MA281: Introduction to Linear Algebra

Dylan C. Beck

Acknowledgements

Primarily, the contents of this document were created in the Fall 2022 semester at Baker University with substantial revisions and reorganization efforts unfolding during the Summer 2023. I express my sincere gratitude toward all of my students in MA281 (Introduction to Linear Algebra), but I am especially indebted to those who assisted in the enhancement of these notes with comments and suggestions, including Sydney Delfelder, Owen Gonzales, Ben Stubblefield, and April Thomas.

Elsewhere, some of the material from Chapter 2 (e.g., regarding the Smith Normal Form, the Rational Canonical Form, and the Jordan Canonical Form) was written independently from August 2018 to January 2022 at the request of students from the Algebra Qualifying Exam Study Group at the University of Kansas. I would like to thank the graduate students there for their motivation and participation in these study sessions — especially Wayne Ng Kwing King, Enrique Salcido, Neethu Suma-Raveendran, and Christopher Wong. I am also grateful to Souvik Dey and Monalisa Dutta for their assistance with these meetings and their peripheral contributions to these notes.

Contents

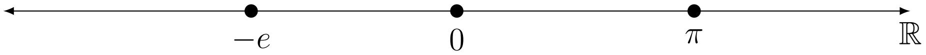
Chapter 1

Vectors and Matrices

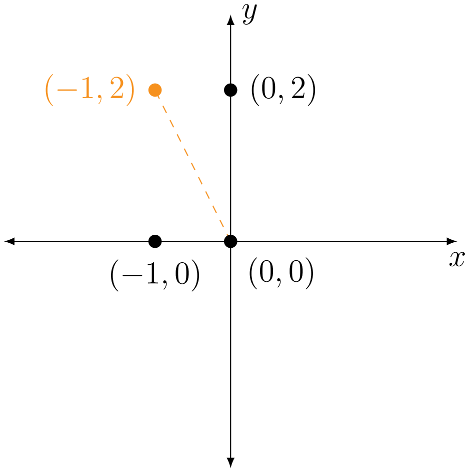
Often, in dealing with real-word problems, we are immediately met with large amounts of data and information. Even an activity as simple as baking a cake requires many ingredients and steps that must be completed in careful order, and the complexity of a task may grow exponentially as the number of inputs increases. One way to efficiently organize data is according to rows and columns in what we will refer to as vectors and matrices. We will demonstrate in this chapter that vectors and matrices admit an arithmetic that yields a highly sophisticated and widely applicable theory.

1.1 Real n-Space

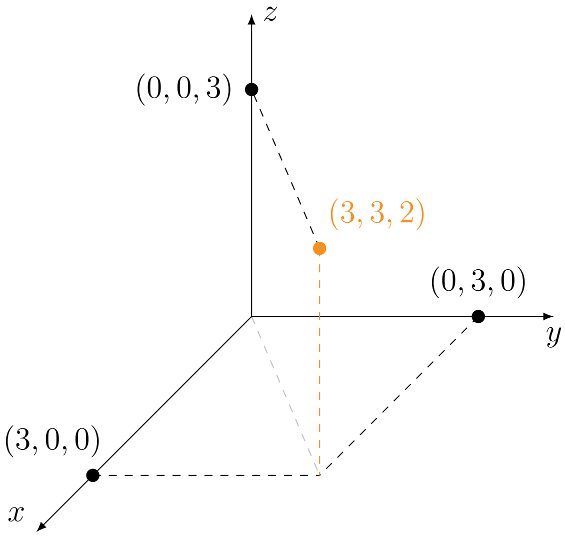
Consider the set $\mathbb{R}$ consisting of real numbers. Like usual, we may geometrically realize $\mathbb{R}$ as a line (the real number line) consisting of **points** x that lie a distance of $|x|$ from the **origin** 0 for each real number x. Explicitly, the point π lies π units to the right of the origin, whereas the point $-e$ lies e units to the left of the origin. Given any pair of real numbers $a \leq b$, the distance between the points a and b along the real number line is given by the length of the closed interval $[a, b]$; we learn in Calculus I that this distance is exactly the real number $b - a$. Consequently, the real numbers $\mathbb{R}$ admit a notion of geometry since we can conceive of things like lines and distances. Below is a visual representation of the real number line with the points $-e$, 0, and π plotted for reference.



Observe that forward and backward are the only two directions along the real number line, hence the geometry of $\mathbb{R}$ is in this sense quite simple. On the other hand, suppose that we want to keep track of both east-west movement and north-south movement. Given that an object lies x units from the origin in the east-west direction and y units in the north-south direction, we may canonically express this data as an **ordered pair** (x, y). Explicitly, if a particle lies 1 unit west and 2 units north of the origin $(0, 0)$, then it lies 1 unit to the left of the origin on the x-axis and 2 units north of the origin on the y-axis; the location of the particle in this case can be written as the ordered pair $(-1, 2)$. We refer to the collection of all ordered pairs of real numbers (x, y) as the **Cartesian product** $\mathbb{R} \times \mathbb{R}$ of the real numbers with itself, i.e., we have that $\mathbb{R} \times \mathbb{R} = \{(x, y) \mid x \text{ and } y \text{ are real numbers}\}$. Graphically, the totality of points in $\mathbb{R} \times \mathbb{R}$ form a plane, so $\mathbb{R} \times \mathbb{R}$ is often called the **Cartesian plane**. Conventionally, the Cartesian plane is denoted by $\mathbb{R}^2$ and referred to also as **real 2-space**.

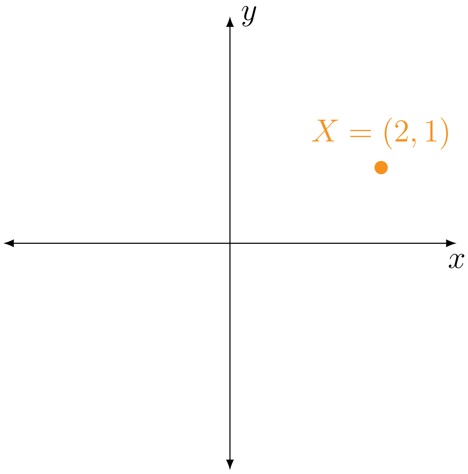


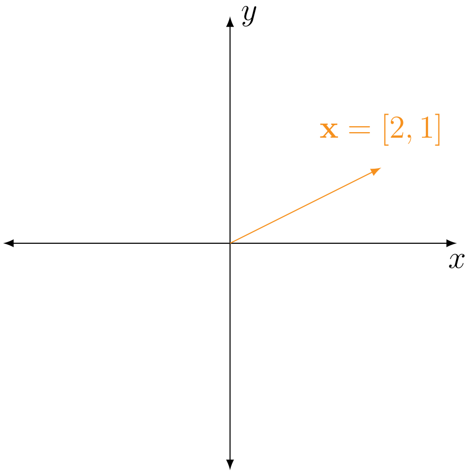
Going one step further, let us keep track of east-west, north-south, and up-down movements. Explicitly, if x measures the location of a particle in the x-axis; y measures the location of a particle in the y-axis; and z measures the location of particle in the z-axis, then the ordered triple (x, y, z) conveniently encodes this information. Like before, if the particle lies 3 units east of the origin, 3 units north of the origin, and 1 unit above the origin, then the particle's location is determined by the ordered triple $(3, 3, 1)$. We denote by $\mathbb{R}^3$ the collection of all ordered triples of real numbers, i.e., we have that $\mathbb{R}^3 = \{(x, y, z) \mid x, y, \text{ and } z \text{ are real numbers}\}$; we refer to $\mathbb{R}^3$ as real 3-space.



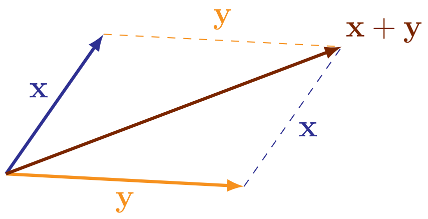
Once and for all, if n is a positive integer, then we will denote by $\mathbb{R}^n$ the collection of all n-**tuples** of real numbers, i.e., we have that $\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) \mid x_1, x_2, \dots, x_n \text{ are real numbers}\}$. We will typically use a capital letter X to denote a real n-tuple $(x_1, x_2, \dots, x_n)$. We refer to the real number x_1 as the first **coordinate** of X; we refer to the real number x_2 as the second coordinate of X; we refer to the real number x_n as the nth coordinate of X; and in general, the real number x_i is called the ith coordinate of X for each integer $1 \leq i \leq n$. Every point in real n-space is uniquely determined by its coordinates: indeed, if we consider any pair of points $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ such that $(x_1, x_2, \dots, x_n) = X = Y = (y_1, y_2, \dots, y_n)$, then each of the coordinates on the left-hand side must be equal to the corresponding coordinate on the right-hand side, i.e., we must have that $x_i = y_i$ for all integers $1 \leq i \leq n$. Even though it is not possible to geometrically visualize points in real n-space for any integer $n \geq 4$, it is still meaningful to discuss this notion. Explicitly, every set of data consisting of n distinct real parameters induces an element of real n-space $\mathbb{R}^n$.

Continuing from a geometric perspective, it is useful to distinguish between points and **vectors** in real n-space. Explicitly, we may view the vector $\mathbf{x} = [x_1, x_2, \dots, x_n]$ corresponding to the point $X = (x_1, x_2, \dots, x_n)$ in real n-space as a ray (or arrow) emanating from the origin and extending to the point $(x_1, x_2, \dots, x_n)$. Explicitly, the vector $\mathbf{x} = [1, 2, 3, 4]$ of $\mathbb{R}^4$ can be represented by the ray extending from the origin $(0, 0, 0, 0)$ to the point $(1, 2, 3, 4)$ in $\mathbb{R}^4$. We refer to the vector $\mathbf{x}$ in this case as lying in **standard position**. We will come to find that despite the mathematical equivalence of points X and vectors $\mathbf{x}$ in real n-space, the benefit of this distinction is that vectors in real n-space are translation-invariant and possess a notion of length. Often, we will restrict our attention to the Cartesian plane $\mathbb{R}^2$ or real 3-space $\mathbb{R}^3$, where we can visualize these vectors.

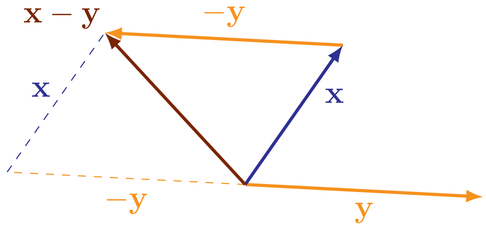




Geometrically, we may prescribe the arithmetic of **vector addition** as follows: to determine the vector sum $\mathbf{x} + \mathbf{y}$ pictorially, visualize $\mathbf{x}$ and $\mathbf{y}$ as rays emanating from the origin; translate $\mathbf{y}$ so that the "foot" of $\mathbf{y}$ lies at the "head" of $\mathbf{x}$; and draw the ray emanating from the "foot" of $\mathbf{x}$ to the "head" of $\mathbf{y}$. Equivalently, one could also determine $\mathbf{x} + \mathbf{y}$ by translating $\mathbf{x}$ so that the "foot" of $\mathbf{x}$ lies at the "head" of $\mathbf{y}$ and subsequently drawing the raw emanating from the "foot" of $\mathbf{y}$ to the "head" of $\mathbf{x}$. Either way, the resulting **vector sum** can be pictured as follows.



We refer to the process of computing the vector sum $\mathbf{x} + \mathbf{y}$ in this manner as the **Parallelogram Law** because the resulting diagram forms a parallelogram. We will in no time describe the algebraic operations of vector addition and scalar multiplication, but for now, we note that for any vector $\mathbf{x}$ emanating from the origin to a point X in real n-space, the point $-X$ is obtained from X by taking the coordinates of X with opposite sign. Explicitly, if we assume that $X = (x_1, x_2, \dots, x_n)$, then $-X = (-x_1, -x_2, \dots, -x_n)$. By identifying the vector $-\mathbf{x}$ in standard position with the ray emanating from the origin to the point $-X$, we find that $-\mathbf{x}$ is nothing more than $\mathbf{x}$ in the "opposite direction.” Consequently, translating $-\mathbf{x}$ so that it overlaps $\mathbf{x}$, the “head” of $-\mathbf{x}$ lies at the “foot” of $\mathbf{x}$ and vice-versa. We may in this way describe **vector subtraction** pictorially as follows.



Even more, **scalar multiplication** of a vector $\mathbf{x}$ by a real number (or **scalar**) α can be visualized by taking the vector $\alpha\mathbf{x}$ as the ray emanating from the origin with length $|\alpha|$ times the length of $\mathbf{x}$ in the same direction of $\mathbf{x}$ if α is positive and in the opposite direction if α is negative. We will henceforth say that two vectors $\mathbf{x}$ and $\mathbf{y}$ in real n-space are **parallel** if there exists a nonzero real number α such that $\mathbf{y} = \alpha\mathbf{x}$. We will say that $\mathbf{x}$ and $\alpha\mathbf{x}$ have the **same direction** if $\alpha > 0$; they have the **opposite direction** if $\alpha < 0$; and the vector $\mathbf{0}$ corresponding to the origin has no direction. Certainly, a pair of vectors in real n-space need not be parallel, hence in general, it might not be possible to say that an arbitrary pair of vectors have the same or opposite direction.

Until now, we have considered vectors in real n-space from a primarily geometric standpoint by way of diagrams and visualizations; however, this might very well come across as unsatisfactory to some readers for several reasons — not least of all that it is difficult to draw vectors in three-space and impossible to picture vectors with more coordinates than that. Bearing this in mind, we turn our attention to an algebraic description of vectors in real n-space. We will to this end represent vectors $\mathbf{v}$ and $\mathbf{w}$ in real n-space according to their coordinates. Explicitly, we will write $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ for some positive integer n and real numbers $v_1, v_2, \ldots, v_n$. Given any positive integer m and any real numbers $w_1, w_2, \ldots, w_m$, the vectors $\mathbf{v}$ and $\mathbf{w}$ are equal (i.e., $\mathbf{v} = \mathbf{w}$) if and only if $m = n$ and $w_i = v_i$ for each integer $1 \leq i \leq n$. Concretely, a pair of vectors expressed in terms of their coordinates are equal if and only if (1.) the number of coordinates of the vectors is the same and (2.) the corresponding coordinates of the vectors are the same. We reserve the notation $\mathbf{0}$ for the **zero vector** whose coordinates are all zero, i.e., $\mathbf{0} = [0, 0, \ldots, 0]$. Crucially, all though we will indiscriminately use the symbol $\mathbf{0}$ to denote the zero vector in all contexts, it is important to realize that the zero vector in real n-space differs as n ranges across all positive integers.

We define vector addition and scalar multiplication **coordinatewise**. Explicitly, for any vectors $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ and $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ in real n-space and any real number α, we declare that

$$\begin{aligned} \mathbf{v}+\mathbf{w} &= [v_1+w_1, v_2+w_2, \ldots, v_n+w_n] \text{ and} \\ \alpha\mathbf{v} &= [\alpha v_1, \alpha v_2, \ldots, \alpha v_n]. \end{aligned}$$

Consequently, it follows that vector subtraction is carried out componentwise, as well.

$$\mathbf{v}-\mathbf{w} = [v_1-w_1, v_2-w_2, \ldots, v_n-w_n]$$

Example 1.1.1. Consider the vectors $\mathbf{u} = [1, 1, -1]$, $\mathbf{v} = [1, 2, 3]$, and $\mathbf{w} = [0, -2, -2]$ in real 3-space. Observe that $\mathbf{u}+\mathbf{v} = [2, 3, 2]$, $-\mathbf{w} = [0, 2, 2]$, $\mathbf{v}-\mathbf{w} = [1, 4, 5]$, and $3\mathbf{u} = [3, 3, -3]$.

Example 1.1.2. Observe that the vectors $\mathbf{u} = [1, 0, -1]$ and $\mathbf{v} = [-3, 0, 3]$ are parallel because we have that $\mathbf{v} = -3\mathbf{u}$, hence $\mathbf{u}$ and $\mathbf{v}$ have the opposite direction; however, the vector $\mathbf{w} = [-1, 1, 1]$ is not parallel to either $\mathbf{u}$ or $\mathbf{v}$. (We will soon see that it is in fact perpendicular to $\mathbf{u}$ and $\mathbf{v}$.)

Considering that vector addition and scalar multiplication in real n-space are determined by the coordinates of the underlying vectors, the following proposition should not come as a surprise.

Proposition 1.1.3 (Properties of Vector Arithmetic in Real n-Space)**.** *Consider any vectors* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$ *in real n-space and any real numbers α and β. We have that*

1.) *vector addition is* **associative**, *i.e.,* $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$;

2.) *vector addition is* **commutative**, *i.e.,* $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$;

3.) *the zero vector* $\mathbf{0}$ *is the* **additive identity**, *i.e.,* $\mathbf{v} + \mathbf{0} = \mathbf{v}$;

4.) *the* **additive inverse** *of* $\mathbf{v}$ *is* $-\mathbf{v}$, *i.e.,* $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$;

5.) *scalar multiplication is associative, i.e.,* $\alpha(\beta\mathbf{v}) = (\alpha\beta)\mathbf{v}$;

6.) *scalar multiplication is* **distributive** *across vector addition, i.e.,* $\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}$;

7.) *scalar multiplication is distributive across scalar addition, i.e.,* $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$; *and*

8.) *the* **multiplicative identity** 1 *preserves scale, i.e.,* $1\mathbf{v} = \mathbf{v}$.

Proof. Each of the above properties can be verified directly by listing the coordinates of the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ and performing the vector addition and scalar multiplication coordinatewise. □

Example 1.1.4. Consider the vectors $\mathbf{u} = [1, 2, 5]$, $\mathbf{v} = [-1, 3, 4]$, and $\mathbf{w} = [3, 1, 6]$ in real 3-space. We can compute $3\mathbf{u} - 5(\mathbf{v} - \mathbf{w})$ according to the Properties of Vector Arithmetic in Real n-Space.

$$\begin{aligned} 3\mathbf{u} - 5(\mathbf{v} - \mathbf{w}) &= 3\mathbf{u} - 5\mathbf{v} + 5\mathbf{w} \\ &= 3[1, 2, 5] - 5[-1, 3, 4] + 5[3, 1, 6] \\ &= [3, 6, 15] + [5, -15, -20] + [15, 5, 30] \\ &= [23, -4, 25] \end{aligned}$$

We could alternatively computed the vector difference $\mathbf{v} - \mathbf{w} = [-4, 2, -2]$ and proceeded as follows.

$$3\mathbf{u} - 5(\mathbf{v} - \mathbf{w}) = 3[1, 2, 5] - 5[-4, 2, -2] = [3, 6, 15] + [20, -10, 10] = [23, -4, 25]$$

Either way, we obtain the same coordinates for the vector $3\mathbf{u} - 5(\mathbf{v} - \mathbf{w})$, as expected.

We refer to the vector $3\mathbf{u} - 5\mathbf{v} + 5\mathbf{w}$ in Example 1.1.4 as a **linear combination** of the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. Generally, for any vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ in real n-space and any scalars $\alpha_1, \alpha_2, \dots, \alpha_k$, we refer to the vector $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_k\mathbf{v}_k$ as the linear combination of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ with **scalar coefficients** $\alpha_1, \alpha_2, \dots, \alpha_k$. Later, these vectors will become a critical object of study; however, for now, it is important to note that every vector $\mathbf{v}$ in real n-space can be written uniquely as a linear combination of the **standard basis vectors** $\mathbf{e}_i$ whose ith coordinate is 1 and whose other coordinates are 0 for all integers $1 \leq i \leq n$. Concretely, there are two standard basis vectors in real 2-space: they are $\mathbf{e}_1 = [1, 0]$ and $\mathbf{e}_2 = [0, 1]$. Likewise, there are three standard basis vectors in real 3-space — namely, $\mathbf{e}_1 = [1, 0, 0]$, $\mathbf{e}_2 = [0, 1, 0]$, and $\mathbf{e}_3 = [0, 0, 1]$. Observe that

$$[v_1, v_2, v_3] = [v_1, 0, 0] + [0, v_2, 0] + [0, 0, v_3] = v_1[1, 0, 0] + v_2[0, 1, 0] + v_3[0, 0, 1] = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3$$

yields an expression of the vector $[v_1, v_2, v_3]$ as a linear combination of the standard basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ with scalar coefficients corresponding to the coordinates of $\mathbf{v}$. By analogy, this process can be carried out for any vector in real n-space with respect to the standard basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$.

Example 1.1.5. Consider the vectors $\mathbf{u} = [1, 2, 5]$, $\mathbf{v} = [-1, 3, 4]$, and $\mathbf{w} = [3, 1, 6]$ in real 3-space. Let us verify that $\mathbf{w}$ is a linear combination of $\mathbf{u}$ and $\mathbf{v}$. By definition, we must find real numbers α and β such that $\mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v}$. Expressing this relation in terms of coordinates yields that

$$[3, 1, 6] = \mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v} = \alpha[1, 2, 5] + \beta[-1, 3, 4] = [\alpha - \beta, 2\alpha + 3\beta, 5\alpha + 4\beta],$$

so it suffices to solve the induced **system of equations** with three equations and two unknowns.

$$\begin{cases} \alpha - \beta = 3 \\ 2\alpha + 3\beta = 1 \\ 5\alpha + 4\beta = 6 \end{cases}$$

By the first equation, it follows that $\alpha = \beta + 3$; substitute this into the second equation to find that

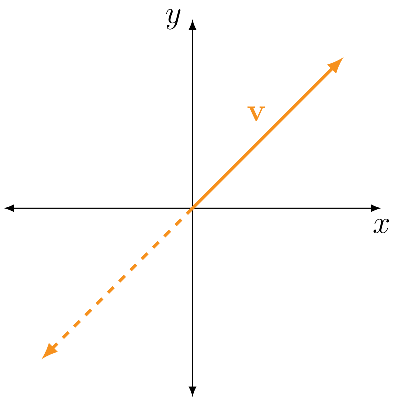
$$1 = 2\alpha + 3\beta = 2(\beta + 3) + 3\beta = 5\beta + 6,$$

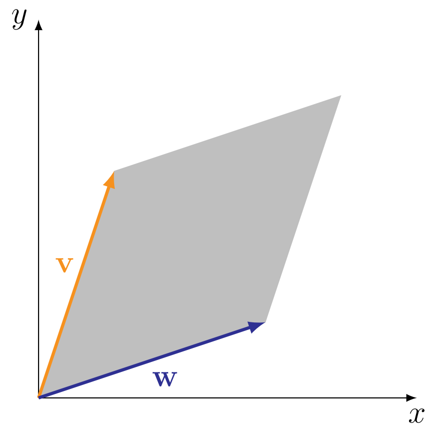
hence we conclude that $\beta = -1$, from which it follows that $\alpha = 2$. We conclude at last that

$$[3, 1, 2] = \mathbf{w} = 2\mathbf{u} - \mathbf{v} = 2[1, 2, 5] - [-1, 3, 4].$$

We remark that the third equation $5\alpha + 4\beta = 6$ was not required to solve this system.

Geometrically, linear combinations of vectors give rise to lines, planes, and hyperplanes. Explicitly, given any nonzero vectors $\mathbf{v}$ and $\mathbf{w}$ in real n-space, the collection $\{\alpha\mathbf{v} \mid \alpha \in \mathbb{R}\}$ of all possible linear combinations of $\mathbf{v}$ is called the **line along v**, and the collection $\{\alpha\mathbf{v} + \beta\mathbf{w} \mid \alpha, \beta \in \mathbb{R}\}$ of all possible linear combinations of $\mathbf{v}$ and $\mathbf{w}$ is called the **plane spanned by v and w**.





Example 1.1.6. Consider the vectors $\mathbf{v} = [1, 3]$ and $\mathbf{w} = [3, 1]$ in real 2-space. By definition, the line along $\mathbf{v}$ is given by the collection of points $(\alpha, 3\alpha)$ such that α is a real number. Consequently, the points $(0, 0)$, $(3, 9)$, and $(-2, -6)$ lie on the line along $\mathbf{v}$; however, the point $(2, 2)$ does not lie on this line: indeed, the point $(2, 2)$ lies on the line along $\mathbf{v}$ if and only if there exists a real number α such that $(2, 2) = (\alpha, 3\alpha)$ if and only if $\alpha = 2$ and $3\alpha = 2$. Because this is impossible, the point $(2, 2)$ does not lie on the lie along $\mathbf{v}$. We say in this case that the system of equations

$$\begin{cases} \alpha = 2 \\ 3\alpha = 2 \end{cases}$$

is **inconsistent** because there is no real number α for which both equations hold.

Likewise, if we wish to determine if the point $(12, 12)$ lies in the planned spanned by $\mathbf{v}$ and $\mathbf{w}$, we seek real numbers α and β such that $[12, 12] = \alpha\mathbf{v} + \beta\mathbf{w} = \alpha[1, 3] + \beta[3, 1] = [\alpha + 3\beta, 3\alpha + \beta]$, hence we must solve the induced system of equations with two equations and two unknowns.

$$\begin{cases} \alpha + 3\beta = 12 \\ 3\alpha + \beta = 12 \end{cases}$$

Like before, if we substitute $\beta = -3\alpha + 12$ from the second equation into the first equation, then

$$12 = \alpha + 3\beta = \alpha + 3(-3\alpha + 12) = -8\alpha + 36$$

implies that $-8\alpha = -24$ so that $\alpha = 3$, from which it follows that $\beta = 3$.

By altering the presentation of our vectors from rows to columns, the relationship between linear combinations of vectors and systems of linear equations becomes all the more evident: by expressing the vectors $\mathbf{v} = [1, 3]$ and $\mathbf{w} = [3, 1]$ of Example 1.1.6 as column vectors, the containment of a point (x, y) within the plane spanned by $\mathbf{v}$ and $\mathbf{w}$ can be determined by solving the vector equation

$$\begin{bmatrix} x \\ y \end{bmatrix} = \alpha\mathbf{v} + \beta\mathbf{w} = \alpha \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \beta \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha + 3\beta \\ 3\alpha + \beta \end{bmatrix}.$$

Comparing the rows of the vectors on the left- and right-hand sides of this equation with the real numbers $x = 12$ and $y = 12$ yields the system of equations from Example 1.1.6.

By analogy to lines and planes spanned by vectors in real n-space, given any nonzero vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ in real n-space, the collection of all possible linear combinations of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ forms a hyperplane called the **span** of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ and denoted by

$$\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\} = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_k\mathbf{v}_k \mid \alpha_1, \alpha_2, \dots, \alpha_k \text{ are real numbers}\}.$$

We will return to discuss the notion of span in greater detail in Section 1.6.

1.2 Vector Magnitude and the Dot Product

Our aim throughout this section is to systematically develop the theory of **Euclidean geometry** in real n-space that was suggested peripherally (and perhaps unsatisfactorily) in the previous section.

We begin with a notion of distance. Given any points $X = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ in real n-space, we define the **distance** between X and Y as the following real number.

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

Consequently, the distance from the origin $O = (0, 0, \dots, 0)$ to the point X is denoted as follows.

$$d(X, O) = \sqrt{x_1^2 + \cdots + x_n^2}$$

We note that this definition of distance is merely a generalization of the length of the hypotenuse of the right triangle formed by the x-axis, the y-axis, and a point in the Cartesian plane: indeed, if we could visualize the right triangle formed by the origin of $\mathbb{R}^n$, the point $(x_1, \dots, x_{n-1}, 0)$, and the point $X = (x_1, \dots, x_{n-1}, x_n)$ in $\mathbb{R}^n$, then the length of its hypotenuse is precisely $d(X, O)$.

Consider the vector $\mathbf{x} = [x_1, \dots, x_n]$ lying in standard position in real n-space. Geometrically, $\mathbf{x}$ can be viewed as the ray emitting from the origin to the point $X = (x_1, \dots, x_n)$, hence the **length** of the vector $\mathbf{x}$ is precisely the distance from the origin O to the point X, i.e., the length of $\mathbf{x}$ is

$$\|\mathbf{x}\| = d(X, O) = \sqrt{x_1^2 + \cdots + x_n^2}.$$

Often, we will rather refer to the quantity $\|\mathbf{x}\|$ as the **magnitude** or **norm** of the vector $\mathbf{x}$.

Example 1.2.1. Consider the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ from Example 1.1.1. Computing the magnitudes of each vector yields $\|\mathbf{u}+\mathbf{v}\| = \sqrt{2^2 + 3^2 + 2^2} = \sqrt{17}$ and $\|-\mathbf{w}\| = \sqrt{0^2 + 2^2 + 2^2} = 2\sqrt{2} = \|\mathbf{w}\|$ and $\|3\mathbf{x}\| = \sqrt{3^2 + 3^2 + (-3)^2} = 3\sqrt{3} = 3\|\mathbf{x}\|$; these last two examples indicate a general phenomenon.

Proposition 1.2.2. *Consider any positive integer n and any vector $\mathbf{x} = [x_1, \dots, x_n]$ in real n-space.*

1.) *We have that $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x}$ is the zero vector.*

2.) *We have that $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all real numbers α.*

Proof. (1.) By definition, we have that $\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2} = 0$ if and only if $x_1^2 + \cdots + x_n^2 = 0$. Clearly, if $\mathbf{x}$ is the zero vector, then $x_1 = \cdots = x_n = 0$ so that $x_1^2 + \cdots + x_n^2 = 0^2 + \cdots + 0^2 = 0$. Conversely, if $\mathbf{x}$ is a nonzero vector, then its ith coordinate x_i must be nonzero for some integer $1 \leq i \leq n$. Considering that the square of a nonzero real number if a positive real number, we have that $x_i^2 > 0$. Even more, the square of any real number is non-negative, hence we have that $\|\mathbf{x}\|^2 = x_1^2 + \cdots + x_n^2 \geq x_i^2 > 0$. We conclude that $\|\mathbf{x}\|$ must be nonzero if $\mathbf{x}$ is nonzero.

(2.) We define $\alpha\mathbf{x} = \alpha[x_1, \dots, x_n] = [\alpha x_1, \dots, \alpha x_n]$. Consequently, the definition of magnitude yields $\|\alpha\mathbf{x}\| = \sqrt{(\alpha x_1)^2 + \cdots + (\alpha x_n)^2} = \sqrt{\alpha^2(x_1^2 + \cdots + x_n)^2} = |\alpha|\sqrt{x_1^2 + \cdots + x_n^2} = |\alpha|\|\mathbf{x}\|$. □

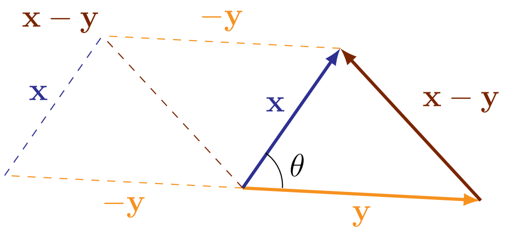
Conventionally, vectors of magnitude 1 are referred to as **unit vectors**. By Proposition 1.2.2, it can be shown that every nonzero vector $\mathbf{x}$ gives rise to a unique unit vector $\frac{1}{\|\mathbf{x}\|}\mathbf{x}$.

Corollary 1.2.3. *Every nonzero vector $\mathbf{x}$ of $\mathbb{R}^n$ induces a unit vector $\frac{1}{\|\mathbf{x}\|}\mathbf{x}$ of $\mathbb{R}^n$.*

Proof. By Proposition 1.2.2, if $\mathbf{x}$ is any nonzero vector of $\mathbb{R}^n$, then $\|\mathbf{x}\|$ is a positive real number. Consequently, we have that $\alpha = \frac{1}{\|\mathbf{x}\|}$ is a positive real number such that $\|\alpha\mathbf{x}\| = \alpha\|\mathbf{x}\| = 1$. □

Example 1.2.4. Consider the vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ from Example 1.2.1. We demonstrated previously that $\|\mathbf{x}+\mathbf{y}\| = \sqrt{17}$ and $\|\mathbf{z}\| = 2\sqrt{2}$, hence $\frac{1}{\sqrt{17}}(\mathbf{x}+\mathbf{y})$ and $\frac{1}{2\sqrt{2}}\mathbf{z}$ are unit vectors of $\mathbb{R}^3$.

Consider any pair of vectors $\mathbf{x}$ and $\mathbf{y}$ lying in standard position in real n-space for some positive integer n. Certainly, if $n = 2$ or $n = 3$, then we could visualize $\mathbf{x}$ and $\mathbf{y}$ in the Cartesian plane $\mathbb{R}^2$ or in the real 3-space $\mathbb{R}^3$ that we occupy, take a protractor, and measure the angle θ formed by the intersection of $\mathbf{x}$ and $\mathbf{y}$ at the origin. Pictorially, we would obtain the following diagram.



By the Law of Cosines, the triangle spanned by the vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x}-\mathbf{y}$ gives the following.

$$\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta)$$

Observe that if $\mathbf{x} = [x_1, \dots, x_n]$ and $\mathbf{y} = [y_1, \dots, y_n]$, then by definition of the magnitude of a vector, it follows that $\|\mathbf{x}\|^2 = x_1^2 + \cdots + x_n^2$ and $\|\mathbf{y}\|^2 = y_1^2 + \cdots + y_n^2$ so that

$$\|\mathbf{x}-\mathbf{y}\|^2 = (x_1-y_1)^2 + \cdots + (x_n-y_n)^2 = x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 - 2(x_1y_1 + \cdots + x_ny_n).$$

Combining this formula with the above equation obtained from the Law of Cosines yields that

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta) = \|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2(x_1y_1 + \cdots + x_ny_n)$$

so that $\|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta) = x_1y_1 + \cdots + x_ny_n$. We refer to the real number $x_1y_1 + \cdots + x_ny_n$ as the **dot product** $\mathbf{x}\cdot\mathbf{y}$ of the real vectors $\mathbf{x}$ and $\mathbf{y}$. Explicitly, if $\mathbf{x} = [x_1, \dots, x_n]$ and $\mathbf{y} = [y_1, \dots, y_n]$, then

$$\mathbf{x}\cdot\mathbf{y} = x_1y_1 + \cdots + x_ny_n.$$

Even more, it is clear from this exposition that the dot product informs the geometry of real n-space.

Proposition 1.2.5. *Given any pair of nonzero vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *lying in standard position in real* n*-space, the angle* θ *of intersection between the vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *at the origin satisfies that*

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right)$$

Essentially, the formula is obtained from the previous paragraph by solving for θ in the identity $\mathbf{x}\cdot\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta)$. Consequently, we will typically refer to the identity $\mathbf{x}\cdot\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta)$ as the **geometric representation** of the dot product. Extending this notion of geometry of vectors in real n-space, we will say that $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** (or **perpendicular**) provided that $\mathbf{x}\cdot\mathbf{y} = 0$. Observe that the angle of intersection between orthogonal vectors is $\cos^{-1}(0) = 90°$.

Example 1.2.6. Consider the vectors $\mathbf{x} = [1, 1, -1]$, $\mathbf{y} = [1, 2, 3]$, and $\mathbf{z} = [0, -2, -2]$ in $\mathbb{R}^3$. By definition of the dot product, we obtain the following identities.

$$\begin{aligned}
\mathbf{x} \cdot \mathbf{x} &= (1)(1) + (1)(1) + (-1)(-1) = 3 \\
\mathbf{x} \cdot \mathbf{y} &= (1)(1) + (1)(2) + (-1)(3) = 0 \\
\mathbf{x} \cdot \mathbf{z} &= (1)(0) + (1)(-2) + (-1)(-2) = 0 \\
\mathbf{y} \cdot \mathbf{z} &= (1)(0) + (2)(-2) + (3)(-2) = -10
\end{aligned}$$

Consequently, it follows that $\mathbf{x}$ is orthogonal to both $\mathbf{y}$ and $\mathbf{z}$, but $\mathbf{x}$ is not orthogonal to itself and $\mathbf{y}$ is not orthogonal to $\mathbf{z}$. Even more, we have that $\mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2$.

Example 1.2.7. Consider the vectors $\mathbf{x} = [1, 2, 0, 2]$ and $\mathbf{y} = [-3, 1, 1, 5]$ in $\mathbb{R}^4$. Even though we cannot visualize $\mathbf{x}$ and $\mathbf{y}$ as rays emitting from the origin because they exist in real 4-space, we can find their angle θ of intersection. By definition of vector magnitude, we have that

$$\begin{aligned}
\|\mathbf{x}\| &= \sqrt{1^2 + 2^2 + 0^2 + 2^2} = \sqrt{9} = 3 \text{ and} \\
\|\mathbf{y}\| &= \sqrt{(-3)^2 + 1^2 + 1^2 + 5^2} = \sqrt{36} = 6.
\end{aligned}$$

By definition of the dot product, it follows that $\mathbf{x} \cdot \mathbf{y} = (1)(-3) + (2)(1) + (0)(1) + (2)(5) = 9$. Consequently, we conclude by the geometric representation of the dot product that

$$\theta = \cos^{-1}\left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right) = \cos^{-1}\left(\frac{9}{(3)(6)}\right) = \cos^{-1}\left(\frac{1}{2}\right) = 60^\circ.$$

Considering that the dot product of vectors in real n-space is determined by the coordinates of the underlying vectors, the following proposition is unsurprising and straightforward to prove.

Proposition 1.2.8 (Properties of the Dot Product in Real n-Space)**.** *Consider any vectors* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{z}$ *lying in standard position in real n-space and any real number α. We have that*

(1.) *the dot product is commutative, i.e.,* $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$*;*

(2.) *the dot product is distributive across vector addition, i.e.,* $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$*;*

(3.) *the dot product is* **homogeneous***, i.e.,* $(\alpha\mathbf{x}) \cdot \mathbf{y} = \alpha(\mathbf{x} \cdot \mathbf{y}) = \mathbf{x} \cdot (\alpha\mathbf{y})$*; and*

(4.) *the dot product is* **non-degenerate***, i.e.,* $\mathbf{x} \cdot \mathbf{x}$ *is nonzero if and only if* $\mathbf{x}$ *is nonzero.*

Even more, the dot product in real n-space satisfies the **Law of Cosines**

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta).$$

Proof. Each of the above properties can be verified directly by listing the coordinates of the vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, computing the dot product, and appealing to familiar properties of real numbers.

Even more, the commutativity and distributivity of the dot product yield that

$$\|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2(\mathbf{x} \cdot \mathbf{y}) = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2(\mathbf{x} \cdot \mathbf{y}).$$

By the geometric representation of the dot product, we find that $\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$ for the angle θ of intersection between $\mathbf{x}$ and $\mathbf{y}$. Considering that the vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x} - \mathbf{y}$ induce a triangle of side lengths $\|\mathbf{x}\|$, $\|\mathbf{y}\|$, and $\|\mathbf{x} - \mathbf{y}\|$, respectively, such that the side of length $\|\mathbf{x} - \mathbf{y}\|$ lies opposite the angle θ of intersection between $\mathbf{x}$ and $\mathbf{y}$, the Law of Cosines holds in this case. □

By applying the Properties of the Dot Product in Real n-Space in the case of orthogonal vectors, we can prove the following important properties of orthogonal vectors.

Proposition 1.2.9 (Properties of Orthogonal Vectors in Real n-Space)**.** *Consider any vectors* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{z}$ *lying in standard position in real* n*-space.*

1.) *If* $\mathbf{x}$ *is orthogonal to* $\mathbf{y}$ *and* $\mathbf{z}$, *then* $\mathbf{x}$ *is orthogonal to* $\mathbf{y}+\mathbf{z}$.

2.) *If* $\mathbf{x}$ *is orthogonal to* $\mathbf{y}$, *then* $\mathbf{x}$ *is orthogonal to* $\alpha\mathbf{y}$ *for all real numbers* α.

3.) *If* $\mathbf{x}$ *is orthogonal to* $\mathbf{y}$, *then their angle of intersection is* $90°$, *i.e.,* $\mathbf{x}$ *and* $\mathbf{y}$ *are perpendicular.*

4.) *If* $\mathbf{x}$ *is orthogonal to* $\mathbf{y}$, *then* $\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2+\|\mathbf{y}\|^2$, *i.e., the* **Pythagorean Theorem** *holds.*

Proof. 1.) By definition, if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal and $\mathbf{x}$ and $\mathbf{z}$ are orthogonal, then $\mathbf{x}\cdot\mathbf{y}=0$ and $\mathbf{x}\cdot\mathbf{z}=0$. By Proposition 1.2.8, it follows that $\mathbf{x}\cdot(\mathbf{y}+\mathbf{z}) = \mathbf{x}\cdot\mathbf{y}+\mathbf{x}\cdot\mathbf{z} = 0$.

2.) By Proposition 1.2.8, we have that $\mathbf{x}\cdot(\alpha\mathbf{y}) = \alpha(\mathbf{x}\cdot\mathbf{y}) = 0$ for all real numbers α.

3.) By Proposition 1.2.5, if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors lying in standard position in real n-space, then the angle θ of intersection between the vectors $\mathbf{x}$ and $\mathbf{y}$ is given by $\theta = \cos^{-1}(0) = 90°$.

4.) By the Law of Cosines and its proof in Proposition 1.2.8, we have that

$$\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2+\|\mathbf{y}\|^2 - 2(\mathbf{x}\cdot\mathbf{y}).$$

Considering that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, we conclude that $2(\mathbf{x}\cdot\mathbf{y})=0$, as desired. □

Example 1.2.10. We determine in this example a unit vector perpendicular to $\mathbf{x} = [-1,3,4]$. By definition, we seek a real vector $\mathbf{y} = [y_1,y_2,y_3]$ such that $\mathbf{x}\cdot\mathbf{y}=0$ and $\|\mathbf{y}\|=1$. Computing the dot product of $\mathbf{x}$ and $\mathbf{y}$, we find that $0 = \mathbf{x}\cdot\mathbf{y} = -y_1+3y_2+4y_3$. We have three variables and only one equation, hence there must be two free variables that we are allowed to set equal to anything that is convenient. We will choose $y_1=0$ and $y_2=-4$; the resulting equation is $3(-4)+4y_3=0$ so that $4y_3 = 3(4)$ and $y_3=3$. Consequently, the vector $\mathbf{y}=[0,-4,3]$ is orthogonal to $\mathbf{x}$; however, its magnitude is $\sqrt{0^2+(-4)^2+3^2}=5$, so it is not a unit vector. By Proposition 1.2.3, we find that $\frac{1}{5}\mathbf{y}$ is a unit vector; it is orthogonal to $\mathbf{x}$ by Proposition 1.2.9 because $\mathbf{y}$ is orthogonal to $\mathbf{x}$.

Example 1.2.11. We determine in this example a unit vector perpendicular to $\mathbf{x} = [-1,3,4]$ and $\mathbf{y}=[2,1,-1]$. Like before in Example 1.2.10, we must solve the following system of equations.

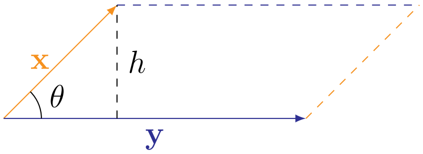
$$\begin{aligned} 0 &= \mathbf{x}\cdot[z_1,z_2,z_3] = -z_1+3z_2+4z_3 \\ 0 &= \mathbf{y}\cdot[z_1,z_2,z_3] = 2z_1+z_2-z_3 \end{aligned}$$

By adding twice the first equation to the second equation, we find that $7z_2+7z_3=0$ or $z_3=-z_2$. We have two equations in three unknowns, so we will have at least one free variable; however, as the arithmetic bears out, we find that z_3 depends on z_2, hence z_2 is a second free variable. By setting $z_1=0$ and $z_2=1$, we find that $z_3=-1$ and $\mathbf{z}=[0,1,-1]$ is orthogonal to $\mathbf{x}$ and $\mathbf{y}$. Considering that $\|\mathbf{z}\| = \sqrt{0^2+1^2+(-1)^2} = \sqrt{2}$, we conclude that $\frac{1}{\sqrt{2}}\mathbf{z}$ is a unit vector orthogonal to $\mathbf{x}$ and $\mathbf{y}$.

Geometrically, we have seen to our pleasant surprise that the dot product in real n-space enjoys many nice properties. But perhaps one of its most astounding features is the following.

Proposition 1.2.12. *Given any nonzero, non-parallel vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *lying in standard position in real* n*-space, the area of the* **parallelogram** *spanned by* $\mathbf{x}$ *and* $\mathbf{y}$ *is* $\|\mathbf{x}\|\|\mathbf{y}\|\sin(\theta)$.

Proof. Pictorially, the parallelogram spanned by $\mathbf{x}$ and $\mathbf{y}$ can be determined as follows.



Observe that the angle θ of intersection between $\mathbf{x}$ and $\mathbf{y}$ satisfies that $h = \|\mathbf{x}\| \sin(\theta)$. Because the area of a parallelogram is the product of its base and its height, it is $h\|\mathbf{y}\| = \|\mathbf{x}\|\|\mathbf{y}\| \sin(\theta)$. □

Before we conclude this section, we state and prove two inequalities regarding vectors in real n-space. Crucially, the following proposition provides a purely algebraic foundation for the geometry of the dot product in real n-space that we have as yet taken for granted (cf. Proposition 1.2.5).

Theorem 1.2.13 (Cauchy-Schwarz Inequality)**.** *Given any vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$*, we have that*

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|.$$

Consequently, the inverse cosine of $\mathbf{x} \cdot \mathbf{y}/\|\mathbf{x}\|\|\mathbf{y}\|$ *is well-defined, and Proposition 1.2.5 is valid.*

Proof. Clearly, if one of $\mathbf{x}$ or $\mathbf{y}$ is zero, then $\mathbf{x} \cdot \mathbf{y} = 0$ and $\|\mathbf{x}\|\|\mathbf{y}\| = 0$, hence the inequality holds. Consequently, we may assume that neither $\mathbf{x}$ nor $\mathbf{y}$ is zero so that $\mathbf{y} \cdot \mathbf{y}$ is nonzero by the Properties of the Dot Product in Real n-Space. Even more, for any real numbers α and β, we have that

$$\|\alpha\mathbf{x} + \beta\mathbf{y}\|^2 = (\alpha\mathbf{x} + \beta\mathbf{y}) \cdot (\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha^2(\mathbf{x} \cdot \mathbf{x}) + 2\alpha\beta(\mathbf{x} \cdot \mathbf{y}) + \beta^2(\mathbf{y} \cdot \mathbf{y})$$

is non-negative. By the above identity with $\alpha = \mathbf{y} \cdot \mathbf{y}$ and $\beta = -(\mathbf{x} \cdot \mathbf{y})$, we find that

$$(\mathbf{x} \cdot \mathbf{x})(\mathbf{y} \cdot \mathbf{y})^2 - (\mathbf{x} \cdot \mathbf{y})^2(\mathbf{y} \cdot \mathbf{y}) \geq 0.$$

Considering that $\mathbf{y} \cdot \mathbf{y}$ is nonzero by assumption, it must be a positive real number. Cancelling one factor of $\mathbf{y} \cdot \mathbf{y}$ from both sides of the above inequality yields that $(\mathbf{x} \cdot \mathbf{y})^2 \leq (\mathbf{x} \cdot \mathbf{x})(\mathbf{y} \cdot \mathbf{y})$. □

Theorem 1.2.14 (Triangle Inequality)**.** *Given any vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *in* $\mathbb{R}^n$*, we have that*

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Proof. By Proposition 1.2.2, we have that $\|\mathbf{x}+\mathbf{y}\|$, $\|\mathbf{x}\|$, and $\|\mathbf{y}\|$ are each non-negative real numbers, hence the desired inequality holds if and only if the inequality $\|\mathbf{x} + \mathbf{y}\|^2 \leq (\|\mathbf{x}\| + \|\mathbf{y}\|)^2$ holds. By Proposition 1.2.8, the left-hand side of this inequality is $(\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = \|\mathbf{x}\|^2 + 2(\mathbf{x} \cdot \mathbf{y}) + \|\mathbf{y}\|^2$. By the Cauchy-Schwarz Inequality, it follows that $2(\mathbf{x} \cdot \mathbf{y}) \leq 2\|\mathbf{x}\|\|\mathbf{y}\|$, hence the inequality holds.

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2(\mathbf{x} \cdot \mathbf{y}) + |\mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2$$ □

1.3 Matrices and Matrix Operations

We will continue to assume throughout this chapter that m and n are positive integers. We refer to a visual representation of any collection of data arranged into m rows and n columns as an $m \times n$ **array**. Each entry of an $m \times n$ array A is a **component** of A. Each component of A can be uniquely identified by specifying its row and column: explicitly, we use the symbol a_{ij} to indicate the component of A that lies in the ith row and jth column. Often, we will refer to a_{ij} as the (i, j)th **entry** of the array A. Collectively, therefore, we may view the array A as **indexed** by its entries a_{ij} for each pair of integers $1 \leq i \leq m$ and $1 \leq j \leq n$. Components of the form a_{ii} are referred to as the **diagonal** entries of A because they lie in the same row and column of A; the collection of all diagonal entries of A is called the **main diagonal** of A. We will adopt the convention that an $m \times n$ array be written using large rectangular brackets, as in each of the following examples.

Example 1.3.1. Consider the case that Alice, Bob, Carly, and Daryl play Bridge together. If Alice and Carly belong to one team and Bob and Daryl belong to the opposing team, then we may encode this information (i.e., these teams) as the two columns of the following 2×2 array A.

$$A = \begin{bmatrix} \text{Alice} & \text{Bob} \\ \text{Carly} & \text{Daryl} \end{bmatrix}$$

Observe that $a_{11} = \text{Alice}$, $a_{12} = \text{Bob}$, $a_{21} = \text{Carly}$, and $a_{22} = \text{Daryl}$. One could just as well swap the rows and columns to display the teams as rows by constructing the following 2×2 array A^T.

$$A^T = \begin{bmatrix} \text{Alice} & \text{Carly} \\ \text{Bob} & \text{Daryl} \end{bmatrix}$$

Our principal concern throughout this course are those $m \times n$ arrays consisting entirely of (real) numbers. Under this restriction, we may refer to an $m \times n$ array as a (real) $m \times n$ **matrix**. Generally, one can define matrices consisting of elements lying in any **ring**, but we will not be so general.

Example 1.3.2. Each real number x may be viewed as a real 1×1 matrix $\begin{bmatrix} x \end{bmatrix}$.

Example 1.3.3. Consider once again the scenario of Example 1.3.1. We may assign to each player a real number called a "skill value" between 0 and 100, e.g., suppose that Alice has skill value 88; Bob has skill value 72; Carly has skill value 95; and Daryl has skill value 90. Under this convention, the matrices of Example 1.3.1 yield new matrices that we could call "skill matrices" as follows.

$$S = \begin{bmatrix} 88 & 72 \\ 95 & 90 \end{bmatrix} \text{ and } S^T = \begin{bmatrix} 88 & 95 \\ 72 & 90 \end{bmatrix}$$

Our previous three examples dealt with **square** matrices, i.e., matrices for which the number of rows and the number of columns were the same (i.e., $m = n$); however, not all matrices are square.

Example 1.3.4. Consider the 1×5 matrix $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}$ of the first five positive integers.

We refer to matrices with only one row as **row vectors**; matrices with only one column are called **column vectors**. We are familiar with some notion of vectors from our study of real n-space in Section 1.1. We may also use the terms (horizontal) n**-tuples** for row vectors with n columns (i.e., $1 \times n$ matrices) and (vertical) m-tuples for column vectors with m rows (i.e., $m \times 1$ matrices).

Like we mentioned in the first paragraph of this section, an $m \times n$ matrix A is uniquely determined by the entry a_{ij} in its ith row and jth column for each pair of integers $1 \leq i \leq m$ and $1 \leq j \leq n$. For instance, the matrix of Example 1.3.4 is the unique matrix with one row whose jth column consists of the integer j for each integer $1 \leq j \leq 5$. Under this identification, we will adopt the one-line notation $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ for the $m \times n$ matrix A with a_{ij} in its ith row and jth column.

Example 1.3.5. Consider the 2×3 matrix whose ith row and jth column consists of the sum $i + j$. We may write this symbolically (in one-line notation) as $\big[i + j\big]_{\substack{1 \leq i \leq 2 \\ 1 \leq j \leq 3}}$ or expanded as follows.

$$\begin{array}{c c} & \begin{array}{ccc} j=1 & j=2 & j=3 \end{array} \\ \begin{array}{c} i=1 \\ i=2 \end{array} & \begin{bmatrix} 1+1 & 1+2 & 1+3 \\ 2+1 & 2+2 & 2+3 \end{bmatrix} \end{array} \text{ or } \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$$

Example 1.3.6. Given any positive integers m and n, there is one and only one matrix consisting entirely of zeros: it is the $m \times n$ **zero matrix** $O_{m \times n}$. Explicitly, we have the following examples.

$$O_{2\times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad O_{2\times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad O_{3\times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad O_{3\times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Often, it is most convenient to simply write O for the zero matrix with the understanding that the number of rows and columns of O is contingent upon the context in which it is discussed.

Example 1.3.7. We refer to the matrix $I_{m \times n} = \big[\delta_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ as the $m \times n$ **identity matrix**, where

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

is the **Kronecker delta**. Put another way, the $m \times n$ identity matrix is the unique $m \times n$ matrix whose (i, j)th component is one for each pair of integers $1 \leq i \leq m$ and $1 \leq j \leq n$ such that $i = j$ and whose other components are all zero. One can also say that $I_{m \times n}$ is the unique $m \times n$ matrix with ones along the main diagonal and zeros elsewhere. Explicitly, we have the following examples.

$$I_{2\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad I_{2\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad I_{3\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad I_{3\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Observe that the only nonzero components of $I_{n \times n}$ lie on the main diagonal, hence $I_{n \times n}$ is a **diagonal matrix**. Explicitly, a diagonal matrix is an $n \times n$ matrix consisting entirely of zeros off the main diagonal. Even more, $I_{n \times n}$ is the unique diagonal $n \times n$ matrix whose nonzero entries are all one. Like with the zero matrix, we will write I for the square identity matrix of the appropriate size.

Example 1.3.8. Given any $m \times n$ matrix $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$, its **matrix transpose** A^T is the $n \times m$ matrix obtained by swapping the rows and columns of A, i.e., we have that $A^T = \big[a_{ji}\big]_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$. Put another way, the (i,j)th entry of A^T is the (j,i)th entry of A, hence the ith row of A^T is precisely the ith column of A. Explicitly, for the matrix A defined in Example 1.3.5, we have the following.

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix} \qquad A^T = \begin{bmatrix} 2 & 3 \\ 3 & 4 \\ 4 & 5 \end{bmatrix}$$

Observe that the first row of A becomes the first column of A^T (and likewise for the second row). Consequently, the transpose of any $1 \times n$ row vector is an $n \times 1$ column vector. We will also refer to A^T simply as the transpose of A; the process of computing A^T is called **transposition**. One other thing to notice is that $I_{m \times n}^T = I_{n \times m}$, hence we have that $I_{n \times n}^T = I_{n \times n}$ or $I^T = I$.

Definition 1.3.9. We say that an $m \times n$ matrix A is **symmetric** if it holds that $A^T = A$. Observe that a matrix is symmetric only if it is square, i.e., a non-square matrix is never symmetric.

Considering that matrices encode numerical data, it is not surprising to find that they induce their own arithmetic. Using one-line notation, matrix addition can be defined as follows.

Definition 1.3.10. Given any $m \times n$ matrices $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ and $B = \big[b_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$, the **matrix sum** of A and B is the $m \times n$ matrix $A + B = \big[a_{ij} + b_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$. Put in words, the matrix sum $A + B$ is the $m \times n$ matrix whose (i,j)th entry is the sum of the (i,j)th entries of A and B.

Caution: the matrix sum is not defined for matrices with different numbers of rows or columns.

Example 1.3.11. We compute the matrix sum of the following 2×3 matrices.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 + -1 & 2 + 0 & 3 + 1 \\ 4 + -1 & 5 + 0 & 6 + 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 4 \\ 3 & 5 & 7 \end{bmatrix}$$

Example 1.3.12. If A is any $m \times n$ matrix, then we have that $A + O_{m \times n} = A = O_{m \times n} + A$. Consequently, we may view $O_{m \times n}$ as the **additive identity** among all $m \times n$ matrices.

Generally, for any real $m \times n$ matrix $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$, we will typically refer to any (real) number c as a **scalar**, and we define the **scalar multiple** of A by the scalar c as $cA = \big[ca_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$. Essentially, we may view this as a generalization of the sum of the matrix A with itself c times.

Example 1.3.13. Given any $m \times n$ matrix $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$, we will write $-A = \big[-a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$. We have that $A + (-A) = O_{m \times n} = -A + A$, and we say that $-A$ is the **additive inverse** of A.

Our next proposition illustrates that matrix transposition and matrix addition are compatible.

Proposition 1.3.14. *Let A and B be any $m \times n$ matrices. We have that $(A + B)^T = A^T + B^T$. Put another way, the transpose of a sum of matrices is the sum of the matrix transposes.*

Proof. By Definition 1.3.10, the (i,j)th entry of $A + B$ is the sum of the (i,j)th entry of A and the (i,j)th entry of B. By Example 1.3.8, the (i,j)th entry of $(A + B)^T$ is the (j,i)th entry of $A + B$, i.e., the sum of the (j,i)th entry of A and the (j,i)th entry of B. But by the same example, this is the sum of the (i,j)th entry of A^T and the (i,j)th entry of B^T. Ultimately, this shows that the (i,j)th entry of $(A + B)^T$ and the (i,j)th entry of $A^T + B^T$ are the same so that $(A + B)^T = A^T + B^T$. □

Even more, if the number of columns (or rows) of a matrix A equals the number of rows (or columns) of a matrix B, then the product of the matrices A and B is defined as follows.

Definition 1.3.15. Given any $m \times n$ matrix $A = \big[a_{ij}\big]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ and any $n \times r$ matrix $B = \big[a_{ij}\big]_{\substack{1 \leq i \leq n, \\ 1 \leq j \leq r}}$ the (left) **matrix product** of A and B is the $m \times r$ matrix AB whose (i, j)th entry is given by

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}.$$

Put in words, the matrix product AB is the $m \times r$ matrix whose (i, j)th entry is the sum of the product of the (i, k)th entry of A and the (k, j)th entry of B for all integers $1 \leq k \leq n$.

Crucially, matrix multiplication is not commutative, i.e., the order of the matrices in the matrix product matters; however, if we assume that $r = m$, then the (right) matrix product BA can be defined analogously. Be sure to note also that the number of rows of AB is the same as the number of rows of A, and the number of columns of AB is the same as the number of columns of B.

Caution: the product is not defined for matrices with an incompatible number of rows and columns.

Example 1.3.16. Consider the following real matrices.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 0 \\ 0 & 1 \\ -1 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 2 \\ -1 & 1 & 3 \end{bmatrix}$$

Considering that A is a 2×3 matrix and B is a 3×2 matrix, both of the products AB and BA can be formed: AB is a 2×2 matrix, and BA is a 3×3 matrix. Explicitly, they are as follows.

$$AB = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1(-1) + 2(0) + 3(-1) & 1(0) + 2(1) + 3(1) \\ 2(-1) + 3(0) + 4(-1) & 2(0) + 3(1) + 4(1) \end{bmatrix} = \begin{bmatrix} -4 & 5 \\ -6 & 7 \end{bmatrix}$$

$$BA = \begin{bmatrix} -1 & 0 \\ 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} -1(1) + 0(2) & -1(2) + 0(3) & -1(3) + 0(4) \\ 0(1) + 1(2) & 0(2) + 1(3) & 0(3) + 1(4) \\ -1(1) + 1(2) & -1(2) + 1(3) & -1(3) + 1(4) \end{bmatrix} = \begin{bmatrix} -1 & -2 & -3 \\ 2 & 3 & 4 \\ 1 & 1 & 1 \end{bmatrix}$$

On the other hand, neither of the matrix products AC or BD exist; however, the matrices CA and DB can be computed because A and B have as many rows as C and D have columns, respectively.

Example 1.3.17. Consider the following real matrices.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

Considering that A and B are both 2×2 matrices, the 2×2 matrices AB and BA can be formed.

$$AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1(-1) + 2(0) & 1(0) + 2(1) \\ 3(-1) + 4(0) & 3(0) + 4(1) \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ -3 & 4 \end{bmatrix}$$

$$BA = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} -1(1) + 0(3) & -1(2) + 0(4) \\ 0(1) + 1(3) & 0(2) + 1(4) \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 3 & 4 \end{bmatrix}$$

Crucially, we note that AB and BA are not equal as matrices, i.e., we have that $AB \neq BA$.

Remark 1.3.18. Example 1.3.16 motivates the following definition of matrix multiplication. Consider a $1 \times n$ row vector $\mathbf{v} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \end{bmatrix}$ and the following $n \times 1$ column vector.

$$\mathbf{w} = \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{n1} \end{bmatrix}$$

We define the **vector dot product** $\mathbf{v} \cdot \mathbf{w}$ of the vectors $\mathbf{v}$ and $\mathbf{w}$ as the 1×1 matrix $\mathbf{v}\mathbf{w}^T$, i.e.,

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}\mathbf{w}^T = \begin{bmatrix} v_{11}w_{11} + v_{12}w_{21} + \cdots + v_{1n}w_{n1} \end{bmatrix}.$$

Given any $m \times n$ matrix A and any $n \times r$ matrix B, the ith row of A may be viewed as the $1 \times n$ vector $A_i = \begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{bmatrix}$ and the jth column of B as the following $n \times 1$ vector.

$$B_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix}$$

Ultimately, under this interpretation, the matrix product AB is defined as the $m \times r$ matrix whose (i, j)th component is the dot product $A_i \cdot B_j = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^n a_{ik}b_{kj}$.

We adapt the following example from the example at the bottom of page 50 of [Lan86].

Example 1.3.19. We say that an $n \times n$ matrix A is a **Markov matrix** if each component of A is a non-negative real number and the sum of each column of A is 1. For instance, the 2×2 matrix

$$A = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix}$$

is a Markov matrix. We may view this Markov matrix as representing a real-life scenario as follows.

Godspeed You! Black Emperor are performing live at the Blue Note in Columbia, Missouri, and Alice and Bob are considering attending the concert. Currently, Alice is 90% certain that she will attend, so she must be 10% certain that she will not attend. On the other hand, Bob is 50% sure he will attend. Consequently, the columns of the matrix A represent Alice and Bob, respectively, and the rows represent their certainty or uncertainty that they will attend the show, respectively.

Even more, suppose that today, Alice has the propensity a to attend the concert and Bob has the propensity b to attend, and tomorrow, Alice has the propensity $0.9a + 0.5b$ to attend the concert and Bob has the propensity $0.1a + 0.5b$ to attend. Under these identifications, tomorrow, the propensity that Alice and Bob will attend the concert is given by the following matrix product.

$$\begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.9a + 0.5b \\ 0.1a + 0.5b \end{bmatrix} = a\begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} + b\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

We could continue to iterate this process to predict the propensity that Alice and Bob will attend the concert on any given day in the future; the resulting model is referred to as a **Markov process**.

Remark 1.3.20. Example 1.3.19 illustrates that if $\mathbf{x}$ is an $n \times 1$ column vector and A is an $m \times n$ matrix, then the $m \times 1$ column vector $A\mathbf{x}$ is simply a linear combination of the columns of A.

$$A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}$$

We will demonstrate now that matrix multiplication is associative and distributive.

Proposition 1.3.21. *If A is any $m \times n$ matrix, B is any $n \times r$ matrix, and C is any $r \times s$ matrix, then the matrix products $A(BC)$ and $(AB)C$ are well-defined; in fact, they are equal.*

Proof. By Definition 1.3.15, we have that BC is an $n \times s$ matrix, hence the matrix product $A(BC)$ is well-defined because the number of columns of A is equal to the number of rows of BC; a similar argument shows that $(AB)C$ is well-defined, hence it suffices to prove that $A(BC) = (AB)C$. By the same definition, the (i, j)th entry of $A(BC)$ is the sum of the products of the (i, k)th entry of A and the (k, j)th entry of BC for all integers $1 \leq k \leq n$, and the (k, j)th entry of BC is the sum of the products of the (k, ℓ)th entry of B and the (ℓ, j)th entry of C for all integers $1 \leq \ell \leq r$. Put into symbols, the previous sentence can be expressed as the double summation identity

$$A(BC)_{ij} = \sum_{k=1}^{n} \sum_{\ell=1}^{r} a_{ik} b_{k\ell} c_{\ell j}.$$

Considering that the order of summation of a finite sum does not matter, it follows that

$$A(BC)_{ij} = \sum_{\ell=1}^{r} \sum_{k=1}^{n} a_{ik} b_{k\ell} c_{\ell j}.$$

Observe that $\sum_{k=1}^{n} a_{ik} b_{k\ell}$ is nothing more than the (i, ℓ)th entry of AB, hence we may view the (i, j)th entry of $A(BC)$ as the sum of the products of the (i, ℓ)th entry of AB and the (ℓ, j)th entry of C for all integers $1 \leq i \leq r$, i.e., it is the (i, j)th entry of $(AB)C$. Ultimately, this shows that the (i, j)th entry of $A(BC)$ and the (i, j)th entry of $(AB)C$ are the same so that $A(BC) = (AB)C$. □

Proposition 1.3.22. *If A is any $m \times n$ matrix and B and C are any $n \times r$ matrices, then the product $A(B + C)$ is well-defined; $A(B + C) = AB + AC$; and $A(cB) = c(AB)$ for all scalars c.*

Proof. By Definition 1.3.10, the matrix sum $B + C$ is an $n \times r$ matrix, hence the product $A(B + C)$ is well-defined because the number of columns of A is equal to the number of rows of $B + C$. By Definition 1.3.15, the (i, j)th entry of $A(B + C)$ is the sum of the products of the (i, k)th entry of A and the $(k, j)$the entry of BC for all integers $1 \leq k \leq n$; the latter is by Definition 1.3.10 the sum of the (k, j)th entry of B and the (k, j)th entry of C. Because multiplication is distributive over addition, the (i, j)th entry of $A(B + C)$ is the sum of the products of the (i, k)th entry of A and the (k, j)th entry of B for all integers $1 \leq k \leq n$ plus the sum of the products of the (i, k)th entry of A and the (k, j)th entry of C for all integers $1 \leq k \leq n$, i.e., it is the sum of the (i, j)th entry of AB and the (i, j)th entry of AC, i.e., it is the (i, j)th entry of $AB + AC$. Because the (i, j)th entry of $A(B + C)$ and the (i, j)th entry of $AB + AC$ are the same, we conclude that $A(B + C) = AB + AC$.

We leave it as an exercise for the reader to demonstrate that $A(cB) = c(AB)$ for all scalars c; however, we remark that inspiration can be found in the proof of Proposition 1.3.21. □

Ultimately, Proposition 1.3.22 implies that matrix multiplication is distributive, i.e., if A is any $m \times n$ matrix, B and C are any $n \times r$ matrices, and c is any scalar, then $A(cB + C) = c(AB) + AC$.

Example 1.3.23. Given any $n \times n$ matrix A, the matrix product of A with itself is denoted simply by A^2; it is an $n \times n$ matrix, hence we may form the matrix product of A^2 with A. By Proposition 1.3.21, we have that $(A^2)A = (AA)A = A(AA) = A(A^2)$; we denote this simply by A^3. Continuing in this manner, the k-fold product of A is $A^k = A^{k-1}A = AA^{k-1}$ for all integers $k \geq 2$. Each of these is an $n \times n$ matrix, so we can scale these matrices and add them together to obtain a **matrix polynomial**. By the distributive property for matrices, matrix polynomials behave familiarly, e.g.,

$$(A - I)(A + I) = A^2 + AI - IA - I^2 = A^2 + A - A - I = A^2 - I \text{ and}$$
$$(A + I)^3 = (A^2 + 2A + I)(A + I) = A^3 + A^2 + 2A^2 + 2A + A + I = A^3 + 3A^2 + 3A + I.$$

Even more, like matrix addition, matrix multiplication is compatible with transposition.

Proposition 1.3.24. *If A is any $m \times n$ matrix and B is any $n \times r$ matrix, then $(AB)^T = B^T A^T$. Put another way, the transpose of a matrix product is the reverse matrix product of the transposes.*

Proof. By Example 1.3.8, the (i, j)th entry of $(AB)^T$ is the (j, i)th AB. By Definition 1.3.15, the (j, i)th entry of AB is the sum of the products of the (j, k)th entry of A and the (k, i)th entry of B for all integers $1 \leq k \leq n$. Considering that scalar multiplication is commutative, this is equal to the sum of the products of the (i, k)th entry of B^T and the (k, j)th entry of A^T for all integers $1 \leq k \leq n$, i.e., it is the (i, j)th entry of $B^T A^T$. We conclude therefore that $(AB)^T = B^T A^T$. □

We conclude with a summary of the matrix operations proved in the previous propositions.

Proposition 1.3.25 (Properties of Matrix Addition, Multiplication, and Transposition)**.** *Consider any matrices A, B, and C such that the following matrix sums and matrix products are well-defined.*

1.) *Matrix addition is associative, i.e., $(A + B) + = A + (B + C)$.*

2.) *Matrix addition is commutative, i.e., $A + B = B + A$.*

3.) *The zero matrix O is the additive identity, i.e., $A + O = A$.*

4.) *The additive inverse of A is $-A$, i.e., $A + (-A) = O$.*

5.) *Matrix multiplication is associative, i.e., $(AB)C = A(BC)$.*

6.) *Matrix multiplication is distributive, i.e., $A(B + C) = AB + AC$ and $(A + B)C = AC + BC$.*

7.) *The multiplicative identity is the identity matrix, i.e., $IA = A$ and $BI = B$.*

8.) *Matrix transposition is distributive across matrix addition, i.e., $(A + B)^T = A^T + B^T$.*

9.) *Matrix transposition is order-reversing, i.e., $(AB)^T = B^T A^T$.*

10.) *Scalar multiplication is associative, i.e., $r(sA) = (rs)A$.*

11.) *Scalar multiplication is distributive across matrix addition, i.e., $r(A + B) = rA + rB$.*

12.) *Scalar multiplication is distributive across scalar addition, i.e., $(r + s)A = rA + sA$.*

13.) *Scalar multiplication is homogeneous, i.e., $(rA)B = r(AB) = A(rB)$.*

1.4 Linear Systems of Equations and Gaussian Elimination

We will continue to assume that m and n are positive integers. If $x_1, \dots, x_n$ are any variables, then a (real) **linear combination** of $x_1, \dots, x_n$ is an expression of the form $a_1x_1 + \cdots + a_nx_n$ for some (real) scalars $a_1, \dots, a_n$. Consequently, a (real) $1 \times n$ **linear equation** is any equation of the form $a_1x_1 + \cdots + a_nx_n = b$ for some (real) scalars $a_1, \dots, a_n$, and b. Even more, a (real) $m \times n$ **system of linear equations** consists of m linear equations in n variables; this is represented as follows.

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

Explicitly, the positive integer m represents the number of equations in the $m \times n$ system of linear equations, and the positive integer n represents the number of variables in each equation.

Example 1.4.1. On 10 June 2022, in Game Four of the 2022 NBA Finals, Stephen Curry scored 43 points. Let x_1 be the number of one-pointers made; let x_2 be the number of two-pointers made; and let x_3 be the number of three-pointers made by Curry in this appearance. Observe that Curry's point total is given by the 1×3 (integer) linear equation $x_1 + 2x_2 + 3x_3 = 43$.

We say that the (real) scalars $\xi_1, \dots, \xi_n$ constitute a **solution** to a (real) $m \times n$ system of linear equations if it holds that $a_{i1}\xi_1 + \cdots + a_{in}\xi_n = b_i$ for each integer $1 \leq i \leq m$.

Example 1.4.2. One can find many solutions to the matrix equation of Example 1.4.1. Explicitly, $\xi_1 = 43$ and $\xi_2 = \xi_3 = 0$ or $\xi_1 = 41$, $\xi_2 = 1$, and $\xi_3 = 0$ give rise to two distinct solutions.

Given more information, we can reduce the number of possible solutions in Example 1.4.1. Using the fact that Curry made seven three-pointers, we may substitute $x_3 = 7$ into our equation $x_1 + 2x_2 + 3x_3 = 43$ to find that $x_1 + 2x_2 + 21 = 43$ or $x_1 + 2x_2 = 22$. Even more, Curry made a combined fifteen free throws and two-pointers. Consequently, we have that $x_1 + x_2 = 15$. Observe that these two equations involving x_1 and x_2 induce the following 2×2 system of linear equations.

$$\begin{aligned} x_1 + 2x_2 &= 22 \\ x_1 + x_2 &= 15 \end{aligned}$$

We may determine the values of x_1 and x_2 that solve the system: we have that $x_1 = 15 - x_2$ so that $22 = x_1 + 2x_2 = (15 - x_2) + 2x_2 = 15 + x_2$; cancelling 15 from both sides gives $x_2 = 7$ and $x_1 = 8$.

Examples 1.4.1 and 1.4.2 highlight the differences between the **general solution** of a system of linear equations as opposed to a **particular solution**. Explicitly, the 1×3 system of equations

$$x_1 + 2x_2 + 3x_3 = 43$$

admits infinitely many solutions: by solving this equation for x_1 in terms of x_2 and x_3, we find that $x_1 = -2x_2 - 3x_3 + 43$, hence the general solution to this system of equations is given by

$$\boldsymbol{\xi} = [-2x_2 - 3x_3 + 43, x_2, x_3] = x_2[-2, 1, 0] + x_3[-3, 0, 1] + [43, 0, 0].$$

Consequently, any choice of real numbers x_2 and x_3 determine a particular solution to this system of linear equations. We will soon revisit this distinction with more sophisticated tools.

Example 1.4.3. Geometrically, linear equations encode lines, planes, and hyperplanes. Explicitly, for any real numbers a and b (not both of which are zero) and any real number c, the solutions of the linear equation $ax + by = c$ form a line (e.g., $2x + y = 3$ is a line with y-intercept $(0, 3)$ and slope -2). Likewise, it is not difficult to verify that for any real numbers a, b, and c (not all of which are zero) and any real number d, the solutions of the linear equation $ax + by + cz = d$ form a plane: indeed, if a is nonzero, then solving for x in this linear equation yields that

$$x = -\frac{b}{a}y - \frac{c}{a}z + d,$$

hence (x, y, z) is a translation of a point lying in the plane spanned by $[0, 1, 0]$ and $[0, 0, 1]$.

Using matrices, we can more efficiently rephrase our above observations concerning $m\times n$ systems of linear equations. Explicitly, observe that a (real) $m \times n$ system of linear equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

gives rise to a $n\times 1$ matrix $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$, an $m\times 1$ matrix $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix}^T$, and an $m \times n$ matrix A whose (i, j)th entry is the coefficient a_{ij} of the jth variable x_j of the ith equation $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$ of the $m \times n$ system of linear equations, i.e., the following $m \times n$ matrix.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

Conversely, the aforementioned matrices A, $\mathbf{x}$, and $\mathbf{b}$ satisfy that $A\mathbf{x} = \mathbf{b}$. We refer to the equation $A\mathbf{x} = \mathbf{b}$ as a (real) $m \times n$ **matrix equation**. Often, the $m \times n$ matrix A and the $m \times 1$ matrix $\mathbf{b}$ are known while the $n \times 1$ matrix $\mathbf{x}$ consists of n variables. Ultimately, we obtain a one-to-one correspondence between (real) $m \times n$ systems of linear equations and $m \times n$ matrix equations.

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \iff A\mathbf{x} = \mathbf{b}, \text{ i.e., } \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Example 1.4.4. We will convert the data of Examples 1.4.1 and 1.4.2 into the language of matrix equations. Consider the matrix $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ whose jth column is the point value of a j-pointer; the matrix $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$ whose jth row is the number of j-pointers made by Curry; and the matrix $\mathbf{b} = \begin{bmatrix} 43 \end{bmatrix}$ consisting of the total points made by Curry. Observe that the linear equation $x_1 + 2x_2 + 3x_3 = 43$ is in one-to-one correspondence with the matrix equation $A\mathbf{x} = \mathbf{b}$.

We say that a (real) $n \times 1$ matrix $\boldsymbol{\xi}$ forms a **solution** to the matrix equation $A\mathbf{x} = \mathbf{b}$ if it holds that $A\boldsymbol{\xi} = \mathbf{b}$; this is a direct analog of a solution of the $m \times n$ system of linear equations.

Example 1.4.5. Rephrasing the results of 1.4.2, the matrices $\boldsymbol{\xi}_1 = \begin{bmatrix} 43 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{\xi}_2 = \begin{bmatrix} 41 & 1 & 0 \end{bmatrix}$ give rise to two distinct solutions of the matrix equation of Example 1.4.4. On the other hand, put into the language of matrix equations, the information that $22 = x_1 + 2x_2$ and $15 = x_1 + x_2$ can be most efficiently synthesized by viewing the coefficients of these linear equations as rows of a matrix. Explicitly, we construct a matrix A whose first row is $\begin{bmatrix} 1 & 2 \end{bmatrix}$, corresponding to the respective coefficients of x_1 and x_2 in the equation $22 = x_1 + 2x_2$; the second row of the matrix A is $\begin{bmatrix} 1 & 1 \end{bmatrix}$, corresponding to the respective coefficients of x_1 and x_2 in the equation $15 = x_1 + x_2$. Once again, the column vector $\mathbf{x}$ consists of the variables x_1 and x_2 in distinct rows, and the column vector $\mathbf{b}$ consists of the integers 22 and 15 in distinct rows. Ultimately, yields the matrix equation

$$A\mathbf{x} = \mathbf{b} \text{ or } \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 22 \\ 15 \end{bmatrix}.$$

Once we have extracted an $m \times n$ matrix equation $A\mathbf{x} = \mathbf{b}$ from a (real) $m \times n$ system of linear equations, our immediate objective is to determine the matrix analog of solving the system. Before we do this, we declare the following three valid operations for systems of linear equations.

Definition 1.4.6 (Elementary Row Operations)**.** *Given any (real) $m \times n$ system of linear equations, the following arithmetic operations are permissible to perform on the system.*

1.) *We may multiply the ith equation by a nonzero (real) scalar c.*

2.) *We may add c times the ith equation to the jth equation for all integers $1 \leq i, j \leq m$.*

3.) *We may interchange the ith and jth equations for all integers $1 \leq i, j \leq m$.*

Consequently, we are looking for matrix analogs of the above three arithmetic operations. Considering that the coefficients of ith equation are encoded in the ith row of the matrix A and the ith row of the matrix $\mathbf{b}$, we may rather consider the **augmented matrix** $\begin{bmatrix} A \mid \mathbf{b} \end{bmatrix}$. By definition, this is simply the matrix A with one additional column in the form of $\mathbf{b}$. We use the bar $\mid$ notation to emphasize that $\mathbf{b}$ is appended as the rightmost column of the matrix A and not originally a column of A. By definition of matrix multiplication, operation (1.) is analogous to left multiplication by the $m \times m$ matrix with (i, i)th entry c; 1 in all other entries of the main diagonal; and 0s elsewhere.

1.) Multiplication of the ith row of an $m \times n$ system of linear equations by a scalar c corresponds to left multiplication of the $m \times (n + 1)$ augmented matrix $\begin{bmatrix} A \mid \mathbf{b} \end{bmatrix}$ by the $m \times m$ matrix with c in row i, column i; 1 in all other entries of the main diagonal; and 0s elsewhere.

Example 1.4.7. We obtain the following augmented matrix for the matrices of Example 1.4.5.

$$\begin{bmatrix} A \mid \mathbf{b} \end{bmatrix} = \left[\begin{array}{cc|c} 1 & 2 & 22 \\ 1 & 1 & 15 \end{array}\right]$$

Consequently, to scale the first equation $x_1 + 2x_2 = 22$ by a factor of c, we multiply this augmented matrix by the 2×2 matrix with c in row 1, column 1; 1 in row 2, column 2; and 0s elsewhere.

$$\left[\begin{array}{cc|c} c & 2c & 22c \\ 1 & 1 & 15 \end{array}\right] = \begin{bmatrix} c & 0 \\ 0 & 1 \end{bmatrix} \left[\begin{array}{cc|c} 1 & 2 & 22 \\ 1 & 1 & 15 \end{array}\right]$$

Likewise, operation (2.) is analogous to left multiplication by the $m \times m$ matrix with c in row j, column i; 1s along the main diagonal; and 0s elsewhere. Explicitly, we obtain the following rule.

2.) Addition of c times the ith row of an $m \times n$ system of linear equations to the jth row of the system corresponds to left multiplication of the $m \times (n+1)$ matrix $\left[A \mid \mathbf{b}\right]$ by the $m \times m$ matrix with c in row j, column i; 1s along the main diagonal; and 0s elsewhere.

Example 1.4.8. Consider the augmented matrix $\left[A \mid \mathbf{b}\right]$ of Example 1.4.7. Observe that in order to subtract the first equation $x_1 + 2x_2 = 22$ from the second equation $x_1 + x_2 = 15$, it suffices to add -1 times the first equation to the second equation. By the previous observation, this can be achieved on the level of matrices by performing the following matrix multiplication.

$$\left[\begin{array}{cc|c} 1 & 2 & 22 \\ 0 & -1 & -7 \end{array}\right] = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \left[\begin{array}{cc|c} 1 & 2 & 22 \\ 1 & 1 & 15 \end{array}\right]$$

Last, operation (3.) is analogous to left multiplication by the $m \times m$ matrix with (i,j)th and (j,i)th entries of 1; 1s along the main diagonal other than in rows i and j; and 0s elsewhere.

3.) Interchanging rows i and j of an $m \times n$ system of linear equations corresponds to left multiplication of the $m \times (n+1)$ matrix $\left[A \mid \mathbf{b}\right]$ by the $m \times m$ matrix with 1 in row j, column i; 1 in row i, column j; 1s along the main diagonal other than rows i and j; and 0s elsewhere.

Example 1.4.9. Once again, consider the augmented matrix $\left[A \mid \mathbf{b}\right]$ of Example 1.4.7. We may interchange the first equation $x_1 + 2x_2 = 22$ and the second equation $x_1 + x_2 = 15$ as follows.

$$\left[\begin{array}{cc|c} 1 & 1 & 15 \\ 1 & 2 & 22 \end{array}\right] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \left[\begin{array}{cc|c} 1 & 2 & 22 \\ 1 & 1 & 15 \end{array}\right]$$

Collectively, we refer to the operations of Definition 1.4.6 as **elementary row operations**; the matrices defined by operations (1.), (2.), and (3.) are therefore called the $m \times m$ **elementary row matrices**. Explicitly, an elementary row matrix is an $m \times m$ matrix obtain by from the $m \times m$ identity matrix I_m by (1.) multiplying any row of I_m by a nonzero scalar c; (2.) adding c times the ith row of I_m to the jth row of I_m; or (3.) interchanging rows i and j of I_m.

Likewise, the operations of Definition 1.4.6 can be defined for the columns of a matrix to obtain the **elementary column operations** and the **elementary column matrices**: we need only swap all instances of "rows" with "columns" and "left multiplication" with "right multiplication."

We will soon see that performing elementary row and column operations on a system of linear equations does not affect the solutions to the system, hence it does not alter the solutions of the underlying matrix equation. Even more, if we employ a sequence of elementary row and column operations to reduce a given augmented matrix to a "relatively simple" form and subsequently interpret the resulting augmented matrix "correctly," then we can easily read off all possible solutions to the underlying system of linear equations. We illustrate this in the case of Example 1.4.8.

Example 1.4.10. Consider the augmented matrix $\left[A \mid \mathbf{b}\right]$ of Example 1.4.8. Converting this back into a system of equations, the second row of the augmented matrix yields that $-x_2 = -7$, hence we conclude that $x_2 = 7$. Consequently, the first row gives that $22 = x_1 + 2x_2 = x_1 + 14$ or $x_1 = 8$. We refer to this as the method of solving a system of linear equations via **back substitution**.

Going forward, we will say that two matrices A and B are **row equivalent** if and only if A can be reduced to B via a sequence of elementary row operations if and only if there exist elementary row matrices $E_1, \dots, E_k$ such that $B = E_k \cdots E_1 A$. Likewise, we make the analogous definition for **column equivalent** matrices. We will write $A \sim B$ if A and B are either row or column equivalent.

Example 1.4.11. By Example 1.4.8 of the previous section, we have that

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$$

are row equivalent because $B = EA$ for the elementary row matrix $E = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$.

By Example 1.4.10, it is clearly advantageous (when possible) to perform a sequence of elementary row operations to reduce a matrix A to a matrix B in which some row has the property that all but one of its entries is nonzero: in this case, the row of B consisting of a single nonzero entry can be used to further reduce A to a matrix possessing more zero entries, as we illustrate next.

Example 1.4.12. Consider the row equivalent matrices A and B of Example 1.4.11. Observe that if we add twice the second row of B to the first row of B, then we obtain the matrix

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

Certainly, matrices with more zero entries are easier to interpret as the collection of coefficients corresponding to some system of linear equations because the variables corresponding to the zeros of the ith row of the matrix do not appear in the ith equation of the system. Even more, the zeros of a matrix inform us about other important properties of the matrix that we will soon discuss. Consequently, we turn our attention in this section to an algorithm that we may employ to reduce a given matrix A to a row equivalent matrix consisting of as many zeros as possible.

We say that a row of an $m \times n$ matrix A is **nonzero** if it contains (at least) one nonzero entry.

Definition 1.4.13. We say that a (real) $m \times n$ matrix A lies in **row echelon form** if and only if

1.) all rows of A consisting entirely of zeros lie beneath the last nonzero row of A and

2.) for any pair of consecutive nonzero rows i and $i+1$, the first nonzero entry of row $i+1$ lies in some column strictly to the right of the column in which the first nonzero entry of row i lies.

Given that A lies in row echelon form, the first nonzero entry of a nonzero row of A is a **pivot**.

Example 1.4.14. Consider the following real matrices.

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 4 \\ 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}$$

Both A and B lie in row echelon form; however, C does not lie in row echelon form because the first nonzero entry of its second row lies in the column directly below the first nonzero of its first row.

We have encountered other instances of matrices in row echelon form, as well: the matrices B of Example 1.4.11 and C of Example 1.4.12 lie in row echelon form; however, the matrix A of Example 1.4.11 does not lie in row echelon form because the first nonzero entry of the second row of A lies directly below the first nonzero entry of the first row of A. Even more, the pivots of the aforementioned matrix B (and C) are 1 in the first row and -1 in the second row. Crucially, the following theorem assures us that it is always possible to reduce any matrix to row echelon form.

Theorem 1.4.15. *Every real matrix is row equivalent to a real matrix in row echelon form.*

Proof. Consider any real $m \times n$ matrix A. Begin by relocating all rows of A consisting entirely of zeros to the bottom of the matrix; interchanging rows corresponds to multiplying on the left by an elementary row matrix, hence the resulting matrix is row equivalent to A. We may disregard all columns of A consisting entirely of zeros because the columns of A do not bear on the row echelon form of A, hence we may assume that the first column of A is nonzero; then, we may find the first nonzero row of A for which the entry in first column of A is nonzero. By interchanging this row with the first row of A, we may ultimately assume that our $m \times n$ matrix A has the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

in which the lowermost rows could consist of zeros and a_{11} is nonzero by assumption. Every nonzero real number has a multiplicative inverse, hence we may subtract $a_{i1}a_{11}^{-1}$ times the first row from the ith row; this corresponds to left multiplication by an elementary row matrix and yields that

$$A \sim \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

for some real numbers $b_{22}, \dots, b_{mn}$. Employing this process with the $(m-1) \times (n-1)$ submatrix

$$B = \begin{bmatrix} b_{22} & \cdots & b_{2n} \\ \vdots & & \vdots \\ b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

and subsequently continuing in this manner, we will eventually reduce A to row echelon form. □

Definition 1.4.16. We say that a matrix lies in **reduced row echelon form** if and only if

1.) it lies in row echelon form;

2.) its pivots are all 1; and

3.) if the jth column contains a pivot, then all of its non-pivot entries are zero. Put another way, the only nonzero entry of any column containing a pivot is the pivot itself.

Corollary 1.4.17. *Every real matrix is row equivalent to a real matrix in reduced row echelon form.*

Proof. By Theorem 1.4.15, every real matrix A is row equivalent to a real matrix B in row echelon form. By multiplying each nonzero row of B by the multiplicative inverse of its pivot, we obtain a row equivalent matrix C whose pivots are all 1. Last, we must ensure that the only nonzero entry of any column containing a pivot is the pivot itself. Observe that if c_{ij} is nonzero and the jth column of C contains a pivot in row k, then we may add $-c_{ij}$ times the kth row of C to the ith row of C to obtain 0 in the ith row and jth column of C. Continuing in this manner yields the result. □

Essentially, the proofs of Theorem 1.4.15 and Corollary 1.4.17 outline the method of **Gaussian Elimination** in systems of linear equations; for completeness, we summarize the results below.

Algorithm 1.4.18 (Gaussian Elimination)**.** Given any nonzero real $m \times n$ matrix A, the following steps will reduce the matrix A to a row equivalent matrix B in reduced row echelon form.

(1.) Begin by relocating all rows of A consisting entirely of zeros to the bottom of the matrix. We may perform this operation because row interchange yields a row equivalent matrix.

(2.) Find the first nonzero row i of the matrix obtained in the previous step for which the entry a_{i1} in first column is nonzero; if this is not the first row, then interchange the first and ith rows of this matrix so that a_{i1} lies in the first row and column of the resulting matrix.

(3.) Multiply the first row of the resulting matrix by the multiplicative inverse a_{i1}^{-1} of the nonzero real number a_{i1} to obtain an entry of 1 in the first row and first column. We may perform this operation because multiplying a row by a nonzero scalar yields a row equivalent matrix.

(4.) If r_j is the component of the jth row and first column of the matrix obtained in step (3.), then add $-r_j$ times the first row of this matrix to the jth row of this matrix for each integer $1 \leq j \leq m$. We may perform this operation because adding a scalar multiple of a row to another row yields a row equivalent matrix. Observe that the only nonzero entry in the first column of the resulting matrix is the pivot of 1 in the first row and first column.

(5.) Repeat steps (2.), (3.), (4.) for the matrix obtained from the resulting matrix of step (4.) by ignoring the first row and first column; if possible, a pivot of 1 is obtained in the second row of this matrix, and all entries of the matrix below this pivot are zero.

(6.) Repeat step (5.) until the row echelon form of A is obtained and all pivots are 1.

(7.) Eliminate any nonzero entry a_{ij} in row i above the pivot 1 in row k by adding $-a_{ij}$ times the kth row of the matrix of step (6.) to the ith row of the matrix.

(8.) Repeat step (7.) until the matrix lies in reduced row echelon form.

We refer to the matrix obtained from this process as the **reduced row echelon form** $\operatorname{RREF}(A)$.

One of the best ways to understand the method of Gaussian Elimination is to practice using it. We illustrate the technique and its applications in the following several examples.

Example 1.4.19. Let us convert the following matrix to reduced row echelon form.

$$A = \begin{bmatrix} 2 & -3 & 7 \\ -1 & 0 & 3 \\ 2 & 1 & 5 \end{bmatrix}$$

Considering that each of the rows of A is nonzero, we may immediately proceed to the second step of the Gaussian Elimination algorithm. Observe that the first nonzero row of A for which the entry in the first column is nonzero is simply the first row of A, so we may proceed to the third step of the algorithm. Explicitly, we multiply the first row of A by $\frac{1}{2}$ (i.e., the multiplicative inverse of 2) to obtain an entry of 1 in the first row and first column of A. We illustrate this as follows.

$$A = \begin{bmatrix} 2 & -3 & 7 \\ -1 & 0 & 3 \\ 2 & 1 & 5 \end{bmatrix} \overset{\frac{1}{2}R_1 \mapsto R_1}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ -1 & 0 & 3 \\ 2 & 1 & 5 \end{bmatrix}$$

We may subsequently reduce all first column entries beneath the first row of the resulting matrix.

$$\begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ -1 & 0 & 3 \\ 2 & 1 & 5 \end{bmatrix} \overset{R_2 + R_1 \mapsto R_2}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & -\frac{3}{2} & \frac{13}{2} \\ 2 & 1 & 5 \end{bmatrix} \overset{R_3 - 2R_1 \mapsto R_3}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & -\frac{3}{2} & \frac{13}{2} \\ 0 & 4 & \frac{3}{2} \end{bmatrix}$$

We have therefore created a pivot of 1 in the first row and first column, so we proceed to do the same for the second row and second column. Explicitly, we multiply the second row of the above matrix by $-\frac{2}{3}$ (i.e., the multiplicative inverse of $-\frac{3}{2}$) to obtain the following row equivalent matrix.

$$\begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & -\frac{3}{2} & \frac{13}{2} \\ 0 & 4 & \frac{3}{2} \end{bmatrix} \overset{-\frac{2}{3}R_2 \mapsto R_2}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 4 & \frac{3}{2} \end{bmatrix}$$

We may then create a pivot of 1 in the second row and second column of this matrix by adding -4 times the second row to the third row, reducing the entry in the third row and second column to 0.

$$\begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 4 & \frac{3}{2} \end{bmatrix} \overset{R_3 - 4R_2 \mapsto R_3}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & \frac{95}{6} \end{bmatrix}$$

Last, we obtain a pivot of 1 in the third row and third column by multiplying by the multiplicative inverse $\frac{6}{95}$ of $\frac{95}{6}$. Ultimately, we obtain the row echelon form of A for which all pivots are 1.

$$\begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & \frac{95}{6} \end{bmatrix} \overset{\frac{6}{95}R_3 \mapsto R_3}{\sim} \begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & 1 \end{bmatrix}$$

We proceed to the seventh and eighth steps of the Gaussian Elimination algorithm. Because there is a pivot in the second row, we eliminate first the nonzero non-pivot entries in the second column.

$$\begin{bmatrix} 1 & -\frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & 1 \end{bmatrix} \overset{R_1 + \frac{3}{2}R_2 \mapsto R_1}{\sim} \begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & 1 \end{bmatrix}$$

Once this is accomplished, we put the matrix in reduced row echelon form as follows.

$$\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & 1 \end{bmatrix} \overset{R_1+3R_3 \mapsto R_1}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{13}{3} \\ 0 & 0 & 1 \end{bmatrix} \overset{R_2+\frac{13}{3}R_3 \mapsto R_2}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Ultimately, the method of Gaussian Elimination illustrates that our original matrix A is in fact row equivalent to the 3×3 identity matrix. We will see in the next section that row equivalence to the $n \times n$ identity matrix is a very important and special property of a square matrix.

Before we conclude this section, we provide two examples that illustrate how all of the topics we have discussed in this section come to bear on the theory of systems of linear equations.

Example 1.4.20. Consider the following real 3×4 system of linear equations.

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 3 \\ x_1 + 2x_3 + 3x_4 &= 4 \\ x_2 + x_4 &= 5 \end{aligned}$$

Converting this system of linear equations into a matrix equation by taking the coefficients of each linear equation as the entries of a 3×4 matrix A, expressing the variables $x_1, \dots, x_4$ as the rows of a 4×1 column vector, and writing the right-hand side as a 3×1 column vector yields the following.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 2 & 3 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$$

Consequently, in order to solve this system of linear equations, it suffices to convert the following 3×5 augmented matrix into its reduced row echelon form by the method of Gaussian Elimination.

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 3 & 4 \\ 0 & 1 & 0 & 1 & 5 \end{array}\right] \overset{R_2-R_1 \mapsto R_2}{\sim} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 3 \\ 0 & -1 & 1 & 2 & 1 \\ 0 & 1 & 0 & 1 & 5 \end{array}\right] \overset{R_2 \leftrightarrow R_3}{\sim} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 3 \\ 0 & 1 & 0 & 1 & 5 \\ 0 & -1 & 1 & 2 & 1 \end{array}\right] \overset{R_3+R_2 \mapsto R_3}{\sim} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 3 \\ 0 & 1 & 0 & 1 & 5 \\ 0 & 0 & 1 & 3 & 6 \end{array}\right]$$

$$\overset{R_1-R_3 \mapsto R_1}{\sim} \left[\begin{array}{cccc|c} 1 & 1 & 0 & -2 & -3 \\ 0 & 1 & 0 & 1 & 5 \\ 0 & 0 & 1 & 3 & 6 \end{array}\right]$$

$$\overset{R_1-R_2 \mapsto R_1}{\sim} \left[\begin{array}{cccc|c} 1 & 0 & 0 & -3 & -8 \\ 0 & 1 & 0 & 1 & 5 \\ 0 & 0 & 1 & 3 & 6 \end{array}\right]$$

Consequently, the 3×4 system is equivalent to the following system in reduced row echelon form.

$$\begin{aligned} x_1 - 3x_4 &= -8 \\ x_2 + x_4 &= 5 \\ x_3 + 3x_4 &= 6 \end{aligned}$$

We obtain the general solution of this system by expressing each of the three variables x_1, x_2, and x_3 in terms of the **free variable** x_4. Crucially, observe that the general solution is given by

$$\boldsymbol{\xi} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3x_4 - 8 \\ -x_4 + 5 \\ -3x_4 + 6 \\ x_4 \end{bmatrix} = x_4 \begin{bmatrix} 3 \\ -1 \\ -3 \\ 1 \end{bmatrix} + \begin{bmatrix} -8 \\ 5 \\ 6 \\ 1 \end{bmatrix}.$$

Consequently, for each assignment of a real number to the free variable x_4, we obtain a unique solution $\boldsymbol{\xi}$. Ultimately, this system of linear equations admits infinitely many solutions, and each solution is determined by the value of x_4 by the above equation. Below are two particular solutions.

$$\text{If } x_4 = 0, \text{ then the particular solution to the system is given by } \boldsymbol{\xi} = \begin{bmatrix} -8 \\ 5 \\ 6 \\ 1 \end{bmatrix}.$$

$$\text{If } x_4 = 3, \text{ then the particular solution to the system is given by } \boldsymbol{\xi} = \begin{bmatrix} 1 \\ 2 \\ -3 \\ 3 \end{bmatrix}.$$

Considering that this system of linear equations admits a solution, we say the system is **consistent**.

Example 1.4.21. Consider the following real 4×3 system of linear equations.

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 0 \\ 4x_1 + 5x_2 + 6x_3 &= 1 \\ 7x_1 + 8x_2 + 9x_3 &= 0 \end{aligned}$$

We obtain an augmented matrix $\left[A \mid \mathbf{b}\right]$ called the **coefficient matrix** corresponding to this system of linear equations by writing down the coefficients of the variables. Each equation is a distinct row. Each variable induces a distinct column. Explicitly, we obtain the following coefficient matrix.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 1 \\ 7 & 8 & 9 & 0 \end{array}\right]$$

We proceed to convert the matrix to reduced row echelon form via Gaussian Elimination.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 1 \\ 7 & 8 & 9 & 0 \end{array}\right] \overset{\substack{R_2 - 4R_1 \mapsto R_2 \\ R_3 - 7R_1 \mapsto R_3}}{\sim} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & -3 & -6 & 1 \\ 0 & -6 & -12 & 0 \end{array}\right] \overset{R_3 - 2R_2 \mapsto R_3}{\sim} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & -3 & -6 & 1 \\ 0 & 0 & 0 & -2 \end{array}\right]$$

We note that from this step, it can be determined that this system of linear equations is **inconsistent**, i.e., it has no solution. Explicitly, observe that the third row of the above augmented matrix implies (on the level of linear equations) that $0 = 0x_1 + 0x_2 + 0x_3 = -2$ — a contradiction.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & -3 & -6 & 1 \\ 0 & 0 & 0 & -2 \end{array}\right] \overset{\substack{-\frac{1}{2}R_3 \mapsto R_3 \\ -\frac{1}{3}R_2 \mapsto R_2}}{\sim} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 2 & -\frac{1}{3} \\ 0 & 0 & 0 & 1 \end{array}\right] \overset{R_2 + \frac{1}{3}R_3 \mapsto R_2}{\sim} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \overset{R_1 - 2R_2 \mapsto R_1}{\sim} \left[\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

1.5 Invertible Matrices

We will assume throughout this section that n is a positive integer. Given any $n \times n$ matrix A, we say that an $n \times n$ matrix L is a **left inverse** of A if it holds that $LA = I$, where we denote by I the $n \times n$ identity matrix. Likewise, we say that an $n \times n$ matrix R is a **right inverse** of A if it holds that $AR = I$. We will establish immediately that every left inverse of A is also a right inverse and vice-versa, hence we may dispense of the distinct notions of left and right inverses of matrices and simply say that an $n \times n$ matrix B is a (two-sided) **inverse** of an $n \times n$ matrix A if it holds that $AB = I = BA$. Our next proposition shows that a two-sided inverse of a matrix A is unique.

Proposition 1.5.1. *Let A be any $n \times n$ matrix. Every left inverse of A is a right inverse of A and vice-versa (provided that both exist). Even more, if A admits a two-sided inverse, then it is unique.*

Proof. Consider any $n \times n$ matrices L and R such that $LA = I = AR$. By Proposition 1.3.21, we have that $R = IR = (LA)R = L(AR) = LI = L$. Consequently, L is a two-sided inverse of A. Even more, if L' is any two-sided inverse of A, then it is a right inverse of A so that $L' = L$. □

Consequently, if an $n \times n$ matrix A admits a (two-sided) inverse, then it is unique, and we may denote it by A^{-1}. We will also say in this case that A is **invertible** (or **non-singular**). Certainly, the zero matrix does not possess an inverse, hence some (and in fact many) matrices are not invertible. We explore next how matrix inverses behave with respect to the matrix operations of Section 1.3.

Proposition 1.5.2. *If A is an invertible $n \times n$ matrix, then $(A^T)^{-1} = (A^{-1})^T$. Put another way, if A is invertible, then A^T is invertible, and its matrix inverse is the transpose of A^{-1}.*

Proof. By Proposition 1.3.24, it follows that $(A^{-1})^T A^T = (AA^{-1})^T = I^T = I$, and we conclude that $(A^T)^{-1} = (A^{-1})^T$ by the uniqueness of the matrix inverse guaranteed by Proposition 1.5.1. □

Proposition 1.5.3. *If $A_1, \dots, A_k$ are any invertible $n \times n$ matrices, then*

$$(A_1 \cdots A_k)^{-1} = A_k^{-1} \cdots A_1^{-1}.$$

Put another way, the product of invertible $n \times n$ matrices is an invertible matrix, and the matrix inverse of the product is the product of the matrix inverses in reverse order.

Proof. By Proposition 1.5.1, it suffices to verify that $(A_k^{-1} \cdots A_1^{-1})(A_1 \cdots A_k) = I$. Considering that $A_i^{-1} A_i = I$ for all integers $1 \leq i \leq k$, we may replace every instance of $A_i^{-1} A_i$ with I; then, using the fact that $IB = B$ for any $n \times r$ matrix B, the result follows after repeating this k times. □

Corollary 1.5.4. *If A is an invertible $n \times n$ matrix, then A^k is invertible for all integers $k \geq 0$.*

Proof. By Proposition 1.5.3, it follows that A^k is invertible with $(A^k)^{-1} = (A^{-1})^k$. □

Corollary 1.5.5. *If A and B are row equivalent, then A is invertible if and only if B is invertible.*

Proof. By definition, an $n \times n$ matrix A is row equivalent to the matrix B if and only if there exist elementary row matrices $E_1, \dots, E_k$ such that $B = E_k \cdots E_1 A$. Considering that $A = E_1^{-1} \cdots E_k^{-1} B$, we conclude that A is invertible if and only if B is invertible by Propositions 1.5.3 and 1.5.6. □

Using the method of Gaussian Elimination, we can determine if an $n \times n$ matrix A admits an inverse, and we may subsequently compute A^{-1} in this way, as well. Before we demonstrate this, we remind the reader that two matrices are row equivalent if and only if there exist some elementary row matrices whose product (on the left) of one matrix gives the other. Explicitly, we have that A and B are row equivalent if and only if there exist elementary row matrices $E_1, \dots, E_k$ such that $B = E_k \cdots E_1 A$. Elementary row matrices are precisely those $n \times n$ matrices obtained from the $n \times n$ identity matrix by performing (at most) one of the following matrix operations.

1.) We may multiply any row of I by a nonzero scalar c.

2.) We may add c times the ith row of I to the jth row of I.

3.) We may interchange any pair of rows i and j of I.

We refer to the above operations as the Elementary Row Operations.

Proposition 1.5.6. *Every elementary row matrix is invertible.*

Proof. Let E be an elementary row matrix. Consider the following three cases.

1.) If E is obtained from I by multiplying the ith row of I by a nonzero scalar c, then E^{-1} is obtained from I by multiplying the ith row of I by the nonzero scalar c^{-1}.

2.) If E is obtained from I by adding c times the ith row of I to the jth row of I, then E^{-1} is obtained from I by adding $-c$ times the ith row of I to the jth row of I.

3.) If E is obtained from I by interchanging rows i and j of I, then E is its own inverse. □

Before we provide several equivalent criteria for the invertibility of a square matrix or establish how to compute a matrix inverse, it is imperative to discuss how theory of systems of linear equations comes to bear on the theory of invertible matrices. Consider the matrix equation $A\mathbf{x} = \mathbf{b}$ for some real $n \times n$ matrix A, the real $n \times 1$ column vector $\mathbf{x}$ whose ith row is a variable x_i, and some real $n \times 1$ column vector $\mathbf{b}$. Crucially, we note that if A is row equivalent to the $n \times n$ identity matrix I, then the matrix equation $A\mathbf{x} = \mathbf{b}$ is consistent (i.e., it admits a solution): indeed, if there exist elementary row matrices $E_1, \dots, E_k$ such that $E_k \cdots E_1 A = I$, then we have that

$$\mathbf{x} = I\mathbf{x} = (E_k \cdots E_1 A)\mathbf{x} = E_k \cdots E_1 (A\mathbf{x}) = E_k \cdots E_1 \mathbf{b}.$$

Conversely, if the matrix equation $A\mathbf{x} = \mathbf{b}$ admits a solution, then A must be row equivalent to the identity matrix. We establish this as follows using a proof by contrapositive.

Theorem 1.5.7. *Given any real $n \times n$ matrix A, the matrix equation $A\mathbf{x} = \mathbf{b}$ admits a solution for every real $n \times 1$ matrix $\mathbf{b}$ if and only if A is row equivalent to the $n \times n$ identity matrix.*

Proof. By the paragraph preceding the statement of the theorem, if A is row equivalent to the $n \times n$ identity matrix, then the matrix equation $A\mathbf{x} = \mathbf{b}$ admits a unique solution for every $n \times 1$ matrix $\mathbf{b}$. Conversely, we will assume that A is not row equivalent to the $n \times n$ identity matrix. Consequently, the nth row of the reduced row echelon form $\operatorname{RREF}(A)$ of the matrix A must be zero. Even more, there exist elementary row matrices $E_1, \dots, E_k$ such that $\operatorname{RREF}(A) = E_k \cdots E_1 A$. By Proposition

1.5.6, each of the $n \times n$ matrices $E_1, \dots, E_k$ is invertible, hence their product $E_k \cdots E_1$ is invertible by Proposition 1.5.3. Consider the real $n \times 1$ matrix $\mathbf{b} = (E_k \dots E_1)^{-1}\mathbf{e}_n$ for the nth standard basis vector $\mathbf{e}_n$ that consists of zeros in each of the first $n-1$ rows and 1 in the nth row. Observe that the matrix equation $A\mathbf{x} = \mathbf{b}$ has no solution: indeed, by construction, we have that

$$\operatorname{RREF}(A)\mathbf{x} = (E_k \cdots E_1 A)\mathbf{x} = E_k \cdots E_1(A\mathbf{x}) = E_k \cdots E_1 \mathbf{b} = \mathbf{e}_n.$$

Considering that the nth row of $\operatorname{RREF}(A)\mathbf{x}$ is 0 and the nth row of $\mathbf{e}_n$ is 1, we have established that it is impossible to obtain a real $n \times 1$ matrix $\mathbf{x}$ for which the matrix equation $A\mathbf{x} = \mathbf{b}$ holds. □

By virtue of Theorem 1.5.7, it follows that any left inverse of an $n \times n$ matrix must be a right inverse, as well. Consequently, the invertibility of a square matrix can be determined by checking whether the matrix can be reduced to the identity matrix. Even more, the unique matrix inverse of a matrix that is row equivalent to the identity matrix is simply the product of the elementary matrices required to put the matrix in reduced row echelon form. We prove this as follows.

Theorem 1.5.8. *Given any $n \times n$ matrices A and B, we have that $AB = I$ if and only if $BA = I$. Explicitly, any left inverse of a square matrix is the unique inverse of the matrix.*

Proof. We will assume first that $AB = I$, and we will demonstrate that $BA = I$. Conversely, we may simply reverse the roles of A and B to find that if $BA = I$, then $AB = I$. Given any $n \times 1$ matrix $\mathbf{b}$, the matrix equation $A\mathbf{x} = \mathbf{b}$ admits a solution $\boldsymbol{\xi} = B\mathbf{b}$: indeed, we have that

$$A\boldsymbol{\xi} = A(B\mathbf{b}) = (AB)\mathbf{b} = I\mathbf{b} = \mathbf{b}.$$

By Theorem 1.5.7, it follows that A is row equivalent to the $n \times n$ identity matrix, hence there exist elementary row matrices $E_1, \dots, E_k$ such that $E_k \cdots E_1 A = I$. By Proposition 1.5.1, in view of the fact that $E_k \cdots E_1$ is a left inverse of A, it follows that $E_k \cdots E_1$ is the unique inverse of A. □

Conversely, we demonstrate that every invertible matrix is row equivalent to the identity matrix. By Corollary 1.4.17, a matrix is row equivalent to its reduced row echelon form. By Corollary 1.5.5, an $n \times n$ matrix A is invertible if and only if $\operatorname{RREF}(A)$ is invertible. Particularly, if $\operatorname{RREF}(A)$ admits any rows consisting entirely of zeros, then it is not invertible (because the last row of $\operatorname{RREF}(A)B$ is zero for all $n \times r$ matrices B), hence the underlying matrix A cannot be invertible. On the other hand, we will establish that if all rows of $\operatorname{RREF}(A)$ are nonzero, then it is invertible, hence A is invertible. Before this, we mention that an **upper-triangular matrix** is an $n \times n$ matrix with the property that the (i, j)th component of the matrix is zero for all integers $1 \leq i < j \leq n$. Put another way, all entries below the main diagonal of an upper-triangular matrix are zero.

Theorem 1.5.9. *Every upper-triangular matrix with nonzero diagonal elements is invertible.*

Proof. By definition, every $n \times n$ upper-triangular matrix U can be written as follows.

$$U = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

By hypothesis that a_{ii} is nonzero for each integer $1 \leq i \leq n$, we may multiply the ith row of the above matrix by a_{ii}^{-1} to obtain an upper-triangular matrix whose pivots are all 1. Each of these products corresponds to multiplication of U (on the left) by an elementary row matrix, hence this process does not come to bear on the existence of an inverse of U. Consequently, we may assume from the beginning that this is the case, i.e., we may restrict our attention to the following situation.

$$U = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ 0 & 1 & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

By Corollary 1.5.5, it suffices to demonstrate that U is row equivalent to the invertible $n \times n$ identity matrix I. We achieve this by furnishing some elementary row operations that reduces U to I. Observe that if we add $-a_{in}$ times the last row of U to the ith row of U, then we obtain a 0 in the (i, n)th component of the resulting matrix. Continuing in this way, we may reduce the nth column of U to zero except in the bottom right-hand corner. Considering that adding any scalar multiple of a row of U to another row of U is a row equivalence, we conclude that U is row equivalent to this matrix. Continuing in this way for each column of U from right to left, it follows that U is row equivalent to the identity matrix. By Theorem 1.5.8, we conclude that U is invertible. □

Corollary 1.5.10. *An $n \times n$ matrix is invertible if and only if it is row equivalent to the $n \times n$ identity matrix. Even more, we may obtain the unique matrix inverse via Gaussian Elimination.*

Proof. By Theorems 1.5.7 and 1.5.8, a matrix that is row equivalent to the identity matrix must be invertible. Conversely, by Proposition 1.5.5, Theorem 1.5.9, and the paragraph that precedes the theorem, an $n \times n$ matrix A is invertible if and only if the upper-triangular matrix $\operatorname{RREF}(A)$ is invertible if and only if $\operatorname{RREF}(A) = I$. Consequently, if A is an invertible $n \times n$ matrix, then there exist elementary row matrices $E_1, \dots, E_k$ such that $E_k \cdots E_1 A = I$, from which we conclude by Theorem 1.5.7 that the unique inverse of A is given by $A^{-1} = E_k \cdots E_1$. □

Corollary 1.5.11. *Every invertible $n \times n$ matrix is a product of elementary row matrices.*

Proof. By the proof of Corollary 1.5.10, every invertible $n \times n$ matrix A admits some elementary row matrices $E_1, \dots, E_k$ such that $E_k \cdots E_1 A = I$. By multiplying both sides on the left by $E_1^{-1} \cdots E_k^{-1}$, we obtain that $A = E_1^{-1} \cdots E_k^{-1}$. By the proof of Proposition 1.5.6, each of the matrices $E_1^{-1}, \dots, E_k^{-1}$ is an elementary row matrix, hence A is the product of elementary row matrices. □

Generally, the method of Gaussian Elimination can in practice be implemented to determine if a square matrix is invertible and to explicitly produce the inverse of such a matrix. Observe that if A is an $n \times n$ matrix, then we may construct the augmented matrix $\left[A \mid I\right]$ by adjoining the $n \times n$ identity matrix I on the right-hand side of A. By performing elementary row operations, we may reduce A to its reduced row echelon form $\operatorname{RREF}(A)$. Consequently, if A is invertible, this will reduce A to I and simultaneously convert I to A^{-1}. Explicitly, this process yields that $\left[A \mid I\right] \sim \left[I \mid A^{-1}\right]$.

Example 1.5.12. Consider the following 2×2 matrix A and the augmented matrix $\left[A \mid I\right]$.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \qquad \left[A \mid I\right] = \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 5 & 0 & 1 \end{array}\right]$$

We will carry out the Gaussian Elimination as follows, listing each elementary row operation.

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 5 & 0 & 1 \end{array}\right] \overset{R_2-3R_1\mapsto R_2}{\sim} \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -1 & -3 & 1 \end{array}\right] \overset{-R_2\mapsto R_2}{\sim} \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & 3 & -1 \end{array}\right] \overset{R_1-2R_2\mapsto R_1}{\sim} \left[\begin{array}{cc|cc} 1 & 0 & -5 & 2 \\ 0 & 1 & 3 & -1 \end{array}\right]$$

Consequently, we find that A is an invertible 2×2 matrix with the following matrix inverse.

$$A^{-1} = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}$$

Example 1.5.13. Consider the following 3×3 matrix A and the augmented matrix $\big[A \mid I\big]$.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix} \qquad \big[A \mid I\big] = \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 \\ 1 & 2 & 2 & 0 & 0 & 1 \end{array}\right]$$

We will carry out the Gaussian Elimination as follows, listing each elementary row operation.

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 \\ 1 & 2 & 2 & 0 & 0 & 1 \end{array}\right] \overset{\substack{R_2-R_1\mapsto R_2 \\ R_3-R_1\mapsto R_3}}{\sim} \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 1 & 1 & -1 & 0 & 1 \end{array}\right] \overset{R_2\leftrightarrow R_3}{\sim} \left[\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{array}\right]$$

$$\overset{\substack{R_1-R_3\mapsto R_1 \\ R_2-R_3\mapsto R_2}}{\sim} \left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 2 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{array}\right]$$

$$\overset{R_1-R_2\mapsto R_1}{\sim} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 2 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{array}\right]$$

By the paragraph preceding Example 1.5.12, we conclude that the inverse of A is given as follows.

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

Example 1.5.14. Let us determine a numerical criterion for which a real 2×2 matrix is invertible by performing Gaussian Elimination to obtain the reduced row echelon form. Consider any matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

such that a, b, c, and d are real numbers. Observe that if $a = 0$ and $c = 0$, then A is not invertible because the first row of the matrix BA will be zero for all real $m \times 2$ matrices B. Consequently, we may assume that a is nonzero. By multiplying the first row of A by a^{-1}, we obtain the following.

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \overset{a^{-1}R_1\mapsto R_1}{\sim} \begin{bmatrix} 1 & a^{-1}b \\ c & d \end{bmatrix}$$

Equivalently, the displayed matrix above is E_1A for the following elementary row matrix

$$E_1 = \begin{bmatrix} a^{-1} & 0 \\ 0 & 1 \end{bmatrix}$$

We may subsequently create a pivot in the first row and first column of E_1A by adding $-c$ times the first row of E_1A to the second row of E_1A. Explicitly, we obtain the following.

$$E_1A = \begin{bmatrix} 1 & a^{-1}b \\ c & d \end{bmatrix} \overset{R_2 - cR_1 \mapsto R_2}{\sim} \begin{bmatrix} 1 & a^{-1}b \\ 0 & d - a^{-1}bc \end{bmatrix}$$

Equivalently, the displayed matrix above is E_2E_1A for the following elementary row matrix.

$$E_2 = \begin{bmatrix} 1 & 0 \\ -c & 1 \end{bmatrix}$$

Observe that if $d - a^{-1}bc = 0$, then the last row of E_2E_1A is zero, hence it is not invertible so that A is not invertible. Consequently, we must have that $d - a^{-1}bc$ is nonzero, i.e., we must have that $ad - bc$ is nonzero. Continuing onward, because $d - a^{-1}bc$ is nonzero, it possesses a multiplicative inverse $(d - a^{-1}bc)^{-1}$. By multiplying the last row of E_2E_1A by $(d - a^{-1}bc)^{-1}$, obtain the following.

$$E_2E_1A = \begin{bmatrix} 1 & a^{-1}b \\ 0 & d - a^{-1}bc \end{bmatrix} \overset{(d - a^{-1}bc)^{-1}R_2 \mapsto R_2}{\sim} \begin{bmatrix} 1 & a^{-1}b \\ 0 & 1 \end{bmatrix}$$

Equivalently, the displayed matrix above is $E_3E_2E_1A$ for the following elementary row matrix.

$$E_3 = \begin{bmatrix} 1 & 0 \\ 0 & (d - a^{-1}bc)^{-1} \end{bmatrix}$$

Last, by adding $-(d - a^{-1}bc)^{-1}$ times the second row of A to the first row of A, we obtain a pivot in the second row and second column. Explicitly, if we multiply $E_3E_2E_1A$ on the left by

$$E_4 = \begin{bmatrix} 1 & -a^{-1}b \\ 0 & 1 \end{bmatrix},$$

then we obtain $E_4E_3E_2E_1A = I_{2\times 2}$ so that $A^{-1} = E_4E_3E_2E_1$. Explicitly, the following holds.

$$A^{-1} = \begin{bmatrix} 1 & -a^{-1}b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (d - a^{-1}bc)^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -c & 1 \end{bmatrix} \begin{bmatrix} a^{-1} & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Consequently, our original matrix A is invertible if and only if $ad - bc$ is nonzero.

1.6 Chapter Overview

This section is currently under construction.

References

[Con22] K. Conrad. *The Minimal Polynomial and Some Applications*. 2022. URL: https://kconrad.math.uconn.edu/blurbs/linmultialg/minpolyandappns.pdf.

[DF04] D.S. Dummit and R.M. Foote. *Abstract Algebra*. 3rd ed. John Wiley & Sons, Inc., 2004.

[FB95] J.B. Fraleigh and R.A. Beauregard. *Linear Algebra*. 3rd ed. Addison-Wesley Publishing Company, 1995.

[HK71] K. Hoffman and R. Kunze. *Linear Algebra*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1971.

[Lan86] S. Lang. *Introduction to Linear Algebra*. 2nd ed. Undergraduate Texts in Mathematics. Springer-Verlag New York, Inc., 1986.

[Lan87] S. Lang. *Introduction to Linear Algebra*. 3rd ed. Undergraduate Texts in Mathematics. Springer-Verlag New York, Inc., 1987.

[McK22] J. McKinno. *The Principal Axis Theorem*. 2022. URL: https://www.math.uwaterloo.ca/~jmckinno/Math225/Week7/Lecture2m.pdf.

[Moo68] J.T. Moore. *Elements of Linear Algebra and Matrix Theory*. International Series in Pure and Applied Mathematics. McGraw-Hill Book Company, 1968.

[Smi17] K.E. Smith. *The Spectral Theorem*. 2017. URL: http://www.math.lsa.umich.edu/~kesmith/SpectralTheoremW2017.pdf.

[Str06] G. Strang. *Linear Algebra and Its Applications*. 4th ed. Cengage Learning, 2006.