
   (b) Deduce that $X^2 - 2$ and $X^3 + X^2 + 1$ are prime in $\mathbb{Q}[X]$.

9. Reduce the fraction 8645/10465 to lowest terms.

10. How many different patterns are there of disjoint cycle structures for permutations of $\{1, 2, 3, 4\}$? Give examples of each, telling how many permutations there are of each kind and what the signs are of each.

11. Prove for $n \geq 2$ that the number of permutations of $\{1, \dots, n\}$ with sign -1 equals the number with sign $+1$.

12. Find all solutions X of the system $AX = B$ when $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ and B is given by
   (a) $B = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, (b) $B = \begin{pmatrix} 5 \\ 3 \\ 2 \end{pmatrix}$, (c) $B = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$.

13. Suppose that a single step in the row reduction process means a single arithmetic operation or a single interchange of two entries. Prove that there exists a constant C such that any square matrix can be transformed into reduced row-echelon form in $\leq Cn^3$ steps, the matrix being of size n-by-n.

14. Compute $A + B$ and AB if $A = \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix}$ and $B = \begin{pmatrix} -4 & 8 \\ -1 & 3 \end{pmatrix}$.

15. Prove that if A and B are square matrices with $AB = BA$, then $(A + B)^n$ is given by the Binomial Theorem: $(A + B)^n = \sum_{k=0}^{n} \binom{n}{k} A^{n-k} B^k$, where $\binom{n}{k}$ is the binomial coefficient $n!/((n-k)!k!)$.

16. Find a formula for the n^{th} power of $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, n being a positive integer.

17. Let D be an n-by-n diagonal matrix with diagonal entries $d_1, \dots, d_n$, and let A be an n-by-n matrix. Compute AD and DA, and give a condition for the equality $AD = DA$ to hold.

18. Fix n, and let E_{ij} denote the n-by-n matrix that is 1 in the $(i, j)^{\text{th}}$ entry and is 0 elsewhere. Compute the product $E_{kl}E_{pq}$, expressing the result in terms of matrices E_{ij} and instances of the Kronecker delta.

19. Verify that if $ad - bc \neq 0$, then $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = (ad-bc)^{-1} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ and that the system $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}$ has the unique solution $\begin{pmatrix} x \\ y \end{pmatrix} = (ad - bc)^{-1} \begin{pmatrix} dp-bq \\ aq-cp \end{pmatrix}$.

20. Which of the following matrices A is invertible? For the invertible ones, find A^{-1}.

    (a) $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$, (b) $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{pmatrix}$, (c) $A = \begin{pmatrix} 7 & 4 & 1 \\ 6 & 4 & 1 \\ 4 & 3 & 1 \end{pmatrix}$.

21. Can a square matrix with a row of 0's be invertible? Why or why not?

22. Prove that if the product AB of two n-by-n matrices is invertible, then A and B are invertible.

23. Let A be a square matrix such that $A^k = 0$ for some positive integer n. Prove that $I + A$ is invertible.

24. Give an example of a set S and functions $f : S \to S$ and $g : S \to S$ such that the composition $g \circ f$ is the identity function but neither f nor g has an inverse function.

25. Give an example of two matrices, A of size 1-by-2 and B of size 2-by-1, such that $AB = I$, I being the 1-by-1 identity matrix. Verify that BA is not the 2-by-2 identity matrix. Give a proof for these sizes that BA can never be the identity matrix.

Problems 26–29 concern least common multiples. Let a and b be positive integers. A **common multiple** of a and b is an integer N such that a and b both divide N. The **least common multiple** of a and b is the smallest positive common multiple of a and b. It is denoted by $\mathrm{LCM}(a, b)$.

26. Prove that a and b have a least common multiple.

27. If a has a prime factorization given by $a = p_1^{k_1} \cdots p_r^{k_r}$, prove that any positive multiple M of a has a prime factorization given by $a = p_1^{m_1} \cdots p_r^{m_r} q_1^{n_1} \cdots q_s^{n_s}$, where $q_1, \dots, q_s$ are primes not in the list $p_1, \dots, p_r$, where $m_j \geq k_j$ for all j, and where $n_j \geq 0$ for all j.

28. (a) Prove that if $a = p_1^{k_1} \cdots p_r^{k_r}$ and $b = p_1^{l_1} \cdots p_r^{l_r}$ are expansions of a and b as products of powers of r distinct primes $p_1, \dots, p_r$, then $\mathrm{LCM}(a, b) = p_1^{\max(k_1, l_1)} \cdots p_r^{\max(k_r, l_r)}$.

    (b) Prove that if N is any common multiple of a and b, then $\mathrm{LCM}(a, b)$ divides N.

    (c) Deduce that $ab = \mathrm{GCD}(a, b)\ \mathrm{LCM}(a, b)$.

29. If $a_1, \dots, a_t$ are positive integers, define their **least common multiple** to be the smallest positive integer M such that each a_j divides M. Give a formula for this M in terms of expansions of $a_1, \dots, a_t$ as products of powers of distinct primes.