Linear Algebra

Mohammad Safdari

Contents

Chapter 1

Systems of Linear Equations

1.1 Fields

Definition 1.1. A **field** is a nonempty set F equipped with two binary operations

$$\begin{array}{c} F \times F \longrightarrow F \\ (a,b) \mapsto a+b \end{array}' \qquad \begin{array}{c} F \times F \longrightarrow F \\ (a,b) \mapsto ab \end{array}'$$

called respectively **addition** and **multiplication**, such that

(i) The operations are *associative* and *commutative*, i.e. for every $a, b, c \in F$

$$\begin{aligned} a+(b+c) &= (a+b)+c, & a(bc) &= (ab)c, \\ a+b &= b+a, & ab &= ba. \end{aligned}$$

(ii) There exist elements $0, 1 \in F$, called respectively *zero* and *identity* of F, such that $1 \neq 0$, and for every $a \in F$

$$a+0 = a, \qquad a1 = a.$$

(iii) Every $a \in F$ has an **opposite**, i.e. there exists $b \in F$ such that

$$a+b = 0.$$

(iv) Every $a \in F - \{0\}$ has an **inverse**, i.e. there exists $c \in F$ such that

$$ac = 1.$$

(v) Multiplication is *distributive* over addition, i.e. for every $a, b, c \in F$

$$a(b+c) = ab + ac.$$

Remark. It is easy to show that the zero and identity of F are unique. Also, for any $a \in F$, its opposite and its inverse (if $a \neq 0$) are unique, and will be denoted by $-a$ and a^{-1} respectively. For the proofs see Section A.1. In addition, note that due to the commutativity we have

$$0 + a = a, \qquad 1a = a, \qquad (-a) + a = 0.$$

We also have $a^{-1}a = 1$, if $a \neq 0$.

Notation. The **subtraction** and the **division** of two elements a, b of a field, are respectively defined as follows

$$a - b := a + (-b), \qquad a/b = \frac{a}{b} := ab^{-1} \text{ when } b \neq 0.$$

Remark. Informally, a field is a structure in which we can perform the four basic arithmetic operations, i.e. addition, subtraction, multiplication, and division.

***Example* 1.2.** $\mathbb{Q}, \mathbb{R}$ are fields with the usual addition and multiplication. $\mathbb{Z}$ is not a field as it has nonzero elements with no (integer) multiplicative inverse, although it has all the other properties of a field.

Proposition 1.3. *Let F be a field. Then for all $a, b, c \in F$ we have*

(i) **(Cancellation Laws)**

$$a + c = b + c \implies a = b,$$
$$ac = bc,\ c \neq 0 \implies a = b.$$

(ii) $0a = 0 = a0$. *And* $ab = 0 \implies a = 0$ *or* $b = 0$.
(iii) $-(-a) = a$, *and* $-(a + b) = (-a) + (-b) = -a - b$.
(iv) *If* $a \neq 0$ *then* $(a^{-1})^{-1} = a$. *And if* $a, b \neq 0$ *then* $(ab)^{-1} = a^{-1}b^{-1}$.
(v) $(-a)b = -ab = a(-b)$, *and* $(-a)(-b) = ab$.
(vi) $-a = (-1)a$, *and for* $a \neq 0$ *we have* $(-a)^{-1} = -a^{-1}$.

Proof. The proofs can be found in Section A.1. ■

Remark. It can happen in a field that $\overbrace{1 + 1 + \cdots + 1}^{p \text{ times}} = 0$ for some prime integer p. In this case we say that the **characteristic** of the field is p. If this does not happen we say that the characteristic of the field is zero. $\mathbb{Q}, \mathbb{R}$ are of characteristic zero.

1.2 Complex Numbers

Definition 1.4. The set $\mathbb{C}$ of **complex numbers** is the set $\mathbb{R}^2$ equipped with the following addition and multiplication

$$(a,b)+(c,d) := (a+c, b+d),$$
$$(a,b)(c,d) := (ac-bd, ad+bc).$$

Theorem 1.5. *$\mathbb{C}$ is a field, whose zero and identity are respectively*

$$(0,0), \text{ and } (1,0).$$

Also the opposite of a complex number $z=(a,b)$ is

$$-z := (-a,-b),$$

and when z is nonzero its inverse is

$$z^{-1} := \Big(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\Big).$$

Proof. Exercise. ■

Remark. It is easy to see that the characteristic of $\mathbb{C}$ is zero, since the characteristic of $\mathbb{R}$ is zero.

Remark. The map $a \mapsto (a,0)$ from $\mathbb{R}$ into $\mathbb{C}$ is a one-to-one map that preserves addition and multiplication, i.e.

$$(a,0)+(b,0) = (a+b,0), \qquad (a,0)(b,0) = (ab,0).$$

Thus $\mathbb{C}$ contains a copy of the field $\mathbb{R}$. We will abuse the notation and denote the element $(a,0)$ by a. We also define $i := (0,1)$. Then any complex number $z=(a,b)$ can be written as

$$z = (a,b) = (a,0)+(0,b) = (a,0)+(0,1)(b,0) = a+ib.$$

Note that we have

$$i^2 = (0,1)^2 = (-1,0) = -1,$$

i.e. i is a square root of -1.

Definition 1.6. Let $z=(a,b)=a+ib$ be a complex number. The real numbers a,b are called the **real part** and the **imaginary part** of z, respectively, and we will denote them by

$$a = \operatorname{Re} z, \qquad b = \operatorname{Im} z.$$

The **conjugate** of z is the complex number

$$\bar{z} := (a, -b) = a - ib.$$

The **modulus** or the **absolute value** of z is the nonnegative real number

$$|z| := \sqrt{a^2 + b^2}.$$

Remark. Note that $|z| \geq 0$, and $|z| = 0 \iff z = 0$.

Remark. We can define the integer powers of complex numbers, as we did in Section A.1 in a more general setting. Then all the basic properties of powers expressed in Theorem A.13 also hold for powers of complex numbers.

Theorem 1.7. *For all $z, w \in \mathbb{C}$ and $n \in \mathbb{Z}$ we have*

(i) $\overline{z + w} = \bar{z} + \bar{w}$.
(ii) $\overline{zw} = \bar{z}\bar{w}$.
(iii) $\bar{\bar{z}} = z$.
(iv) $z\bar{z} = |z|^2$, *hence* $z^{-1} = |z|^{-2}\bar{z}$.
(v) $|\bar{z}| = |z|$.
(vi) $|z + w| \leq |z| + |w|$.
(vii) $|zw| = |z||w|$.
(viii) $|\operatorname{Re} z| \leq |z|$, *and* $|\operatorname{Im} z| \leq |z|$.
(ix) $z = \bar{z}$ *if and only if* $z \in \mathbb{R}$.
(x) $z + \bar{z} = 2 \operatorname{Re} z$, *and* $z - \bar{z} = 2i \operatorname{Im} z$.
(xi) $|z^n| = |z|^n$ *(when $z = 0$ we assume $n > 0$).*
(xii) $\overline{z^n} = \bar{z}^n$ *(when $z = 0$ we assume $n > 0$).*

Proof. Exercise. ■

Remark. Suppose $z = a + ib$ is nonzero. Then $r := |z| > 0$. Now $\frac{z}{r}$ has modulus one, so it belongs to the unit circle in $\mathbb{C}$. Hence there is a unique $\theta \in [0, 2\pi)$ such that $\frac{z}{r} = e^{i\theta} = \cos\theta + i \sin\theta$. Therefore

$$z = re^{i\theta} = r(\cos\theta + i \sin\theta).$$

This is called the **polar representation** of z. The number θ is called the **argument** of z, and is denoted by $\arg z$. In fact θ is the signed angle between the segment connecting z and 0, and the half line of nonnegative real numbers.

Remark. Suppose $z = re^{i\theta}$ and $w = se^{i\phi}$. Then we have

$$zw = rse^{i(\theta+\phi)}.$$

The interpretation of this formula is that when you multiply a complex number w by a complex number z, you scale the modulus of w by the modulus of z, and you rotate w around the origin by the angle $\arg z$.

Definition 1.8. A field F is called **algebraically closed**, if every nonconstant polynomial with coefficients in F has at least one root in F.

***Example* 1.9.** By the fundamental theorem of algebra, $\mathbb{C}$ is an algebraically closed field. But the fields $\mathbb{R}, \mathbb{Q}$ are not algebraically closed. For example, the polynomial $x^2 + 1$ does not have a root in $\mathbb{R}$, and the polynomial $x^2 - 2$ does not have a root in $\mathbb{Q}$.

1.3 Matrices

Definition 1.10. Let F be a field, and $m, n \in \mathbb{N}$. An $m \times n$ **matrix** with entries in F is a function

$$A : \{(i, j) : i, j \in \mathbb{N},\ i \leq m,\ j \leq n\} \to F.$$

We denote by A_{ij} (or $A_{i,j}$) the value of A at (i, j), and call it the ij-th **entry** of A. The matrix A is usually denoted as a rectangular array of elements of F with m rows and n columns

$$A = [A_{ij}] = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}.$$

The $1 \times n$ matrix $[A_{i1}, \ldots, A_{in}]$ is called the i-th **row** of A, and is denoted by $A_{i,\cdot}$. Also, the $m \times 1$ matrix

$$\begin{bmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{bmatrix}$$

is called the j-th **column** of A, and is denoted by $A_{\cdot,j}$. A $1 \times n$ matrix is also called a **row vector**, and an $m \times 1$ matrix is also called a **column vector**. The set of $m \times n$ matrices with entries in F is denoted by $F^{m \times n}$. The **size** of a matrix $A \in F^{m \times n}$ is $m \times n$.

Remark. We know that F^n is the set of *ordered n-tuples* of elements of F. In order to make this precise, we can define F^n to be the set of functions

$$a : \{1, 2, \ldots, n\} \to F.$$

Then we denote by a_i the value of a at i, and we call it the i-th **component** of a. We will denote a by the following familiar notation

$$a = (a_1, \ldots, a_n),$$

and we also call it a **vector**. We can identify F^n with both $F^{1\times n}$ and $F^{n\times 1}$ via the maps

$$(a_1,\dots,a_n)\mapsto [a_1,\dots,a_n],$$
$$(a_1,\dots,a_n)\mapsto \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.$$

In particular, we always identify F with $F^{1\times 1}$. We also refer to the i,1-th entry of a column vector, or the 1,i-th entry of a row vector, as the i-th **component** of them.

Remark. Note that as matrices are functions into F, it suffices to define them by specifying their ij-th entry for every i, j. Also, when we want to show that two matrices are equal, it is enough to check the equality of their ij-th entry for each i, j. The same things apply to the elements of F^n.

Definition 1.11. Let F be a field, and $m, n \in \mathbb{N}$. The $m \times n$ **zero matrix** is a matrix whose entries are all zero. We often denote the zero matrix simply by 0. A **square matrix** is a matrix for which $m = n$, i.e. a matrix that has the same number of rows and columns. The **(main) diagonal** of a square matrix A is the n-tuple $(A_{11}, A_{22}, \dots, A_{nn}) \in F^n$. The entries A_{ii} are referred to as the *diagonal entries* of A. The square matrix A is called **upper triangular** if $A_{ij} = 0$ for $j < i$. In other words, the entries of A below its main diagonal are zero, so A has the form

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}.$$

Similarly, a square matrix A is called **lower triangular** if $A_{ij} = 0$ for $j > i$. A **diagonal matrix** is a square matrix A for which $A_{ij} = 0$ when $i \neq j$, so it has the form

$$\begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}.$$

A special diagonal matrix is the $n \times n$ **identity matrix**, which is defined by

$$I_{ij} = (I_n)_{ij} := \begin{cases} 0 & i \neq j, \\ 1 & i = j. \end{cases}$$

Definition 1.12. Let F be a field, and $m, n \in \mathbb{N}$. The **addition** of two $m \times n$ matrices A, B with entries in F, is defined by

$$(A + B)_{ij} := A_{ij} + B_{ij}.$$

The **multiplication** of an $m \times n$ matrix A with an $n \times l$ matrix B is an $m \times l$ matrix AB, which is defined by

$$(AB)_{ij} := \sum_{k=1}^{n} A_{ik} B_{kj}.$$

The **scalar multiplication** of $a \in F$ and $A \in F^{m \times n}$ is defined by

$$(aA)_{ij} := aA_{ij}.$$

The **transpose** of an $m \times n$ matrix A is the $n \times m$ matrix $A^{\intercal}$ that satisfies

$$(A^{\intercal})_{ij} := A_{ji}.$$

Notation. For a matrix A we set $-A := (-1)A$, so $(-A)_{ij} = -A_{ij}$. Also, for two $m \times n$ matrices A, B we set $A - B := A + (-B)$.

Remark. Remember that we can identify F^n with both $F^{n \times 1}$ and $F^{1 \times n}$. These identifications allow us to apply the above operations to the elements of F^n. In particular the addition and scalar multiplication on F^n are defined as follows

$$\begin{aligned}(a_1, \ldots, a_n) + (b_1, \ldots, b_n) &:= (a_1 + b_1, \ldots, a_n + b_n),\\ a(a_1, \ldots, a_n) &:= (aa_1, \ldots, aa_n),\end{aligned}$$

where $a \in F$ and $(a_1, \ldots, a_n), (b_1, \ldots, b_n) \in F^n$. In addition, the zero vector is $0 = (0, \ldots, 0)$, and we set $-(a_1, \ldots, a_n) := (-a_1, \ldots, -a_n)$. Note that these operations will also have the properties stated in the next theorem, since they are equivalent to the operations on matrices.

Remark. Let $A, B \in F^{m \times n}$ and $a \in F$. It is easy to show that for every i, j we have

$$\begin{aligned}(A + B)_{i,.} &= A_{i,.} + B_{i,.}, \qquad (aA)_{i,.} = aA_{i,.}, \qquad (A_{i,.})^{\intercal} = A^{\intercal}_{.,i},\\ (A + B)_{.,j} &= A_{.,j} + B_{.,j}, \qquad (aA)_{.,j} = aA_{.,j}, \qquad (A_{.,j})^{\intercal} = A^{\intercal}_{j,.}.\end{aligned}$$

Theorem 1.13. *Let F be a field. Then for all $L \in F^{p \times m}$, $A, B, E \in F^{m \times n}$, $C \in F^{n \times l}$, and $a, b \in F$ we have*

(i) *The addition of matrices is associative and commutative, i.e.*

$$A + (B + E) = (A + B) + E, \qquad A + B = B + A.$$

(ii) *Let* $0 \in F^{m\times n}$ *be the zero matrix, then*

$$A + 0 = A, \qquad A + (-A) = 0.$$

(iii) $1A = A$, *and* $I_m A = A = AI_n$.

(iv) *We have*

$$L(A + B) = LA + LB, \qquad (A + B)C = AC + BC.$$

(v) *We have*

$$\begin{aligned} a(A + B) &= aA + aB, & (a + b)A &= aA + bA, \\ (aA)C &= a(AC) = A(aC), & a(bA) &= (ab)A. \end{aligned}$$

(vi) *If* A *or* C *is the zero matrix, then* AC *is the zero matrix. Also, if* a *is zero, or* A *is the zero matrix, then* aA *is the zero matrix.*

(vii) *We have*

$$\begin{aligned} (A + B)^\intercal &= A^\intercal + B^\intercal, & (aA)^\intercal &= aA^\intercal, \\ (AC)^\intercal &= C^\intercal A^\intercal, & (A^\intercal)^\intercal &= A. \end{aligned}$$

Proof. The proofs can be found in Section A.2. ∎

Remark. As a consequence of the above theorem, we can easily show by induction that if $A_1, \ldots, A_k \in F^{n\times n}$ then we have

$$(A_1 \cdots A_k)^\intercal = A_k^\intercal \cdots A_1^\intercal.$$

Theorem 1.14. *The multiplication of matrices is associative, i.e. for any field* F *and all matrices* $A \in F^{p\times m}$, $B \in F^{m\times n}$, *and* $C \in F^{n\times l}$, *we have*

$$(AB)C = A(BC).$$

Proof. We have

$$\begin{aligned} \big((AB)C\big)_{ij} &= \sum_{k=1}^{n} (AB)_{ik} C_{kj} = \sum_{k=1}^{n} \Big(\sum_{l=1}^{m} A_{il} B_{lk} \Big) C_{kj} \\ &= \sum_{k=1}^{n} \sum_{l=1}^{m} A_{il} B_{lk} C_{kj} = \sum_{l=1}^{m} \sum_{k=1}^{n} A_{il} B_{lk} C_{kj} \\ &= \sum_{l=1}^{m} A_{il} \Big(\sum_{k=1}^{n} B_{lk} C_{kj} \Big) = \sum_{l=1}^{m} A_{il} (BC)_{lj} = \big(A(BC)\big)_{ij}. \end{aligned}$$

∎

***Example* 1.15.** Let $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ be matrices in $F^{2\times 2}$, for some field F. Then we have

$$AB = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = BA.$$

Hence the multiplication of matrices is not in general commutative. This example also shows that the product of two nonzero matrices can be zero. Hence the cancellation law does not hold for matrix multiplication, i.e. for $A, B, C \in F^{n\times n}$

$$AC = BC \not\Longrightarrow A = B.$$

Theorem 1.16. *Suppose F is a field, and $A \in F^{m\times n}$, $C \in F^{n\times l}$. Then we have*

$$(AC)_{ij} = A_{i,.}C_{.,j}, \qquad (AC)_{.,j} = AC_{.,j}, \qquad (AC)_{i,.} = A_{i,.}C.$$

Remark. In other words, the j-th column of AC is the product of A and the j-th column of C. And the i-th row of AC is the product of the i-th row of A, and C.

Proof. Since $A_{i,.}$ and $C_{.,j}$ are respectively $1\times n$ and $n\times 1$ matrices, their product is a 1×1 matrix, i.e. an element of F; and we have

$$(A_{i,.}C_{.,j})_{1,1} = \sum_{k\leq n}(A_{i,.})_{1,k}(C_{.,j})_{k,1} = \sum_{k\leq n} A_{i,k}C_{k,j} = (AC)_{ij}.$$

Similarly, $(AC)_{.,j}$ and $(AC)_{i,.}$ are respectively $m\times 1$ and $1\times l$ matrices. Hence we have

$$\big((AC)_{.,j}\big)_{i,1} = (AC)_{i,j} = \sum_{k\leq n} A_{ik}C_{kj} = \sum_{k\leq n} A_{ik}(C_{.,j})_{k,1} = (AC_{.,j})_{i,1},$$
$$\big((AC)_{i,.}\big)_{1,j} = (AC)_{i,j} = \sum_{k\leq n} A_{ik}C_{kj} = \sum_{k\leq n}(A_{i,.})_{1,k}C_{kj} = (A_{i,.}C)_{1,j}. \qquad \blacksquare$$

***Exercise* 1.17.** Suppose $A \in F^{m\times n}$ and $C \in F^{n\times l}$. Show that

$$AC = \sum_{k\leq n} A_{.,k}C_{k,.}.$$

Solution. Note that $A_{.,k}$ and $C_{k,.}$ are respectively $m\times 1$ and $1\times l$ matrices. Hence their product is an $m\times l$ matrix. Now we have

$$\Big(\sum_{k\leq n} A_{.,k}C_{k,.}\Big)_{i,j} = \sum_{k\leq n}\big(A_{.,k}C_{k,.}\big)_{i,j} = \sum_{k\leq n}\sum_{s=1}^{1}(A_{.,k})_{i,s}(C_{k,.})_{s,j}$$
$$= \sum_{k\leq n}(A_{.,k})_{i,1}(C_{k,.})_{1,j} = \sum_{k\leq n} A_{i,k}C_{k,j} = (AC)_{i,j},$$

as desired. $\blacksquare$

Notation. Let $j, n \in \mathbb{N}$, and suppose $j \leq n$. We denote by e_j the column vector in $F^{n\times 1}$ whose components are all zero except for its j-th component which is one, i.e.

$$e_j := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{-th row.}$$

We call this the j-th vector of the *standard basis* of $F^{n\times 1}$. We also have

$$e_j^\intercal = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \in F^{1\times n}.$$

We call this the j-th vector of the standard basis of $F^{1\times n}$. We also sometimes abuse the notation and call e_j's or $e_j^\intercal$'s the standard basis vectors of F^n. Note that we use the same notation for every n. For example e_1 can be any of the followings

$$[1],\ \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},\ \dots .$$

But this should cause no confusion, since the value of n is usually evident from the context.

Remark. Let I be the identity matrix. Then we have $I_{i,.} = e_i^\intercal$, and $I_{.,j} = e_j$.

Theorem 1.18. *Suppose F is a field, and $A \in F^{m\times n}$. Let $x = [x_1, \dots, x_n]^\intercal \in F^{n\times 1}$ be a column vector, and let $y = [y_1, \dots, y_m] \in F^{1\times m}$ be a row vector. Then we have*

$$Ax = \sum_{j\leq n} x_j A_{.,j}, \qquad yA = \sum_{i\leq m} y_i A_{i,.}.$$

In particular for $e_j \in F^{n\times 1}$ and $e_i^\intercal \in F^{1\times m}$ we have

$$Ae_j = A_{.,j}, \qquad e_i^\intercal A = A_{i,.}.$$

Remark. We say that Ax is a *linear combination* of the columns of A, and yA is a linear combination of the rows of A.

Proof. We know that Ax and yA are respectively $m\times 1$ and $1\times n$ matrices. Then we have

$$(Ax)_{i,1} = \sum_{j\leq n} A_{ij}x_j = \sum_{j\leq n} x_j (A_{.,j})_{i,1} = \Big(\sum_{j\leq n} x_j A_{.,j}\Big)_{i,1},$$
$$(yA)_{1,j} = \sum_{i\leq m} y_i A_{ij} = \sum_{i\leq m} y_i (A_{i,.})_{1,j} = \Big(\sum_{i\leq m} y_i A_{i,.}\Big)_{1,j}.$$

The final statement of the theorem is a trivial consequence of the above relations, and the special form of the standard basis vectors. ■

***Exercise* 1.19.** Show that $A \in F^{n\times n}$ is diagonal if and only if for every $j \leq n$ we have

$$A_{\cdot,j} = A_{jj}e_j.$$

Remark. Similarly we can show that A is diagonal if and only if for every $i \leq n$ we have $A_{i,\cdot} = A_{ii}e_i^{\intercal}$.

Solution. Suppose A is diagonal. Then we have

$$A_{\cdot,j} = [A_{1j}, \ldots, A_{jj}, \ldots, A_{nj}]^{\intercal} = [0, \ldots, 0, A_{jj}, 0, \ldots, 0]^{\intercal} = A_{jj}e_j.$$

Conversely if $A_{\cdot,j} = A_{jj}e_j$ then we get $A_{\cdot,j} = A_{jj}e_j = [0, \ldots, 0, A_{jj}, 0, \ldots, 0]^{\intercal}$. Hence $A_{ij} = 0$ for every $i \neq j$. Thus A is diagonal. ■

Definition 1.20. Let F be a field. A square matrix $A \in F^{n\times n}$ is called **invertible** if there is $B \in F^{n\times n}$ such that

$$AB = I_n = BA.$$

We say B is an **inverse** of A. Also, we say two matrices $A, C \in F^{n\times n}$ **commute** if

$$AC = CA.$$

Theorem 1.21. *Suppose F is a field, and $A, C \in F^{n\times n}$ are invertible matrices. Then*

(i) *The inverse of A is unique, and we denote it by A^{-1}.*

(ii) *A^{-1} and $A^{\intercal}$ are also invertible, and*

$$(A^{-1})^{-1} = A, \qquad (A^{\intercal})^{-1} = (A^{-1})^{\intercal}.$$

(iii) *AC is also invertible, and*

$$(AC)^{-1} = C^{-1}A^{-1}.$$

Proof. The proofs can be found in Sections A.1, and A.2. ■

Remark. As a consequence of the above theorem, we can easily show by induction that if $A_1, \ldots, A_k \in F^{n\times n}$ are invertible then $A_1 \cdots A_k$ is also invertible, and

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

***Example* 1.22.** Let $a, b, c, d \in F$. Consider the following 2×2 matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \; B = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Then it is easy to show by direct computation that

$$AB = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = BA.$$

Now suppose $ad - bc \neq 0$. Then the above equation implies that A is invertible, and

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

On the other hand, if $ad - bc = 0$ then we have $AB = 0$. Therefore A cannot be invertible, since otherwise we would have $B = IB = A^{-1}AB = A^{-1}0 = 0$. But this implies that $A = 0$, and hence $I = A^{-1}A = A^{-1}0 = 0$, which is a contradiction.

***Exercise* 1.23.** Suppose $A, B \in F^{n \times n}$ are diagonal matrices. Show that AB is also a diagonal matrix, and $(AB)_{jj} = A_{jj}B_{jj}$ for every $j \leq n$. Then conclude that $AB = BA$. In other words, conclude that diagonal matrices commute.

Solution. For $i \neq j$ we have

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj} = 0 \cdot 0 + \cdots + A_{ii} \cdot 0 + \cdots + 0 \cdot B_{jj} + \cdots + 0 \cdot 0 = 0.$$

And for $i = j$ we have

$$(AB)_{jj} = \sum_{k=1}^{n} A_{jk}B_{kj} = 0 \cdot 0 + \cdots + A_{jj}B_{jj} + \cdots + 0 \cdot 0 = A_{jj}B_{jj}.$$

Hence AB is diagonal. Thus BA is also diagonal. Furthermore for $i \neq j$ we have

$$(BA)_{jj} = B_{jj}A_{jj} = A_{jj}B_{jj} = (AB)_{jj}, \qquad (BA)_{ij} = 0 = (AB)_{ij}.$$

Therefore $AB = BA$ as desired. ■

Definition 1.24. Let F be a field. For $m \in \mathbb{N}$ we inductively define the **powers** of a square matrix $A \in F^{n \times n}$ to be

$$A^0 := I_n, \; A^1 := A, \; \ldots \; A^m := A^{m-1}A.$$

Also, for every polynomial

$$p(x) = a_0 + a_1 x + \cdots + a_m x^m$$

with coefficients $a_i \in F$, we define

$$p(A) := a_0 I_n + a_1 A + \cdots + a_m A^m.$$

We say that the matrix $p(A)$ is a *polynomial in* A.

Theorem 1.25. *Suppose F is a field, and $A, C \in F^{n\times n}$. Then for all nonnegative integers m, k we have*

(i) *If A commutes with C, then A^m commutes with C^k.*

(ii) *If A is invertible, then A^m is also invertible and*

$$(A^m)^{-1} = (A^{-1})^m.$$

(iii) $A^m A^k = A^{m+k}$.

(iv) $(A^m)^k = A^{mk}$.

(v) *If A, C commute, then we have $(AC)^m = A^m C^m$.*

(vi) *For any two polynomials p, q with coefficients in F we have*

$$(p+q)(A) = p(A) + q(A), \qquad (pq)(A) = p(A)q(A).$$

As a result, $p(A)$ and $q(A)$ always commute.

Remark. The significance of part (vi) is that the addition and multiplication of polynomials convert to the addition and multiplication of matrices via the map $p \mapsto p(A)$.

Proof. The proofs can be found in Sections A.1, and A.5. ■

Remark. As a consequence of the above theorem, we can easily show by induction that if $p_1, \ldots, p_k$ are polynomials with coefficients in F, then we have

$$\begin{aligned}(p_1 + \cdots + p_k)(A) &= p_1(A) + \cdots + p_k(A),\\ (p_1 p_2 \cdots p_k)(A) &= p_1(A)p_2(A)\cdots p_k(A).\end{aligned}$$

Definition 1.26. For a matrix $A \in \mathbb{C}^{m\times n}$ we define its **conjugate transpose** $A^* \in \mathbb{C}^{n\times m}$ by

$$(A^*)_{ij} := \overline{A_{ji}}.$$

Remark. Note that if $A \in \mathbb{R}^{m\times n}$ then $A^* = A^\intercal$.

Proposition 1.27. *Let $\lambda \in \mathbb{C}$, $A, B \in \mathbb{C}^{m\times n}$, and $C \in \mathbb{C}^{n\times l}$.*

(i) *We have*

$$\begin{aligned}(A+B)^* &= A^* + B^*, \qquad (\lambda A)^* = \bar{\lambda} A^*,\\ (AC)^* &= C^* A^*, \qquad (A^*)^* = A.\end{aligned}$$

(ii) *For every i, j we have*

$$(A_{i,.})^* = A^*_{.,i}, \qquad (A_{.,j})^* = A^*_{j,.}.$$

(iii) *If $A \in \mathbb{C}^{n\times n}$ is invertible, then A^* is also invertible, and we have*

$$(A^*)^{-1} = (A^{-1})^*.$$

Proof. **(i)** We have $\big((A^*)^*\big)_{ij} = \overline{(A^*)_{ji}} = \overline{\overline{A_{ij}}} = A_{ij}$. Also

$$\big((\lambda A)^*\big)_{ij} = \overline{(\lambda A)_{ji}} = \bar{\lambda}\overline{A_{ji}} = \bar{\lambda}(A^*)_{ij} = (\bar{\lambda}A^*)_{ij}.$$

In addition, we have

$$\begin{aligned}\big((AC)^*\big)_{ij} = \overline{(AC)_{ji}} &= \overline{\sum_{k\le n} A_{jk}C_{ki}} = \sum_{k\le n} \overline{A_{jk}}\,\overline{C_{ki}} \\ &= \sum_{k\le n} \overline{C_{ki}}\,\overline{A_{jk}} = \sum_{k\le n} (C^*)_{ik}(A^*)_{kj} = (C^*A^*)_{ij}.\end{aligned}$$

The other one is similar.

(ii) Note that $(A_{i,.})^*$ and $A^*_{.,i}$ have the same size. Now we have

$$\big((A_{i,.})^*\big)_{j,1} = \overline{(A_{i,.})_{1,j}} = \overline{A_{ij}} = A^*_{ji} = (A^*_{.,i})_{j,1}.$$

The other one is similar.

(iii) We have

$$(A^{-1})^*A^* = (AA^{-1})^* = I^* = I.$$

Similarly we have $A^*(A^{-1})^* = I$. Hence we get the desired result due to the uniqueness of the inverse of a matrix. ∎

Remark. As a consequence of the above proposition, we can easily show by induction that if $A_1, \ldots, A_k \in \mathbb{C}^{n\times n}$ then we have

$$(A_1 \cdots A_k)^* = A_k^* \cdots A_1^*.$$

1.4 Systems of Linear Equations

A *system of m linear equations in n unknowns*, or simply a *linear system*, is an expression of the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \quad\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m, \end{cases}$$

where a_{ij}, b_i are elements of a field F. The unknowns are also called *variables*. Let $A \in F^{m\times n}$ be the matrix whose ij-th entry is a_{ij}. Also let $b \in F^{m\times 1}$ be the column vector whose i-th component is b_i. Then we can write the system in the matrix form

$$Ax = b,$$

where x is the column vector whose j-th component is x_j. A **solution** of the system is a column vector $x \in F^{n\times 1}$ such that $Ax = b$. The goal is to find all the solutions of a system, or to show that no solution exists. The matrix A is called the **coefficient matrix** of the system. Let $[A|b]$ be the $m \times (n+1)$ matrix whose j-th column is $A_{\cdot,j}$ for $j \leq n$, and its $(n+1)$-th column is b. We call $[A|b]$ the **augmented matrix** of the system.

Remark. Formally, a system of linear equations can be defined to be the augmented matrix associated to it.

A system $[A|b]$ is called **homogeneous** if $b = 0$, and *nonhomogeneous* if $b \neq 0$. Note that $x = 0$ is always a solution of a homogeneous system $[A|0]$, since $A0 = 0$. The zero solution is called the *trivial* solution of the homogeneous system. A nonzero solution of a homogeneous system is called a *nontrivial* solution.

A system of linear equations is called *consistent* if it has at least one solution, and is called *inconsistent* if it has no solution. Two systems of m linear equations in n unknowns are called **equivalent** if they have the same set of solutions. Note that the set of solutions of a system can be empty too.

Remark. It is easy to see that the equivalence of systems of linear equations is an equivalence relation, i.e. any system is equivalent to itself, and if $[A|b]$ is equivalent to $[A'|b']$ then $[A'|b']$ is equivalent to $[A|b]$ too. Also, if $[A|b]$ is equivalent to $[A'|b']$, and $[A'|b']$ is equivalent to $[A''|b'']$, then $[A|b]$ is equivalent to $[A''|b'']$ too.

Our strategy for solving a system of linear equations is to change the system into an equivalent system that is easier to solve. Then we repeat this process until we get a system which is equivalent to our original system, and can be solved with a few simple calculations. The tools that we use to accomplish this are the operations defined below.

Definition 1.28. An **elementary row operation** is an operation performed on a matrix, that is of one of the following three types

(i) Interchanging two rows of the matrix.
(ii) Multiplying a row by a nonzero constant.
(iii) Adding a multiple of one row to another row.

An **elementary matrix** is a square matrix obtained from the identity matrix by applying one elementary row operation. The type of an elementary matrix is the type of the elementary row operation that produced the matrix.

Notation. Suppose $a, c \in F$, and $a \neq 0$. The elementary matrices have the following forms

$$\begin{matrix} \\ i\text{-th row} \to \\ \vdots \\ j\text{-th row} \to \end{matrix} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & a & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & c & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

So, the first matrix is obtained by interchanging the i-th row and the j-th row of the identity matrix; the second matrix is obtained by multiplying the i-th row of the identity matrix by the nonzero constant a; and the third matrix is obtained by adding the j-th row of the identity matrix multiplied by c, to its i-th row. We denote the first elementary matrix by $E(i, j)$, the second elementary matrix by $E(i; a)$, and the third elementary matrix by $E(i, j; c)$. Note that we always assume $i \neq j$. Also note that the identity matrix I is also an elementary matrix, since for example $I = E(i; 1)$ for any i.

Remark. Let E be one of the above elementary matrices. Then for $k \neq i, j$ we have

$$E_{k,\cdot} = I_{k,\cdot} = e_k^\intercal,$$

i.e. the k-th row of E is the k-th vector of the standard basis. We also have

$$\begin{aligned} &(E(i,j))_{i,\cdot} = e_j^\intercal, \qquad (E(i;a))_{i,\cdot} = ae_i^\intercal, \qquad (E(i,j;c))_{i,\cdot} = e_i^\intercal + ce_j^\intercal. \\ &(E(i,j))_{j,\cdot} = e_i^\intercal, \qquad (E(i;a))_{j,\cdot} = e_j^\intercal, \qquad (E(i,j;c))_{j,\cdot} = e_j^\intercal. \end{aligned}$$

Proposition 1.29. *Let $A \in F^{m\times n}$, and let B be the matrix obtained from A by applying one elementary row operation. Then $B = EA$, where E is the elementary matrix corresponding to that elementary row operation.*

Proof. Suppose B is obtained from A by adding its j-th row multiplied by c, to its i-th row. The other two cases are similar. For $k \neq i$ we have $B_{k,\cdot} = A_{k,\cdot}$. We also have $B_{i,\cdot} = A_{i,\cdot} + cA_{j,\cdot}$. Now let $E = E(i, j; c) \in F^{m\times m}$. Then $EA \in F^{m\times n}$. Furthermore we know that $(EA)_{k,\cdot} = E_{k,\cdot}A$ for any $k \leq m$. When $k \neq i$ we have

$$(EA)_{k,\cdot} = E_{k,\cdot}A = e_k^\intercal A = A_{k,\cdot} = B_{k,\cdot}.$$

And when $k = i$ we have

$$(EA)_{i,\cdot} = E_{i,\cdot}A = (e_i^\intercal + ce_j^\intercal)A = e_i^\intercal A + ce_j^\intercal A = A_{i,\cdot} + cA_{j,\cdot} = B_{i,\cdot}.$$

Therefore we must have $EA = B$. ∎

Proposition 1.30. *Let $E \in F^{m\times m}$ be an elementary matrix. Then E is invertible, and its inverse is also an elementary matrix that has the same type as E. In fact we have*

$$E(i,j)^{-1} = E(i,j), \qquad E(i;a)^{-1} = E(i;a^{-1}), \qquad E(i,j;c)^{-1} = E(i,j;-c),$$

where $a, c \in F$ and $a \neq 0$.

Remark. A trivial consequence of the above two propositions is that an elementary row operation is invertible, and its inverse has the same type as itself.

Proof. We will only prove the third identity; the other two are similar. We have to show that

$$E(i,j;c)E(i,j;-c) = I = E(i,j;-c)E(i,j;c).$$

Let $E = E(i,j;c)$, $E' = E(i,j;-c)$, and $A = EE'$. Then we have $A_{kl} = E_{k,\cdot}E'_{\cdot,l}$. When $k \neq i$ we have $E_{k,\cdot} = e_k^\intercal = E'_{k,\cdot}$. Hence

$$A_{kl} = e_k^\intercal E'_{\cdot,l} = (E'_{\cdot,l})_{k,\cdot} = E'_{k,l} = \begin{cases} 1 & l = k, \\ 0 & l \neq k. \end{cases}$$

When $k = i$ we have $E_{i,\cdot} = e_i^\intercal + ce_j^\intercal$ and $E'_{i,\cdot} = e_i^\intercal - ce_j^\intercal$. Thus

$$\begin{aligned} A_{il} &= (e_i^\intercal + ce_j^\intercal)E'_{\cdot,l} = e_i^\intercal E'_{\cdot,l} + ce_j^\intercal E'_{\cdot,l} \\ &= (E'_{\cdot,l})_{i,\cdot} + c(E'_{\cdot,l})_{j,\cdot} = E'_{i,l} + cE'_{j,l} = \begin{cases} 1 + c\cdot 0 = 1 & l = i \neq j, \\ -c + c\cdot 1 = 0 & l = j \neq i, \\ 0 + c\cdot 0 = 0 & l \neq i, j. \end{cases} \end{aligned}$$

Therefore $EE' = A = I$. We can similarly show that $E'E = I$. Hence E is invertible, and $E^{-1} = E'$ as desired. ∎

Proposition 1.31. *Suppose $A \in F^{m\times n}$, $E \in F^{m\times m}$, and $b \in F^{m\times 1}$. If E is invertible then the linear system $[EA|Eb]$ is equivalent to the linear system $[A|b]$.*

Proof. Suppose $x \in F^{n\times 1}$ is a solution of $[A|b]$. Then $Ax = b$. Hence $(EA)x = E(Ax) = Eb$, i.e. x is also a solution of $[EA|Eb]$. Conversely if x is a solution of $[EA|Eb]$ then $EAx = Eb$. Thus

$$Ax = IAx = E^{-1}E(Ax) = E^{-1}(EAx) = E^{-1}Eb = Ib = b,$$

so x is a solution of $[A|b]$ too. Note that the above calculations also show that if one of the systems has no solution, then the other system cannot have a solution either. Therefore we have shown that the two systems have the same set of solutions, i.e. they are equivalent. ∎

Note that in particular when E is an elementary matrix, $[EA|Eb]$ is equivalent to $[A|b]$, since elementary matrices are invertible. Hence if we apply an elementary row operation on a linear system we obtain an equivalent system. Because applying an elementary row operation on a linear system, is the same as multiplying the system by the elementary matrix associated to the elementary row operation.

Now our strategy is to perform a sequence of elementary row operations on the augmented matrix of a linear system, in order to obtain an equivalent system that can be solved easily. The final form of the augmented matrix that we wish to obtain, is described below.

Definition 1.32. Let $A \in F^{m\times n}$. We say A is in **reduced row echelon form** if it satisfies the following conditions

(i) Every nonzero row of A is above every zero row of A, i.e. if $A_{i,\cdot} \neq 0$ and $A_{j,\cdot} = 0$, then $i < j$.

(ii) The first nonzero entry of a nonzero row is 1, i.e. if $A_{ij} \neq 0$ and $A_{il} = 0$ for every $l < j$, then $A_{ij} = 1$. In this case, A_{ij} is called the **leading entry** of the row $A_{i,\cdot}$.

(iii) A leading entry is the only nonzero entry in its column, i.e. if $A_{ij} = 1$ is the leading entry of the i-th row, then $A_{kj} = 0$ for every $k \neq i$.

(iv) The leading entry of a nonzero row, is in a column to the right of the column containing the leading entry of any row above it. In other words, if A_{ij} and A_{kl} are the leading entries of the rows $A_{i,\cdot}$ and $A_{k,\cdot}$ respectively, then $i < k$ implies $j < l$.

Remark. The number of leading entries of A is obviously less than or equal to the number of rows of A. Also note that by condition 3, the number of leading entries of A cannot be more than the number of columns of A either, because any column can contain at most one leading entry.

***Example* 1.33.** The following matrix is in reduced row echelon form

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The leading entries are highlighted in bold. Note that the 1 in the last column is not a leading entry. The zero matrix is also a trivial example of a matrix in reduced row echelon form.

Proposition 1.34. *Suppose $A \in F^{m\times n}$ is in reduced row echelon form. Let $A_{1,j_1}, \dots A_{k,j_k}$ be the leading entries of the nonzero rows of A. Then we have $k \leq \min\{m, n\}$, and*

$$A_{\cdot,j_i} = e_i,$$

for every $i \leq k$. In addition if $j < j_i$ then

$$A_{.,j} = [A_{1j}, A_{2j}, \ldots, A_{i-1,j}, 0, \ldots, 0]^{\intercal}.$$

In other words, $A_{lj} = 0$ if $l \geq i$. And if $j \geq j_k$ then $A_{lj} = 0$ for $l > k$.

Remark. Note that when $j < j_1$ then we have $A_{.,j} = 0$.

Proof. It is obvious that $k \leq m$. We also have $k \leq n$, since $j_1 < j_2 < \cdots < j_k \leq n$. Now note that the column $A_{.,j_i}$ has only one nonzero component, which is its i-th component; and this nonzero component is 1. Hence we have $A_{.,j_i} = e_i$.

Furthermore if $j < j_i$, then the nonzero entries of the column $A_{.,j}$ are on the rows $A_{1,.}, \ldots, A_{i-1,.}$. In other words, $A_{lj} = 0$ for $l \geq i$. Because otherwise for some $l \geq i$ the leading entry of the l-th row is A_{lj} where $j < j_i$, i.e. the leading entry of the l-th row is in a column to the left of the column containing the leading entry of the i-th row which lies above the l-th row, or is equal to the l-th row. But this is in contradiction with the 4th condition of a matrix in reduced row echelon form, or simply with the fact that a row cannot have two leading entries. Finally note that when $j > j_k$ we have $A_{lj} = 0$ for $l > k$, because the rows of A below the k-th row are zero. ■

Theorem 1.35. *Suppose $B = [A|b] \in F^{m\times(n+1)}$ is the augmented matrix of a system of linear equations. Also suppose that B is in reduced row echelon form. Let $B_{1,j_1}, \ldots B_{k,j_k}$ be the leading entries of the nonzero rows of B. If $j_k = n+1$ then the linear system $[A|b]$ has no solution. Otherwise, let $B_{.,l_1}, \ldots, B_{.,l_{n-k}}, B_{.,n+1}$ be the columns of B that do not contain a leading entry. Then the set of solutions of the linear system $[A|b]$ is the set of vectors of the form*

$$v_0 + x_{l_1} v_{l_1} + \cdots + x_{l_{n-k}} v_{l_{n-k}},$$

where for $p \leq n-k$, $x_{l_p} \in F$ is arbitrary, and $v_0, v_{l_p} \in F^{n\times 1}$ are given by

$$(v_0)_i = \begin{cases} B_{q,n+1} & i = j_q, q \leq k, \\ \\ 0 & i = l_{\tilde{p}}, \tilde{p} \leq n-k, \end{cases} \qquad (v_{l_p})_i = \begin{cases} -B_{q,l_p} & i = j_q, q \leq k, \\ 1 & i = l_p, \\ 0 & i = l_{\tilde{p}}, \tilde{p} \neq p, \tilde{p} \leq n-k. \end{cases}$$

Remark. Note that we require the augmented matrix of the system to be in reduced row echelon form, not its coefficient matrix. Also note that when the system is homogeneous we have $v_0 = 0$, since $B_{.,n+1} = b = 0$.

Remark. The variables $x_{l_1}, \ldots, x_{l_{n-k}}$ are called the **free variables** of the system. Also, the expression

$$v_0 + x_{l_1} v_{l_1} + \cdots + x_{l_{n-k}} v_{l_{n-k}},$$

is sometimes called the *general solution* of the system.

Proof. If $j_k = n+1$ then the k-th row of B is $[0\,0\,\ldots\,0\,1]$. The equation corresponding to this row is $0x_1+\cdots+0x_n = 1$, which obviously has no solution. So let us assume that $j_k \leq n$. Thus in particular we have $k \leq n$, since $j_1 < j_2 < \cdots < j_k \leq n$. We also know that the first k rows of B are nonzero, and the rest are zero. The corresponding equation to a zero row is $0x_1 + \cdots + 0x_n = 0$, which is trivially satisfied for all $x_j \in F$. Thus we have to find the solutions to the equations corresponding to the nonzero rows. Consider the q-th row where $q \leq k$. Then we have

$$0x_1 + \cdots + 0x_{j_q-1} + 1x_{j_q} + \sum_{j>j_q} B_{qj}x_j = B_{q,n+1}.$$

But note that $B_{qj} = 0$ when $j = j_{\tilde{q}}$ for $\tilde{q} \neq q$, and $B_{q,l_p} = 0$ when $l_p < j_q$. Hence the above equation is equivalent to

$$x_{j_q} = B_{q,n+1} - \sum_{l_p>j_q} B_{q,l_p}x_{l_p} = B_{q,n+1} - \sum_{p\geq 1} B_{q,l_p}x_{l_p}. \qquad (*)$$

Therefore if we assign arbitrary values to the free variables $x_{l_1}, \ldots, x_{l_{n-k}}$, then the vector

$$y := v_0 + x_{l_1}v_{l_1} + \cdots + x_{l_{n-k}}v_{l_{n-k}}$$

is a solution of the system, since it is easy to see that $y_{l_p} = x_{l_p}$, and

$$y_{j_q} = B_{q,n+1} - \sum_{p\geq 1} B_{q,l_p}x_{l_p} = B_{q,n+1} - \sum_{p\geq 1} B_{q,l_p}y_{l_p}.$$

Conversely if $x \in F^{n\times 1}$ is a solution of the system, then the relation $(*)$ must hold for every $q \leq k$; and there is no restriction on the l_p-th component of x for every $p \leq n-k$. Now it is easy to see that

$$x = v_0 + x_{l_1}v_{l_1} + \cdots + x_{l_{n-k}}v_{l_{n-k}},$$

since the equality holds for every component. Hence the set of solutions of the system is exactly the set of vectors described in the theorem. ∎

Remark. Consider a consistent linear system whose augmented matrix is in reduced row echelon form. It is clear from the previous theorem that this system has a unique solution if and only if it has no free variables.

***Example* 1.36.** Although the concrete description of the solutions of a system given in the previous theorem has some theoretical applications, the best way to apply that result in numerical examples is to imitate its proof. For example consider the following augmented matrix of a linear system

$$\left[\begin{array}{cccccccc|c} 0 & \mathbf{1} & 3 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

The free variables of this system are x_1, x_3, x_4, x_6, x_8. We also have

$$\begin{aligned} x_2 &= 3 - 3x_3 - x_6, \\ x_5 &= 2 + 2x_6, \\ x_7 &= 9 - x_8. \end{aligned}$$

Now to produce the vector v_0 we make every free variable zero. Also to produce a vector v_{l_p} we put $x_{l_p} = 1$, and $x_{l_q} = 0$ for $q \neq p$, and we ignore the constant terms i.e. we consider v_0 to be 0. Hence the general solution of the above system is

$$\begin{bmatrix} 0 \\ 3 \\ 0 \\ 0 \\ 2 \\ 0 \\ 9 \\ 0 \end{bmatrix} + x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ -3 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_6 \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_8 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{bmatrix},$$

where x_1, x_3, x_4, x_6, x_8 can have any values. ■

1.5 Gaussian Elimination

Remember that our strategy to solve a system of linear equations is to perform a sequence of elementary row operations on the augmented matrix of the system, in order to obtain the equivalent system whose augmented matrix is in reduced row echelon form. Then we can easily solve the system as we have shown above. There are several algorithms to convert a matrix to its reduced row echelon form. Here we describe the method called **Gaussian elimination**. We describe this algorithm step by step, and simultaneously we apply each step to a given matrix, to clarify the method.

(i) Let $A \in F^{m \times n}$ be a given matrix. Consider the first nonzero column of A from the left, i.e. the nonzero column $A_{.,j}$ with smallest j. Let A_{ij} be the first nonzero entry of the column $A_{.,j}$ from the top, i.e. the nonzero entry A_{ij} with smallest i. Then interchange the first row of A with its i-th row. In the resulting matrix, the $1, j$-th entry is nonzero. Next multiply the first row by the inverse of this nonzero entry. Then in the resulting matrix, the $1, j$-th entry is one. This entry is called a **leading entry**. In the following example,

the leading entries are highlighted in bold.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 2 & 6 & 0 & 8 & -14 & -2 & -2 \\ 0 & 1 & 3 & 0 & 5 & -9 & 3 & 3 \\ 0 & 3 & 9 & 0 & 11 & -19 & -4 & -4 \end{bmatrix} \to \begin{bmatrix} 0 & 2 & 6 & 0 & 8 & -14 & -2 & -2 \\ 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 1 & 3 & 0 & 5 & -9 & 3 & 3 \\ 0 & 3 & 9 & 0 & 11 & -19 & -4 & -4 \end{bmatrix}$$
$$\to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 1 & 3 & 0 & 5 & -9 & 3 & 3 \\ 0 & 3 & 9 & 0 & 11 & -19 & -4 & -4 \end{bmatrix}.$$

(ii) Now add suitable multiples of the first row to the other rows, in order to make the entries below the leading entry and in its column, zero.

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 1 & 3 & 0 & 5 & -9 & 3 & 3 \\ 0 & 3 & 9 & 0 & 11 & -19 & -4 & -4 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 0 & 0 & 0 & 1 & -2 & 4 & 4 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & -1 \end{bmatrix}.$$

(iii) Then ignore the first row, and apply the previous steps to the rest of the matrix, which is a matrix in $F^{(m-1)\times n}$. In the following example, the ignored rows are highlighted in gray.

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -3 & -3 \\ 0 & 0 & 0 & 0 & 1 & -2 & 4 & 4 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & -1 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 1 & -2 & 4 & 4 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & -1 \end{bmatrix}$$
$$\to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 \end{bmatrix}.$$

Then ignore the first two rows, and apply the previous steps to the rest of the matrix, which is a matrix in $F^{(m-2)\times n}$. Repeat this process by ignoring more rows, until there is no row left, or the remaining rows are all zero.

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

(iv) Now consider the last leading entry, i.e. the leading entry in the last nonzero row. Then add suitable multiples of this row to the other rows, in order to

make the entries above the leading entry and in its column, zero.

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & -1 & -1 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Repeat this process with the leading entry of the row above the last nonzero row, and then with the leading entries of the rows above it, until you reach the leading entry of the first row.

$$\begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 4 & -7 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Remark. The steps 1-3 are called the **forward phase** of the algorithm, and the step 4 is called the **backward phase**.

Remark. If instead of step 4, in steps 2 and 3 we make all the entries in the columns of leading entries zero, we obtain the *Gauss-Jordan elimination.* This method can be used to solve linear systems too, but it requires more arithmetic operations compared to Gaussian elimination. Because when in this algorithm we make the entries above a leading entry zero, we have to multiply and add some entries that would have been zero in the backward phase of the Gaussian elimination. So the Gaussian elimination is more efficient, and hence it is the preferred computational method.

Remark. Let us describe the Gaussian elimination in a different way. This description allows us to look at this algorithm as a function on the space of matrices $F^{m\times n}$. Let $S_i : F^{m\times n} \to F^{m\times n}$ be the function that switches the i-th row of a matrix A with the first row in the sequence $A_{i,.}, A_{i+1,.}, \ldots, A_{m,.}$ that has the leftmost nonzero entry. If all the rows $A_{i,.}, A_{i+1,.}, \ldots, A_{m,.}$ are zero, then S_i does not change A. Let $R_i : F^{m\times n} \to F^{m\times n}$ be the function that multiplies the i-th row of a matrix A with the inverse of the first nonzero entry of its i-th row. If the i-th row of A is zero, then R_i does not change A.

Finally let $U_i : F^{m\times n} \to F^{m\times n}$ be the function that makes every entry above the first nonzero entry of the i-th row, zero; and let $D_i : F^{m\times n} \to F^{m\times n}$ be the function that makes every entry below the first nonzero entry of the i-th row, zero. Both U_i, D_i add suitable multiples of the i-th row to other rows to produce their output. And if the i-th row of the matrix is zero, then U_i, D_i do not change it. Now it is easy to see that the application of the Gaussian elimination to a matrix

is the same as the action of the following function

$$[U_2 \circ \cdots \circ U_{m-1} \circ U_m] \circ R_m \circ [D_{m-1} \circ R_{m-1} \circ S_{m-1}] \circ \cdots \\ \cdots \circ [D_2 \circ R_2 \circ S_2] \circ [D_1 \circ R_1 \circ S_1].$$

Note that the leftmost group of functions in the above formula represents the backward phase of the algorithm, and the other groups represent the part of the forward phase that creates and employs the i-th leading entry (if there is any). Similarly, it is easy to see that the Gauss-Jordan elimination has the same output as the following function

$$[U_m \circ R_m] \circ [U_{m-1} \circ D_{m-1} \circ R_{m-1} \circ S_{m-1}] \circ \cdots \\ \cdots \circ [U_2 \circ D_2 \circ R_2 \circ S_2] \circ [D_1 \circ R_1 \circ S_1].$$

Definition 1.37. Let $A, B \in F^{m\times n}$. We say B is the **reduced row echelon form** of A if B is in reduced row echelon form, and there is an invertible matrix $E \in F^{m\times m}$ such that $B = EA$.

Remark. Every matrix A has a unique reduced row echelon form B. The existence of B, through an algorithm to find it, is proved in the next theorem. We will prove the uniqueness of B in Theorem 2.46, after we develop the appropriate tools. Note that the uniqueness of the reduced row echelon form means that if B, B' are two matrices in reduced row echelon form, and E, E' are invertible matrices such that $B = EA$, $B' = E'A$, then $B = B'$. But it is not true that E is also uniquely determined by A. In fact different algorithms for finding B, produce different invertible matrices E.

Theorem 1.38. *Gaussian elimination converts every matrix to its reduced row echelon form.*

Remark. In fact we will show that for $A \in F^{m\times n}$, Gaussian elimination produces a finite sequence of elementary matrices $E_1, \ldots, E_k \in F^{m\times m}$ such that $E_k \cdots E_1 A$ is the reduced row echelon form of A.

Proof. Let $A \in F^{m\times n}$. Note that each operation in the Gaussian elimination is an elementary row operation, which corresponds to multiplication from the left by an elementary matrix. Therefore if $B \in F^{m\times n}$ is the output of the Gaussian elimination, then $B = EA$, where $E \in F^{m\times m}$ is the product of finitely many elementary matrices. But the elementary matrices are invertible, hence their product is invertible too, i.e. E is invertible. Therefore it suffices to show that B is in reduced row echelon form. We prove this by induction on the number of rows m. When $m = 1$, the Gaussian elimination converts the matrix A, which is a row vector, to a row

vector whose first nonzero entry is one; or to the zero row vector when A is itself zero. Thus the Gaussian elimination converts A to a reduced row echelon matrix.

Now suppose the claim is true for $m-1$, and we want to prove it for m. Let $\tilde{A} \in F^{m\times n}$ be the matrix that is obtained from A by applying the first and second steps of Gaussian elimination. Let $A' \in F^{(m-1)\times n}$ be the matrix that is obtained from $\tilde{A}$ by removing the first row, i.e. $A'_{i,\cdot} = \tilde{A}_{i+1,\cdot}$ for $i \leq m-1$. Note that when we apply the third step of the algorithm to $\tilde{A}$, we are actually applying the forward phase of the Gaussian elimination to A'. Also when we apply the backward phase to $\tilde{A}$, we are applying the backward phase to A', and in addition we add some multiples of nonzero rows of A' to the first row of $\tilde{A}$ to make some of its entries zero. Thus in the process of applying the Gaussian elimination to A, we have applied the Gaussian elimination to A' too. Note that in the Gaussian elimination, other than its first step, we do not interchange any row with the first row. So if we let $B' \in F^{(m-1)\times n}$ be the matrix that is obtained from B by removing the first row, then B' is the output of the Gaussian elimination applied to A'. Therefore by the induction hypothesis B' is in reduced row echelon form.

Hence we only need to conclude that B is also in reduced row echelon form. We have to check the four conditions of a matrix in reduced row echelon form. If A is zero, then B is zero too, and it is in reduced row echelon form trivially. Otherwise A has at least a nonzero row, so the first row of $\tilde{A}$ is nonzero. The first row of B is obtained from the first row of $\tilde{A}$ by adding to it some multiples of the rows below it. But when we formed $\tilde{A}$, we made the entries below the leading entry of the first row, zero. So the leading entry of the first row will not change when we add a multiple of another row to the first row. In other words, the first row of B is nonzero. Thus as the zero rows of B' are below every nonzero row, B has this property too. Furthermore, the leading entry of every nonzero row of B is one, since this is true in B'; and we made the leading entry of the first row of B one, when we applied the first step of the algorithm to A.

Now consider a leading entry in B. If it is the leading entry of the first row, then we have made every other entry in its column zero, when we applied the second step of the algorithm. If the leading entry belongs to B', then every other entry in its column is zero by induction hypothesis. But we have to note that this column is in B', i.e. we cannot conclude anything from the induction hypothesis about the entry on this column and the first row of B. However we made these entries zero during the application of the backward phase of the algorithm. So the leading entries of B are the only nonzero entry in their columns.

Finally it remains to show that the leading entry of each row of B is to the right of the leading entry of any row above it. If the two rows belong to B', then this is true due to the induction hypothesis. Thus we only need to show that the leading entry of the first row of B is to the left of the leading entry of any other row. But when we applied the first step of the algorithm to A, we chose the first leading

entry in the first nonzero column, and we interchanged its row with the first row. Then in the second step we made all the entries in the column of the first leading entry zero. Hence in B' the first nonzero column is to the right of the column that contains the leading entry of the first row of B. Therefore the leading entries of B' are all to the right of the leading entry of the first row of B, as desired. ■

Remark. The above theorem is in particular true when we apply the Gaussian elimination to the augmented matrix of a linear system. Then we convert the system through a sequence of elementary row operations, to a system whose augmented matrix is in reduced row echelon form. And we have seen before that we can easily solve such systems. At the end we have found the set of solutions of the initial system, since the elementary row operations do not change the set of solutions of a system. Note that due to this last property, we do not need to use the uniqueness of the reduced row echelon form of the augmented matrix, in order to be certain that we will not get a different set of solutions if we use a different algorithm. Because if there were more than one reduced row echelon form for the augmented matrix, then all of them would have produced equivalent linear systems, with the same set of solutions, due to Proposition 1.31, since they are all of the form EA, where A is the augmented matrix of the system, and E is an invertible matrix.

Chapter 2

Vector Spaces

2.1 Vector Spaces

Definition 2.1. A **vector space** over a field F is a nonempty set V equipped with two operations

$$\begin{array}{l} V \times V \longrightarrow V \\ (v, w) \mapsto v + w \end{array} , \qquad \begin{array}{l} F \times V \longrightarrow V \\ (a, v) \mapsto av \end{array} ,$$

called respectively **(vector) addition** and **scalar multiplication**, such that

(i) Addition is *associative*, i.e. for every $v, w, u \in V$ we have

$$(v + w) + u = v + (w + u).$$

(ii) Addition is *commutative*, i.e. for every $v, w \in V$ we have

$$v + w = w + v.$$

(iii) V has an additive identity, i.e. there is an element $0 \in V$ such that for every $v \in V$ we have

$$v + 0 = v.$$

(iv) Every $v \in V$ has an **additive inverse**, i.e. there exists $w \in V$ such that

$$v + w = 0.$$

(v) For every $a \in F$ and every $v, w \in V$ we have

$$a(v + w) = av + aw.$$

(vi) For every $a, b \in F$ and every $v \in V$ we have

$$(a + b)v = av + bv.$$

(vii) For every $a, b \in F$ and every $v \in V$ we have

$$a(bv) = (ab)v.$$

(viii) For every $v \in V$ we have

$$1v = v,$$

where 1 is the identity of F.

Remark. The elements of the vector space V are called **vectors**. Note that the nature of the elements of V can be quite different from our intuitive notion of vectors in three dimensional Euclidean space. For example the elements of a vector space can be functions, or sequences, or more complicated objects. Nevertheless, we will refer to the elements of any vector space as vectors. In contrast, the elements of the field F are usually referred to as **scalars**.

Remark. The properties (v) and (vi) mean that scalar multiplication is *distributive* over vector addition and addition of scalars.

Remark. As we will show below, the additive identity of a vector space is unique, and will be called the **zero vector**. Also, the additive inverse of any vector v is unique, and will be denoted by $-v$. Note that due to the commutativity we also have

$$0 + v = v, \qquad (-v) + v = 0.$$

In addition, for two vectors v, w we define

$$w - v := w + (-v).$$

Remark. Let V be a vector space over a field F. When $F = \mathbb{R}$ we say V is a *real vector space*, and when $F = \mathbb{C}$ we say V is a *complex vector space*.

***Example* 2.2.** Suppose F is a field. Then F^n equipped with the standard componentwise addition and scalar multiplication, is a vector space over F. Also, $F^{m\times n}$ equipped with the addition and scalar multiplication of matrices, is a vector space over F. Recall that we can consider the elements of F^n as column vectors, i.e. as $n \times 1$ matrices, via the identification

$$(a_1, \ldots, a_n) \mapsto \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.$$

We will usually use this convention unless otherwise specified. Note that this identification preserves both addition and scalar multiplication. Also, remember that we formally defined F^n and $F^{m\times n}$ to be respectively the set of functions from $\{1, \ldots, n\}$ and $\{(i, j) : i \leq m,\, j \leq n\}$ into F. Therefore F^n and $F^{m\times n}$ are special cases of the vector spaces described in the next example.

***Example* 2.3.** Let $\mathcal{S}$ be a nonempty set, and let F be a field. Then the space of functions from $\mathcal{S}$ into F, i.e.

$$F^{\mathcal{S}} := \{f : \text{for every function } f : \mathcal{S} \to F\},$$

is a vector space over F. The addition and scalar multiplication on this space are defined as follows

$$(f+g)(s) := f(s) + g(s), \qquad (af)(s) := af(s),$$

where $f, g \in F^{\mathcal{S}}$ and $a \in F$. Note that $f+g$ and af are functions, so in order to define them we have to specify their values at every $s \in \mathcal{S}$. We leave it as an exercise, to check that $F^{\mathcal{S}}$ is indeed a vector space with these operations. We only mention that the zero of this vector space is the zero function, i.e. the function that maps every $s \in \mathcal{S}$ to $0 \in F$. Also, the additive inverse of a function f is the function $(-f)(s) := -f(s)$.

***Example* 2.4.** Another special case of the above example, is when $\mathcal{S} = \mathbb{N}$. Then $F^{\mathbb{N}}$ is the space of all sequences in F. In this case, the above operations are actually the componentwise addition and scalar multiplication of sequences, and they will make $F^{\mathbb{N}}$ a vector space over F. We will also denote this vector space by F^{∞}.

***Example* 2.5.** Let $F[x]$ be the space of polynomials with coefficients in a field F. Then $F[x]$ equipped with the standard addition of polynomials and multiplying polynomials by a constant, is a vector space over F. For the details see Section A.3.

Proposition 2.6. *Let V be a vector space over a field F. Then for every $a \in F$ and every $v, w, u \in V$ we have*

(i) **(Cancellation Law)** *$v + w = u + w$ implies $v = u$.*
(ii) *V has a unique additive identity.*
(iii) *Every vector v has a unique additive inverse.*
(iv) $0v = 0$.
(v) $a0 = 0$.
(vi) *$av = 0$ implies $a = 0$ or $v = 0$.*
(vii) $(-1)v = -v$.

Remark. Note that we use 0 to denote both the zero vector, and the zero of F. But it should always be clear from the context which one is intended.

Proof. **(i)** Suppose y is an additive inverse of w. Then we can add y to both sides of $v + w = u + w$ to obtain $(v + w) + y = (u + w) + y$. Now by associativity of addition we have $v + (w + y) = u + (w + y)$. Since $w + y = 0$, we get $v + 0 = u + 0$. Therefore $w = u$.

(ii) Suppose $0, \tilde{0}$ are both additive identities of V. Then we have $\tilde{0} = \tilde{0}+0$, since 0 is an additive identity. We also have $0 + \tilde{0} = 0$, since $\tilde{0}$ is an additive identity. However we know that $\tilde{0} + 0 = 0 + \tilde{0}$, because addition is commutative. Therefore we must have $\tilde{0} = 0$, as desired.

(iii) Suppose $w, \tilde{w}$ are both additive inverses of v. Then

$$w + v = 0 = \tilde{w} + v.$$

Thus by cancellation law we get $w = \tilde{w}$.

(iv) We have

$$0 + 0v = 0v = (0 + 0)v = 0v + 0v.$$

Hence by cancellation law we get $0 = 0v$.

(v) We have

$$0 + a0 = a0 = a(0 + 0) = a0 + a0.$$

Thus again by cancellation law we get $0 = a0$.

(vi) If $a = 0$ then there is nothing to prove. Now if $a \neq 0$ then we have

$$v = 1v = (a^{-1}a)v = a^{-1}(av) = a^{-1}0 = 0.$$

(vii) We have

$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = 0.$$

Hence the result follows from the uniqueness of the additive inverse of v. ■

Remark. It is easy to show by induction that

$$a(v_1 + \cdots + v_k) = av_1 + \cdots + av_k,$$
$$(a_1 + \cdots + a_k)v = a_1v + \cdots + a_kv,$$

for $v, v_i \in V$, and $a, a_i \in F$.

2.2 Subspaces and Linear Combinations

Notation. In the rest of this chapter, we assume that F is a field, and V is a vector space over F.

Definition 2.7. Let $W \subset V$. Then we can restrict the vector addition and scalar multiplication of V to W. If with these restricted operations, W becomes a vector space, then we say W is a **subspace** of V.

Remark. The vector addition and scalar multiplication of V are functions as follows

$$+ : V \times V \to V, \qquad \cdot : F \times V \to V.$$

So when we talk about their restrictions to W, we mean

$$+|_{W\times W} : W \times W \to V, \qquad \cdot|_{F\times W} : F \times W \to V.$$

Note that we do not know a priori that the image of the above restricted functions are inside W. But this is actually true as we will prove below.

Theorem 2.8. *Let $W \subset V$. Then W is a subspace of V if and only if all of the following three conditions hold*

(i) $0 \in W$, *where* 0 *is the zero vector of* V.
(ii) W *is closed under addition, i.e. if* $u, v \in W$ *then* $u + v \in W$.
(iii) W *is closed under scalar multiplication, i.e. if* $a \in F$ *and* $u \in W$ *then* $au \in W$.

Proof. First suppose W satisfies the above three conditions. Then W is nonempty. In addition, since W is closed under addition and scalar multiplication, we have

$$+|_{W\times W} : W \times W \to W, \qquad \cdot|_{F\times W} : F \times W \to W.$$

Hence W is equipped with a vector addition and a scalar multiplication. It is obvious that these operations satisfy the properties (i),(ii), and (v)–(viii) of the definition of a vector space; because these properties hold for all the vectors in V, so they also hold for all the vectors in W. Thus we only need to show that the properties (iii),(iv) of the definition of a vector space also hold in W, i.e. we have to show that W has an additive identity, and every $w \in W$ has an additive inverse in W. Since the zero vector of V belongs to W, it is obvious that W has an additive identity. On the other hand, every $w \in W$ has an additive inverse in W, because W is closed under scalar multiplication, and therefore we have $-w = (-1)w \in W$. Thus the above restricted operations make W a vector space. Hence W is a subspace of V.

Now suppose W is a subspace of V. Then W is a vector space with the restricted operations

$$+|_{W\times W} : W \times W \to V, \qquad \cdot|_{F\times W} : F \times W \to V.$$

But the operations of a vector space must take values inside the vector space, i.e. the image of the above functions must be W. Therefore W must be closed under both vector addition and scalar multiplication. On the other hand, W must be nonempty, since it is a vector space. Let $w \in W$. Then we have $0 = 0w \in W$, since W is closed under scalar multiplication. Thus W satisfies the three conditions stated in the theorem. ∎

Remark. The following proposition is an easy criterion to check whether a given subset is a subspace.

Proposition 2.9. *Suppose $W \subset V$ is nonempty. Then W is a subspace of V if and only if for all $u, v \in W$ and $a \in F$ we have $u + av \in W$.*

Proof. If W is a subspace then we have $av \in W$, since W is closed under scalar multiplication. Now we have $u+av \in W$ because W is closed under vector addition. Conversely suppose that W satisfies the property stated in the proposition. Then if we set $a = 1$ we get $u + v \in W$ for every $u, v \in W$. Hence W is closed under addition. Now let w be a vector in W. Then we have $0 = w + (-1)w \in W$. Therefore by setting $u = 0$ we obtain $av = 0 + av \in W$ for every $v \in W$ and $a \in F$. Thus W is closed under scalar multiplication. Hence by the previous theorem W is a subspace. ■

***Example* 2.10.** It is easy to see that $\{0\}$ is a subspace of V. This subspace is called the zero subspace.

***Example* 2.11.** Let $A \in F^{m\times n}$. Then the set of solutions of the homogeneous system of linear equations $Ax = 0$ is a subspace of F^n. Because $x = 0$ is a solution of the system, so the set of solutions is nonempty. Also if x, y are solutions of the system, and $a \in F$, then we have

$$A(x + ay) = Ax + aAy = 0 + a0 = 0.$$

Thus $x + ay$ is also a solution of the system, as desired.

***Example* 2.12.** It is easy to show that the set of polynomials in $F[x]$ whose degree is less than or equal to some given integer n, is a subspace of $F[x]$.

Theorem 2.13. *The intersection of a nonempty family of subspaces of V is a subspace of V.*

Proof. Suppose $\{W_\alpha : \alpha \in I\}$ is a nonempty family of subspaces of V. Let

$$W := \bigcap_{\alpha\in I} W_\alpha = \{v \in V : v \in W_\alpha \text{ for every } \alpha\}.$$

Then we have $0 \in W$, since $0 \in W_\alpha$ for every α. Now let $u, v \in W$ and $a \in F$. Then $u, v \in W_\alpha$ for every α. Hence we have $u + v, au \in W_\alpha$ for every α, because each W_α is a subspace. Thus $u+v, au \in W$ by the definition of intersection of sets. Therefore W is a subspace. ■

Definition 2.14. Let $\mathcal{S}, W \subset V$. We say the subspace **spanned** by $\mathcal{S}$ is W if the following conditions hold:

(i) $\mathcal{S} \subset W$,
(ii) W is a subspace of V,
(iii) W is the smallest subspace of V, with respect to inclusion, that contains $\mathcal{S}$, i.e. if W' is a subspace of V that contains $\mathcal{S}$, then we must have $W \subset W'$.

The subspace W is also called the subspace **generated** by $\mathcal{S}$, or the **span** of $\mathcal{S}$; and is denoted by $\text{span}(\mathcal{S})$. A vector space, or a subspace of a vector space, is called **finitely generated** if it is spanned by a finite set.

Remark. Note that a priori we do not know that every subset of a vector space generates a subspace; because it is not obvious that we can always find the smallest subspace that contains a given subset. However, the following theorem shows that this is always possible, and the span of every subset of a vector space exists.

Theorem 2.15. *Let $\mathcal{S} \subset V$. Then* $\text{span}(\mathcal{S})$ *exists, i.e. there is a unique subspace of V that contains $\mathcal{S}$, and is contained in any subspace containing $\mathcal{S}$.*

Proof. Let W be the intersection of all subspaces containing $\mathcal{S}$. Note that V itself is a subspace containing $\mathcal{S}$, so the family of all subspaces containing $\mathcal{S}$ is nonempty, and therefore their intersection is defined. Also note that by Theorem 2.13, W is a subspace of V, since it is the intersection of a nonempty family of subspaces of V. In addition, W contains $\mathcal{S}$; because it is the intersection of a nonempty family of sets, and each one of those sets contains $\mathcal{S}$.

Now suppose W' is a subspace of V that contains $\mathcal{S}$. Then $W \subset W'$, because W is the intersection of all subspaces containing $\mathcal{S}$, and W' is one of the subspaces that contains $\mathcal{S}$. Hence by definition we have $W = \text{span}(\mathcal{S})$. Finally, let us show that W is the only subspace that satisfies all the conditions in the definition of $\text{span}(\mathcal{S})$. Suppose $\tilde{W}$ is also a smallest subspace that contains $\mathcal{S}$. Then we must have $W \subset \tilde{W}$; since $\tilde{W}$ is a subspace that contains $\mathcal{S}$, and W is a smallest subspace containing $\mathcal{S}$. On the other hand, we must have $\tilde{W} \subset W$; since W is a subspace that contains $\mathcal{S}$, and $\tilde{W}$ is a smallest subspace containing $\mathcal{S}$. Therefore $W = \tilde{W}$, as desired. ∎

***Example* 2.16.** We have $\text{span}(\emptyset) = \{0\}$, since $\{0\}$ is a subspace that contains $\emptyset$, and is contained in any other subspace.

***Example* 2.17.** Let $v \in V$. We claim that $\text{span}(\{v\}) = \{av : a \in F\}$. Let us denote $\{av : a \in F\}$ by W. It is obvious that $v \in W$, since $v = 1v$. Furthermore, W is a subspace; because it contains v, so it is nonempty. In addition, if $u, w \in W$ and $c \in F$, then we have $u = av$ and $w = bv$, for some $a, b \in F$; hence we have

$$u + cw = av + cbv = (a + cb) \in W.$$

Thus W is a subspace. Now suppose W' is a subspace that contains v. Then for every $a \in F$ we have $av \in W'$, since W' is closed under scalar multiplication.

Therefore we have $W \subset W'$. So W is the smallest subspace that contains v; hence it is $\text{span}(\{v\})$. In Theorem 2.21, we will examine a more general version of this example. Also, note that if we put $v = 0$ in this example, we will get $\text{span}(\{0\}) = \{0\}$; because $a0 = 0$, for every $a \in F$.

Definition 2.18. A **list** of vectors in V is a finite sequence of elements of V, i.e. it is a function from the set $\{1, \dots, k\}$ to V, for some $k \in \mathbb{N}$.

Remark. Note that a list of vectors in V can be considered as an element of V^k for some $k \in \mathbb{N}$. But we denote a list by simply writing the vectors in its image, like $v_1, v_2, \dots, v_k$.

Remark. We need to work with lists of vectors instead of sets of vectors, since we want to allow repetition of vectors, and more importantly we want each vector to have a precise position among other vectors i.e. we want the vectors to have an order. In the sequel, whenever we use a set of vectors in a notion that is defined for lists of vectors, we implicitly assume that we have arranged the elements of the set in an arbitrary sequence. Although we have to check that the notion does not depend on the particular order of the elements of the set, which is the case for all the notions we define here. Finally, for convenience we consider the empty set to be a list of vectors too, called the *empty list*.

Remark. When we talk about the span of a list of vectors $v_1, \dots, v_k$, we mean the span of the set of vectors in that list, i.e. the span of the set $\{v_1, \dots, v_k\}$. The span of $v_1, \dots, v_k$ is usually denoted by $\text{span}(v_1, \dots, v_k)$.

Definition 2.19. A **linear combination** of a finite list of vectors $v_1, \dots, v_k \in V$ is a vector of the form

$$a_1 v_1 + \cdots + a_k v_k,$$

where $a_1, \dots, a_k \in F$. We also denote the above vector by $\sum_{j=1}^{k} a_j v_j$.

Remark. Note that by the generalized associativity and commutativity rules, the above expression is unambiguously defined and is independent of the order of the summands. See Section A.6.

Remark. Also note that a list of vectors has finitely many vectors, so we are not defining linear combinations of infinitely many vectors. In fact, adding infinitely many vectors requires some notion of limit, which is not available in an arbitrary vector space.

Proposition 2.20. *The subspaces of a vector space are closed under forming linear combinations, i.e. if $W \subset V$ is a subspace and $v_1, \dots, v_k \in W$ then*

$$a_1 v_1 + \cdots + a_k v_k \in W$$

for every $a_1, \dots, a_k \in F$.

Proof. The proof is by induction on k. The case of $k = 1$ is trivially true, since W is closed under scalar multiplication. Suppose the claim is true for some k, and we want to prove it for $k + 1$. Let $v_1, \ldots, v_{k+1} \in W$ and $a_1, \ldots, a_{k+1} \in F$. Then by the induction hypothesis we have $a_1v_1 + \cdots + a_kv_k \in W$. On the other hand we have $a_{k+1}v_{k+1} \in W$, since W is closed under scalar multiplication. Finally we have

$$a_1v_1 + \cdots + a_kv_k + a_{k+1}v_{k+1} \in W,$$

because W is closed under addition. ■

Theorem 2.21. *Suppose $v_1, \ldots, v_n \in V$. Then the subspace generated by $v_1, \ldots, v_n$ equals the set of all linear combinations of $v_1, \ldots, v_n$, i.e.*

$$\operatorname{span}(v_1, \ldots, v_n) = \{a_1v_1 + \cdots + a_nv_n : \textit{for every } a_1, \ldots, a_n \in F\}.$$

Proof. Let $W := \{a_1v_1 + \cdots + a_nv_n : \text{for every } a_1, \ldots, a_n \in F\}$. In order to show that $W = \operatorname{span}(v_1, \ldots, v_n)$, we have to prove that W is the smallest subspace with respect to inclusion that contains $v_1, \ldots, v_n$. First let us show that W is a subspace containing $v_1, \ldots, v_n$. It is obvious that W contains $v_1, \ldots, v_n$, since if we take $a_j = 1$, and $a_i = 0$ for every $i \neq j$, then we get

$$v_j = 0v_1 + \cdots + 0v_{j-1} + 1v_j + 0v_{j+1} + \cdots + 0v_n \in W.$$

Thus in particular W is nonempty. Now suppose $u, v \in W$ and $a \in F$. Then by definition of W we have $u = \sum_{j=1}^{n} a_jv_j$ and $v = \sum_{j=1}^{n} b_jv_j$, for some $a_j, b_j \in F$. Hence we have

$$u + av = \sum_{j=1}^{n} a_jv_j + a\sum_{j=1}^{n} b_jv_j = \sum_{j=1}^{n}(a_jv_j + ab_jv_j) = \sum_{j=1}^{n}(a_j + ab_j)v_j \in W.$$

Therefore W is a subspace, as desired. Now let W' be a subspace that contains $v_1, \ldots, v_n$. Then by the previous proposition we know that W' contains every linear combination of $v_1, \ldots, v_n$, so $W \subset W'$. Thus W is the smallest subspace that contains $v_1, \ldots, v_n$; hence we have $W = \operatorname{span}(v_1, \ldots, v_n)$. ■

Remark. Let $\mathcal{S} \subset V$ be an arbitrary set of vectors. Then we can similarly show that the subspace generated by $\mathcal{S}$ equals the set of all linear combinations of any list of vectors in $\mathcal{S}$, i.e.

$$\operatorname{span}(\mathcal{S}) = \{a_1v_1 + \cdots + a_kv_k : \text{for every } k \in \mathbb{N}, v_1, \ldots, v_k \in \mathcal{S}, \text{and } a_1, \ldots, a_k \in F\}.$$

Remark. In contrast to the definition of $\operatorname{span}(\mathcal{S})$ in terms of subspaces containing $\mathcal{S}$, the above theorem gives a concrete description of $\operatorname{span}(\mathcal{S})$ that only uses the elements of $\mathcal{S}$ itself.

Proposition 2.22. *Suppose $\mathcal{S} \subset V$, and $\mathcal{A} \subset \operatorname{span}(\mathcal{S})$. Then $\operatorname{span}(\mathcal{A}) \subset \operatorname{span}(\mathcal{S})$.*

Proof. $\operatorname{span}(\mathcal{S})$ is a subspace that contains $\mathcal{A}$, so it must contain $\operatorname{span}(\mathcal{A})$ too. ■

Remark. Let us give another proof for the above proposition by using the description of span in terms of linear combinations. Suppose $\sum_{j=1}^{n} a_j v_j \in \operatorname{span}(\mathcal{A})$ where $v_1, \ldots, v_n \in \mathcal{A} \subset \operatorname{span}(\mathcal{S})$. Then we have $v_j = \sum_{k=1}^{m} b_{jk} u_k$, where $u_1, \ldots, u_m \in \mathcal{S}$. Note that we can use the same set of vectors in $\mathcal{S}$ for all v_j's, since there are only finitely many v_j's, and we can set $b_{jk} = 0$ if u_k did not appear in the expansion of v_j. Hence we have

$$\sum_{j=1}^{n} a_j v_j = \sum_{j=1}^{n} a_j \left(\sum_{k=1}^{m} b_{jk} u_k \right) = \sum_{k=1}^{m} \left(\sum_{j=1}^{n} a_j b_{jk} \right) u_k \in \operatorname{span}(\mathcal{S}).$$

Notice that the essence of the above argument is that a linear combination of several linear combinations of some vectors is itself a linear combination of those vectors.

Proposition 2.23. *Suppose $v_1, \ldots, v_k, b \in F^n$. Let $A \in F^{n \times k}$ be the matrix whose j-th column is v_j. Then $b \in \operatorname{span}(v_1, \ldots, v_k)$ if and only if the linear system $Ax = b$ has a solution $x \in F^k$.*

Remark. Note that we put the vectors v_j in the columns of A, not its rows.

Proof. Suppose $b \in \operatorname{span}(v_1, \ldots, v_k)$. Then there are scalars $x_1, \ldots, x_k \in F$ such that $b = x_1 v_1 + \cdots + x_k v_k$. Let $x := [x_1, \ldots, x_k]^\intercal \in F^k$. Then we have

$$Ax = x_1 A_{\cdot,1} + \cdots + x_k A_{\cdot,k} = x_1 v_1 + \cdots + x_k v_k = b.$$

So the system has a solution. On the other hand if the system has a solution $x \in F^k$, then we have $b = Ax = x_1 A_{\cdot,1} + \cdots + x_k A_{\cdot,k} = x_1 v_1 + \cdots + x_k v_k$, where $x_1, \ldots, x_k$ are the components of x. Hence b is a linear combination of $v_1, \ldots, v_k$. ■

***Exercise* 2.24.** Suppose W is a subspace of V, and U is a subset of W. Show that U is a subspace of W if and only if it is a subspace of V.

Solution. The proof is a simple application of Proposition 2.9. Let us restate that proposition here, with a slight change, in order to make our argument more clear. It says that if $U \subset V$ is nonempty, then U is a subspace of V if and only if for all $u, v \in V$ and $a \in F$ we have

$$u, v \in U \implies u + av \in U.$$

Now suppose U is a subspace of W. Then U is nonempty, and $U \subset W \subset V$. Let $a \in F$. If $u, v \in V$ also belong to U, then we have $u, v \in U \subset W$. Hence $u + av \in U$, and therefore U is a subspace of V.

Conversely, suppose U is a subspace of V. Then U is nonempty, and $U \subset W$. Let $a \in F$. If $u, v \in W$ also belong to U, then we have $u, v \in U \subset V$. Hence $u + av \in U$, and therefore U is a subspace of W. ■

2.3 Linear Independence

Suppose we have a subspace W, and we know that the vectors $v_1, \ldots, v_n$ generate W, i.e. we know that $W = \text{span}(v_1, \ldots, v_n)$. Now, a natural question that arises is that whether we can find a simpler set that generates W, i.e. a set that generates W, and has fewer elements. For example, suppose $W = \text{span}(v, 2v)$, for some vector v. Then by Theorem 2.21 we have

$$W = \{av + 2bv : a, b \in F\}.$$

However, $av + 2bv = (a + 2b)v \in \text{span}(v)$. Thus $W \subset \text{span}(v)$. On the other hand, it is obvious that $\text{span}(v) \subset W$, since W is a subspace that contains v. Hence we have $W = \text{span}(v)$, i.e. we can generate W with only one vector, namely v. Therefore, the set of generators $v, 2v$ for W is not optimal, i.e. we can reduce it to a smaller set that generates W.

In general, suppose we have subspace, and a set that generates it. We want to know whether this given set of generators contains unnecessary vectors or not. The following notion helps us to answer this question.

Definition 2.25. A finite list of vectors $v_1, \ldots, v_k \in V$ is called **linearly independent** if for every $a_1, \ldots, a_k \in F$ we have

$$a_1v_1 + \cdots + a_kv_k = 0 \implies a_j = 0 \text{ for every } j.$$

We also consider the empty list to be linearly independent.

Oppositely, a finite list of vectors $v_1, \ldots, v_k \in V$ is called **linearly dependent** if it is not linearly independent, i.e. if there are $a_1, \ldots, a_k \in F$, where at least one of the a_j's is nonzero, such that

$$a_1v_1 + \cdots + a_kv_k = 0.$$

Remark. As with linear combinations, here we prefer to work with lists of vectors instead of sets of vectors, since we want the vectors to have an order, and we want to allow repetitions. Similarly, whenever we talk about the linear independence, or the linear dependence, of a set of vectors, we implicitly assume that we have arranged the elements of the set in an arbitrary sequence. It is easy to see that these notions do not depend on the particular order of the elements of the set.

Remark. In these notes we only consider finite linearly independent sets. But let us mention that an arbitrary subset $\mathcal{S}$ of a vector space is called linearly independent if every finite subset of $\mathcal{S}$ is linearly independent. And $\mathcal{S}$ is called linearly dependent if it has a finite linearly dependent subset.

Remark. If there is a repetition in a list of vectors $v_1, \ldots, v_k$, say for example $v_i = v_j$, then this list is linearly dependent. Because we have

$$0v_1 + \cdots + 0v_{i-1} + 1v_i + 0v_{i+1} + \cdots + 0v_{j-1} + (-1)v_j + 0v_{j+1} + \cdots + 0v_k = 0.$$

Also if one of the vectors in the list is zero, say $v_i = 0$, then the list is linearly dependent, since

$$0v_1 + \cdots + 0v_{i-1} + 1v_i + 0v_{i+1} + \cdots + 0v_k = 0.$$

Thus if a list is linearly independent then its vectors are all nonzero and distinct.

***Example* 2.26.** Let $u, v \in V$. If u is linearly dependent then there is a nonzero $a \in F$ such that $au = 0$. Hence we must have $u = 0$. Conversely, it is obvious that if $u = 0$ then it is linearly dependent, since for example $1 \cdot 0 = 0$. Thus u is linearly dependent if and only if it is zero. Equivalently, u is linearly independent if and only if it is nonzero.

Now if u, v are linearly dependent then there are $a, b \in F$ such that $au + bv = 0$, and at least one of the a, b is nonzero. If $a \neq 0$ then we have $u = -a^{-1}bv$, and if $b \neq 0$ then we have $v = -b^{-1}au$. Thus one of the u, v is a scalar multiple of the other. Conversely if one of the u, v is a scalar multiple of the other, say $u = cv$ for some $c \in F$, then we have $1u - cv = 0$. Therefore u, v are linearly dependent. Hence u, v are linearly dependent if and only if one of them is a scalar multiple of the other.

Proposition 2.27. *A list $v_1, \ldots, v_k \in V$ is linearly dependent if and only if one of its elements is in the span of the others, i.e. for some j we have*

$$v_j \in \operatorname{span}(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k).$$

Proof. If for some j we have $v_j \in \operatorname{span}(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k)$, then we must have

$$v_j = a_1v_1 + \cdots + a_{j-1}v_{j-1} + a_{j+1}v_{j+1} + \cdots + a_kv_k,$$

for some $a_1, \ldots, a_k \in F$. But then we have

$$a_1v_1 + \cdots + a_{j-1}v_{j-1} + (-1)v_j + a_{j+1}v_{j+1} + \cdots + a_kv_k = 0.$$

Hence $v_1, \ldots, v_k$ are linearly dependent, since the coefficient of v_j is nonzero.

Conversely if $v_1, \ldots, v_k$ are linearly dependent, then we have $a_1v_1 + \cdots + a_kv_k = 0$ for some $a_1, \ldots, a_k \in F$, and at least one of the a_i's is nonzero. Suppose $a_j \neq 0$. Then we have

$$v_j = (-a_j^{-1}a_1)v_1 + \cdots + (-a_j^{-1}a_{j-1})v_{j-1} + (-a_j^{-1}a_{j+1})v_{j+1} + \cdots + (-a_j^{-1}a_k)v_k.$$

Therefore $v_j \in \operatorname{span}(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k)$. ■

Proposition 2.28. *A subset of a linearly independent set is linearly independent.*

Proof. Suppose $v_1, \ldots, v_k$ is linearly independent, and $v_{j_1}, \ldots, v_{j_n}$ is a subset of $v_1, \ldots, v_k$. Suppose $a_{j_1} v_{j_1} + \cdots + a_{j_n} v_{j_n} = 0$ for some $a_{j_i} \in F$. Then we have

$$0v_1 + \cdots + 0v_{j_1-1} + a_{j_1} v_{j_1} + 0v_{j_1+1} + \cdots + a_{j_2} v_{j_2} + \cdots + a_{j_n} v_{j_n} + \cdots + 0v_k = 0.$$

Hence all the coefficients of the above linear combination must be zero, since $v_1, \ldots, v_k$ is linearly independent. In particular we have $a_{j_1} = \cdots = a_{j_n} = 0$. Therefore $v_{j_1}, \ldots, v_{j_n}$ is linearly independent too. ■

Theorem 2.29. *Suppose* $w_1, \ldots, w_n \in V$, and $u_1, \ldots, u_m \in \operatorname{span}(w_1, \ldots, w_n)$. If $m > n$ then $u_1, \ldots, u_m$ *are linearly dependent.*

Remark. An equivalent way of stating the above result is that if $u_1, \ldots, u_m \in \operatorname{span}(w_1, \ldots, w_n)$, and $u_1, \ldots, u_m$ are linearly independent, then $m \leq n$.

Proof. Let $W := \operatorname{span}(w_1, \ldots, w_n)$. The proof is by induction on n. First suppose that $n = 1$. In this case we have $W := \operatorname{span}(w_1)$, so $u_j = a_j w_1$ for $j = 1, \ldots, m$ and $a_j \in F$. If one of the u_j's is 0, for example $u_1 = 0$, then the list is linearly dependent as we have $1u_1 + 0u_2 + \cdots + 0u_m = 0$. Thus suppose that u_j's are all nonzero. Therefore a_j's are all nonzero too. But then we have

$$a_2 u_1 - a_1 u_2 + 0u_3 + \cdots + 0u_m = a_2(a_1 w_1) - a_1(a_2 w_1) = (a_2 a_1 - a_1 a_2) w_1 = 0.$$

Hence the list is linearly dependent, since the coefficients of u_1, u_2 in the above relation are nonzero.

Now suppose that the theorem is true for $n = k - 1$. We want to deduce it for $n = k$. We know that u_j's are in the span of w_i's. Hence for each $j \leq m$ we have

$$u_j = a_{j1} w_1 + \cdots + a_{jk} w_k,$$

where $a_{ji} \in F$. If the coefficient of w_k is zero for all j, then u_j's are in the span of $k-1$ vectors, and therefore they are linearly dependent by the induction hypothesis. So suppose one of the a_{jk}'s, say a_{mk}, is nonzero. Then for $j = 1, \ldots, m-1$ set

$$v_j = u_j - (a_{jk} a_{mk}^{-1}) u_m. \qquad (*)$$

These are $m - 1$ vectors, and they are in the span of $k - 1$ vectors $w_1, \ldots, w_{k-1}$, since we made the coefficient of w_k zero. But we have $m > k$ so $m - 1 > k - 1$. Hence by the induction hypothesis v_j's are linearly dependent. This means that there are scalars $c_j \in F$, where at least one of them is nonzero, such that

$$c_1 v_1 + \cdots + c_{m-1} v_{m-1} = 0.$$

Now by using $(*)$ and rearranging the terms in the above relation we get

$$c_1u_1 + \cdots + c_{m-1}u_{m-1} - (c_1a_{1k}a_{mk}^{-1} + \cdots + c_{m-1}a_{m-1k}a_{mk}^{-1})u_m = 0.$$

Therefore u_j's are also linearly dependent, as desired. ■

Proposition 2.30. *Suppose $\mathcal{B} \subset V$ is a linearly independent set, and $v \in V$. Then $\mathcal{B} \cup \{v\}$ is linearly independent if and only if $v \notin \text{span}(\mathcal{B})$.*

Remark. An equivalent way of stating the above result is that if $\mathcal{B}$ is linearly independent, then $\mathcal{B} \cup \{v\}$ is linearly dependent if and only if $v \in \text{span}(\mathcal{B})$.

Proof. We will prove the contrapositive of the proposition which is stated in the above remark. If $v \in \text{span}(\mathcal{B})$ then $\mathcal{B} \cup \{v\}$ is linearly dependent by Proposition 2.27. Conversely suppose that $\mathcal{B} \cup \{v\}$ is linearly dependent. Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$. Then there are scalars $a_1, \ldots, a_n, a$, where at least one of them is nonzero, such that $a_1v_1 + \cdots + a_nv_n + av = 0$. If $a = 0$ then $a_1v_1 + \cdots + a_nv_n = 0$. But $\mathcal{B}$ is linearly independent, so we must have $a_j = 0$ for all j, which is contrary to our assumption. Therefore $a \neq 0$. Hence we have $v = -a^{-1}a_1v_1 - \cdots - a^{-1}a_nv_n$. Thus $v \in \text{span}(\mathcal{B})$ as desired. ■

Proposition 2.31. *Suppose $v_1, \ldots, v_k \in F^n$. Let $A \in F^{n\times k}$ be the matrix whose j-th column is v_j. Then $v_1, \ldots, v_k$ are linearly independent if and only if the homogeneous linear system $Ax = 0$ has only the trivial solution $x = 0$.*

Remark. Note that we put the vectors v_j in the columns of A, not its rows.

Proof. Suppose $x := [x_1, \ldots, x_k]^\intercal \in F^k$. Then we have

$$Ax = x_1A_{\cdot,1} + \cdots + x_kA_{\cdot,k} = x_1v_1 + \cdots + x_kv_k.$$

Therefore $Ax = 0$ if and only if $x_1v_1 + \cdots + x_kv_k = 0$. Now note that $x \neq 0$ if and only if at least one of the x_j's is nonzero. Hence the existence of a nontrivial solution for the system $Ax = 0$ is equivalent to the linear dependence of $v_1, \ldots, v_k$. The contrapositive of this statement is our desired result. ■

2.4 Bases and Dimension

Definition 2.32. A finite list of vectors $v_1, \ldots, v_n \in V$ is called a **basis** for V, if it is linearly independent and generates V.

Remark. In general a set $\mathcal{B} \subset V$ is a basis, if it is linearly independent and generates V. But we only consider finite bases in these notes.

Remark. Note that a list of vectors has an order, so a basis has an order too. Hence some authors use the term *ordered basis*. Also note that no vector can be repeated in a linearly independent list, so the elements of a basis are all distinct. The next theorem shows that why bases are important and useful.

Theorem 2.33. *Let $v_1, \ldots, v_n \in V$ be a list of vectors. Then $v_1, \ldots, v_n$ is a basis for V if and only if every vector $v \in V$ can be written as a unique linear combination of $v_1, \ldots, v_n$.*

Remark. The uniqueness in the above theorem means that if for some $a_i, b_i \in F$ we have

$$a_1 v_1 + \ldots + a_n v_n = v = b_1 v_1 + \ldots + b_n v_n,$$

then $b_i = a_i$ for each i. In other words the n-tuple $(a_1, \ldots, a_n)$ is uniquely determined by v.

Proof. Suppose $v_1, \ldots, v_n$ is a basis for V. Let $v \in V$. Then $v \in \operatorname{span}(v_1, \ldots, v_n)$, so there are $a_i \in F$ such that $v = a_1 v_1 + \ldots + a_n v_n$. Now suppose for $b_i \in F$ we also have $v = b_1 v_1 + \ldots + b_n v_n$. Then by subtracting these two equations we get

$$(a_1 - b_1)v_1 + \ldots + (a_n - b_n)v_n = 0.$$

But $v_1, \ldots, v_n$ are linearly independent, so for every i we have $a_i = b_i$ as desired.

Conversely suppose that every vector $v \in V$ can be written as a unique linear combination of $v_1, \ldots, v_n$. Then $v \in \operatorname{span}(v_1, \ldots, v_n)$. Thus $v_1, \ldots, v_n$ generate V. Next suppose $a_1 v_1 + \ldots + a_n v_n = 0$ for some $a_i \in F$. Then we have

$$a_1 v_1 + \ldots + a_n v_n = 0 = 0v_1 + \ldots + 0v_n.$$

But 0 must be written uniquely as a linear combination of $v_1, \ldots, v_n$. Therefore $a_i = 0$ for every i. Hence $v_1, \ldots, v_n$ are linearly independent, and consequently they form a basis for V. ∎

***Example* 2.34.** The list of vectors $e_1, \ldots, e_n$ is a basis for F^n, called its *standard basis*. To see this let $x = [x_1, \ldots, x_n]^\intercal \in F^n$. Then we have $x = x_1 e_1 + \cdots + x_n e_n$. So $e_1, \ldots, e_n$ generate F^n. On the other hand, suppose $x_1 e_1 + \cdots + x_n e_n = 0$ for some $x_1, \ldots, x_n \in F$. Then we have $[x_1, \ldots, x_n]^\intercal = x_1 e_1 + \cdots + x_n e_n = 0$. Hence $x_j = 0$ for every j. Thus $e_1, \ldots, e_n$ are also linearly independent.

Theorem 2.35. *Suppose $v_1, \ldots, v_n$ and $u_1, \ldots, u_m$ are bases for V. Then $m = n$.*

Remark. In other words, the number of vectors in a basis is uniquely determined by the vector space.

Proof. We have $u_1, \ldots, u_m \in \text{span}(v_1, \ldots, v_n)$, and we know that $u_1, \ldots, u_m$ are linearly independent. Hence by Theorem 2.29 we must have $m \leq n$. Similarly we can show that $n \leq m$. Thus $m = n$. ■

Definition 2.36. If a vector space V has a finite basis, then the **dimension** of V is the number of vectors in its basis. A vector space is called **finite dimensional** if it has a finite basis, and is called **infinite dimensional** otherwise.

Notation. We know that the number of vectors in a basis is independent of the basis, so the dimension of a finite dimensional vector space is uniquely determined by the vector space. If V is a finite dimensional vector space over a field F, we denote its dimension by

$$\dim V.$$

And when we want to emphasize the field of scalars, we denote the dimension of V by

$$\dim_F V,$$

and we call it the dimension of V over F.

***Example* 2.37.** We have $\dim F^n = n$, since F^n has a basis with n elements, i.e. its standard basis. We also have $\dim F^{m\times n} = mn$. To see this consider the matrices E_{ij} for $i \leq m$ and $j \leq n$, where the entries of E_{ij} are all zero except its ij-th entry which equals 1. Then it is easy to show that $\{E_{ij} : i \leq m, j \leq n\}$ is a basis for $F^{m\times n}$.

***Example* 2.38.** Consider the zero vector space $\{0\}$. Then its dimension is 0, since the empty set $\emptyset$ is a linearly independent set that spans $\{0\}$.

Theorem 2.39. *Suppose that V is the span of a finite set $\mathcal{S} \subset V$, and $\mathcal{B} \subset V$ is a linearly independent set. Then there is $\mathcal{A} \subset \mathcal{S}$ such that $\mathcal{B} \cup \mathcal{A}$ is a basis for V.*

Remark. An important consequence of the above theorem is that in a finitely generated vector space we can extend a linearly independent set $\mathcal{B}_1$ to a basis, and we can reduce a finite spanning set $\mathcal{S}_1$ to a basis. For the first claim consider an arbitrary finite spanning set $\mathcal{S}$ for the space. Then there is $\mathcal{A}_1 \subset \mathcal{S}$ such that $\mathcal{B}_1 \cup \mathcal{A}_1$ is a basis, i.e. we have extended $\mathcal{B}_1$ to a basis. Now for the second claim consider the linearly independent set $\mathcal{B} = \emptyset$. Then there is $\mathcal{A} \subset \mathcal{S}_1$ such that $\mathcal{A} = \emptyset \cup \mathcal{A}$ is a basis, i.e. we have reduced $\mathcal{S}_1$ to a basis.

Proof. Suppose $\mathcal{S} = \{u_1, \ldots, u_m\}$ spans V. We are looking for a set $\mathcal{A} \subset \mathcal{S}$ such that $\mathcal{B} \cup \mathcal{A}$ is a basis. This means that $\mathcal{B} \cup \mathcal{A}$ must spans V, and be linearly independent. Consider the class of sets $\tilde{\mathcal{A}} \subset \mathcal{S}$ such that $\mathcal{B} \cup \tilde{\mathcal{A}}$ is linearly independent. Since $\mathcal{S}$ is finite, it has finitely many subsets. Therefore there are finitely many sets $\tilde{\mathcal{A}}$ in the above class. Also the above class is nonempty, since it contains

$\emptyset \subset \mathcal{S}$. Now, in this finite class consider a set with the greatest number of elements, and call it $\mathcal{A}$. Note that there might be several sets with the greatest number of elements, but we only need one of them.

We claim that $\mathcal{B} \cup \mathcal{A}$ is a basis. First note that $\mathcal{B} \cup \mathcal{A}$ is linearly independent, since we have chosen $\mathcal{A}$ among the subsets of $\mathcal{S}$ whose union with $\mathcal{B}$ is linearly independent. Thus we only need to show that $\mathcal{B} \cup \mathcal{A}$ spans V. Now if $\mathcal{A} = \mathcal{S}$ then $\mathcal{S} \subset \mathcal{B} \cup \mathcal{A}$. Hence by Proposition 2.22 we have

$$V = \operatorname{span}(\mathcal{S}) \subset \operatorname{span}(\mathcal{B} \cup \mathcal{A}) \subset V \implies \operatorname{span}(\mathcal{B} \cup \mathcal{A}) = V.$$

Otherwise there is $v \in \mathcal{S} - \mathcal{A}$. Then $\mathcal{A} \cup \{v\}$ is a subset of $\mathcal{S}$ with more elements than $\mathcal{A}$. Thus $\mathcal{B} \cup \mathcal{A} \cup \{v\}$ is linearly dependent, since $\mathcal{A}$ has the greatest number of elements amongst the subsets of $\mathcal{S}$ whose union with $\mathcal{B}$ is linearly independent. Therefore by Proposition 2.30 we must have $v \in \operatorname{span}(\mathcal{B} \cup \mathcal{A})$. As v was arbitrary we obtain that $\mathcal{S} \subset \mathcal{B} \cup \mathcal{A}$. Hence $\operatorname{span}(\mathcal{B} \cup \mathcal{A}) = V$ as before. ■

***Exercise* 2.40.** Suppose V is finite dimensional, and $v_1, \ldots, v_k \in V$ is a nonempty linearly independent set of vectors that is not a basis for V. Show that there is more than one set of vectors $w_1, \ldots, w_m \in V$ such that $v_1, \ldots, v_k, w_1, \ldots, w_m$ is a basis for V. Is the number of such sets $w_1, \ldots, w_m$ necessarily infinite?

Theorem 2.41. *Every finitely generated vector space has a basis.*

Remark. As a consequence, every finitely generated vector space is finite dimensional.

Proof. Suppose $\mathcal{S}$ is a finite set that spans the space. Consider the linearly independent set $\mathcal{B} = \emptyset$. Then by Theorem 2.39 there is $\mathcal{A} \subset \mathcal{S}$ such that $\mathcal{A} = \emptyset \cup \mathcal{A}$ is a basis, i.e. the space has a basis. ■

Remark. Let us give another proof for the above theorem, that describes an algorithm for finding a basis. Suppose V is a finitely generated vector space, i.e. $V = \operatorname{span}(w_1, \ldots, w_m)$. Consider the list $w_1, \ldots, w_m$. If all the vectors in the list are zero then $V = \{0\}$, and $\emptyset$ is a basis for V. So we assume that some of the vectors in the list are nonzero. We start with the first nonzero vector in the list, say it is w_j. Then we remove the vectors $w_1, \ldots, w_{j-1}$ from the list. Next we consider w_{j+1}. If w_{j+1} is in the span of w_j, we remove it from the list; otherwise we keep it. Now suppose we have gone through the list, and kept the vectors $w_j, w_{j_2}, \ldots, w_{j_k}$. If the next vector, say w_l, is in the span of $w_j, w_{j_2}, \ldots, w_{j_k}$, we remove it from the list; otherwise we keep it. We repeat this process until there is no more vector to consider. Then we have a list $w_j, w_{j_2}, \ldots, w_{j_n}$.

Note that the above list is linearly independent by our construction. Because the first vector is nonzero, and therefore it is linearly independent. And at each

step we added a vector that is not in the span of previous vectors, so we kept the list linearly independent by Proposition 2.30. Finally note that the removed vectors are in the span of $w_j, w_{j_2}, \ldots, w_{j_n}$, so each w_i is in the span of $w_j, w_{j_2}, \ldots, w_{j_n}$. Hence by Proposition 2.22 we have

$$V = \text{span}(w_1, \ldots, w_m) \subset \text{span}(w_j, \ldots, w_{j_n}) \subset V.$$

Thus $w_j, w_{j_2}, \ldots, w_{j_n}$ is a basis for V.

We can even modify the above algorithm to prove Theorem 2.39. Suppose $\mathcal{B} = \{v_1, \ldots, v_k\}$ is linearly independent, and $\mathcal{S} = \{w_1, \ldots, w_m\}$ spans V. Then we start with the list of elements of $\mathcal{B}$, and we go through the list of elements of $\mathcal{S}$ one by one. At each step if a vector belongs to the span of the list we ignore it; otherwise we add it to the list. At the end we have constructed a basis for V by adding some elements of $\mathcal{S}$ to $\mathcal{B}$. The proof of this fact is similar to the above argument.

Remark. When $V = F^n$, we have to solve a linear system at each step of the above algorithm, to determine whether a vector is in the span of the previous vectors. But we can actually find a basis among $w_1, \ldots, w_m$ by solving only one system of linear equations. So the above algorithm is not very useful in this case. We will describe the faster method in Theorem 2.47.

Remark. If V is not finitely generated, then we can apply a modified version of the above algorithm to produce linearly independent sets of vectors with arbitrarily large number of elements. To do this we start with a nonzero vector $v_1 \in V$, which we know is linearly independent. Suppose we have chosen $v_1, \ldots, v_k$, and they are linearly independent. Then $V \neq \text{span}(v_1, \ldots, v_k)$, since V is not finitely generated. Hence there is $v_{k+1} \in V - \text{span}(v_1, \ldots, v_k)$. Thus by Proposition 2.30, $v_1, \ldots, v_k, v_{k+1}$ is linearly independent too. Therefore by repeating this process we can build linearly independent sets with any number of elements.

Theorem 2.42. *Suppose that V is a finite dimensional vector space, and* $\dim V = n$. *Then*

(i) *Any linearly independent subset of V has at most n elements, and if it has n elements then it is a basis.*

(ii) *Any set that generates V has at least n elements, and if it has n elements then it is a basis.*

Remark. A consequence of the above theorem is that a maximal linearly independent set is a basis, and a minimal spanning set is also a basis. These facts are actually true in infinite dimensional vector spaces too, and can be used to prove the existence of a basis for those spaces.

Proof. **(i)** Suppose $\mathcal{B} \subset V$ is linearly independent. Then we have seen that there is a basis $\tilde{\mathcal{B}} \supset \mathcal{B}$. Hence the number of elements of $\mathcal{B}$ is at most the number of

elements of $\tilde{\mathcal{B}}$, which is $n = \dim V$. If $\mathcal{B}$ has exactly n elements, then we must have $\mathcal{B} = \tilde{\mathcal{B}}$, so $\mathcal{B}$ must be a basis.

(ii) Suppose $\mathcal{S} \subset V$ is a spanning set. Then we have seen that there is a basis $\tilde{\mathcal{S}} \subset \mathcal{S}$. Hence the number of elements of $\mathcal{S}$ is at least the number of elements of $\tilde{\mathcal{S}}$, which is $n = \dim V$. If $\mathcal{S}$ has exactly n elements, then we must have $\mathcal{S} = \tilde{\mathcal{S}}$, so $\mathcal{S}$ must be a basis. ■

***Example* 2.43.** The space F^∞ is not finite dimensional, therefore it is not finitely generated either. To see this suppose to the contrary that F^∞ is finite dimensional. Let $n := \dim F^\infty$. But this leads to a contradiction, since F^∞ contains the following linearly independent list of $n+1$ vectors

$$(1,0,0,\dots),(0,1,0,\dots),\dots,(0,0,\dots,0,\overset{\overset{n+1\text{-th}}{\downarrow}}{1},0,\dots).$$

Theorem 2.44. *Suppose that W is a subspace of a finite dimensional vector space V. Then*

(i) *W is finite dimensional, and* $\dim W \leq \dim V$.

(ii) *If* $\dim W = \dim V$ *then* $W = V$.

Proof. **(i)** Let $n = \dim V$. Every linearly independent set $\mathcal{A} \subset W$ is also a linearly independent subset of V. Therefore $\mathcal{A}$ has at most n elements. Let $\mathcal{B}$ be a linearly independent subset of W that has the greatest number of elements. Note that there might be several linearly independent sets with the greatest number of elements, but we only need one of them.

Now let $w \in W$. If $w \notin \operatorname{span}(\mathcal{B})$ then $\mathcal{B} \cup \{w\}$ is linearly independent by Proposition 2.30. Also note that $w \notin \mathcal{B}$, since $\mathcal{B} \subset \operatorname{span}(\mathcal{B})$. Thus $\mathcal{B} \cup \{w\}$ is a subset of W that has more elements than $\mathcal{B}$, which is a contradiction. So we must have $w \in \operatorname{span}(\mathcal{B})$. Hence $W \subset \operatorname{span}(\mathcal{B})$, since w was arbitrary. On the other hand, by Proposition 2.20 we have $\operatorname{span}(\mathcal{B}) \subset W$, since W is a subspace containing $\mathcal{B}$. Therefore $\mathcal{B}$ spans W, and as it is linearly independent, $\mathcal{B}$ is a finite basis for W. Finally note that $\mathcal{B}$ has at most n elements, so $\dim W \leq n$.

(ii) Let $n = \dim W = \dim V$. Let $\mathcal{B}$ be a basis of W. Then $\mathcal{B}$ is a linearly independent subset of V that has n elements. Hence $\mathcal{B}$ is a basis for V. Therefore $V = \operatorname{span}(\mathcal{B}) = W$. ■

Proposition 2.45. *Let $A, B \in F^{m\times n}$, and suppose B is the reduced row echelon form of A. Let $B_{1,j_1},\dots B_{k,j_k}$ be the leading entries of the nonzero rows of B. Then the columns $A_{\cdot,j_1},\dots,A_{\cdot,j_k}$ are linearly independent, and form a basis for* $\operatorname{span}(A_{\cdot,1},\dots,A_{\cdot,n})$. *In addition for every $j \leq n$ we have*

$$A_{\cdot,j} = \sum_{l\leq k} B_{lj} A_{\cdot,j_l}.$$

Also, for every $i \leq k$ the columns $A_{.,j_1}, \ldots, A_{.,j_{i-1}}$ form a basis for

$$\operatorname{span}(A_{.,1}, \ldots, A_{.,j_i-1}).$$

Remark. In other words, if $j < j_i$ then $A_{.,j}$ is a linear combination of $A_{.,j_1}, \ldots, A_{.,j_{i-1}}$. (Note the difference between j_{i-1} and $j_i - 1$.) In particular we have $A_{.,j} = 0$ when $j < j_1$.

Remark. Note that as a consequence of this proposition we get

$$k = \dim \operatorname{span}(A_{.,1}, \ldots, A_{.,n}).$$

Proof. We know that $B = EA$ for some invertible matrix $E \in F^{m \times m}$. We also know that $j_1 < j_2 < \cdots < j_k \leq n$. By Proposition 1.34, we also know that $B_{.,j_i} = e_i$ for every $i \leq k$, and when $j < j_i$ we have

$$B_{.,j} = [B_{1j}, B_{2j}, \ldots, B_{i-1,j}, 0, \ldots, 0]^\intercal = \sum_{l \leq i-1} B_{lj} e_l = \sum_{l \leq i-1} B_{lj} B_{.,j_l}. \qquad (*)$$

Note that when $j < j_1$ we have $B_{.,j} = 0$. Also when $j \geq j_k$ we have to set $i = k+1$ in the above equation.

Now suppose $\sum_{i \leq k} a_i A_{.,j_i} = 0$ for some $a_i \in F$. Then we have

$$\begin{aligned} \sum_{i \leq k} a_i e_i &= \sum_{i \leq k} a_i B_{.,j_i} = \sum_{i \leq k} a_i (EA)_{.,j_i} \\ &= \sum_{i \leq k} a_i E A_{.,j_i} = E\Big(\sum_{i \leq k} a_i A_{.,j_i} \Big) = E0 = 0. \end{aligned}$$

Thus $a_i = 0$ for every i. Hence $A_{.,j_1}, \ldots, A_{.,j_k}$ are linearly independent. Next suppose $j < j_i$. Then if we multiply the equation $(*)$ by E^{-1} we get

$$\begin{aligned} A_{.,j} = (E^{-1}B)_{.,j} = E^{-1} B_{.,j} &= E^{-1}\Big(\sum_{l \leq i-1} B_{lj} B_{.,j_l} \Big) \\ = \sum_{l \leq i-1} B_{lj} E^{-1} B_{.,j_l} = \sum_{l \leq i-1} B_{lj} (E^{-1}B)_{.,j_l} &= \sum_{l \leq i-1} B_{lj} A_{.,j_l}. \qquad (**) \end{aligned}$$

Hence $A_{.,j_1}, \ldots, A_{.,j_{i-1}}$ generate $\operatorname{span}(A_{.,1}, \ldots, A_{.,j_i-1})$. So they form a basis for $\operatorname{span}(A_{.,1}, \ldots, A_{.,j_i-1})$, since they are linearly independent. If we simply assume that $j \leq n$ then we can set $i = k+1$ in the above equation, and conclude that $A_{.,j_1}, \ldots, A_{.,j_k}$ is a basis for $\operatorname{span}(A_{.,1}, \ldots, A_{.,n})$. Also note that when $j < j_1$ the above equation becomes $A_{.,j} = (E^{-1}B)_{.,j} = E^{-1}B_{.,j} = E^{-1}0 = 0$.

Now let $j \leq n$. Suppose $j < j_i$ for some i; or $j > j_k$, in which case we set $i = k+1$. Then the equation $(**)$ gives us

$$A_{.,j} = \sum_{l \leq i-1} B_{lj} A_{.,j_l} = \sum_{l \leq k} B_{lj} A_{.,j_l},$$

since $B_{lj} = 0$ for $l \geq i$. ∎

Theorem 2.46. *The reduced row echelon form of a matrix is unique.*

Remark. Let $A \in F^{m\times n}$. Remember that the reduced row echelon form of A is a matrix $B \in F^{m\times n}$ which is in reduced row echelon form, such that we have $B = EA$ for some invertible matrix $E \in F^{m\times m}$. Also note that the uniqueness of the reduced row echelon form of A means that if $B, B' \in F^{m\times n}$ are two matrices in reduced row echelon form, and $E, E' \in F^{m\times m}$ are invertible matrices such that $B = EA$, $B' = E'A$, then $B = B'$. But it is not true that E is also uniquely determined by A.

Proof. Let $A, B \in F^{m\times n}$, and suppose B is the reduced row echelon form of A. Let $B_{1,j_1}, \dots B_{k,j_k}$ be the leading entries of the nonzero rows of B. We know that $j_1 < j_2 < \cdots < j_k \leq n$. By Proposition 1.34, we also know that $B_{\cdot,j_i} = e_i$ for every $i \leq k$. Furthermore, the previous proposition implies that $A_{\cdot,j_1}, \dots, A_{\cdot,j_k}$ are linearly independent, and form a basis for $\operatorname{span}(A_{\cdot,1}, \dots, A_{\cdot,n})$. Also, for every $i \leq k$ the columns $A_{\cdot,j_1}, \dots, A_{\cdot,j_i-1}$ form a basis for $\operatorname{span}(A_{\cdot,1}, \dots, A_{\cdot,j_i-1})$. In addition we have $A_{\cdot,j} = 0$ when $j < j_1$.

Now let us show that B is uniquely determined by A. First note that $j_1, \dots, j_k$ are uniquely determined by A. We prove this by induction. We know that $A_{\cdot,1} = A_{\cdot,2} = \cdots = A_{\cdot,j_1-1} = 0$. But $A_{\cdot,j_1}$ is linearly independent, so it is nonzero. Thus $A_{\cdot,j_1}$ is the first nonzero column of A. Suppose we have shown that $j_1, \dots, j_{i-1}$ are uniquely determined by A. Then we know that $A_{\cdot,j} \in \operatorname{span}(A_{\cdot,j_1}, \dots, A_{\cdot,j_{i-1}})$ for $j < j_i$. On the other hand, $A_{\cdot,j_1}, \dots, A_{\cdot,j_i}$ are linearly independent, so $A_{\cdot,j_i} \notin \operatorname{span}(A_{\cdot,j_1}, \dots, A_{\cdot,j_{i-1}})$. Therefore $A_{\cdot,j_i}$ is the first column after $A_{\cdot,j_1}, \dots, A_{\cdot,j_{i-1}}$ that does not belong to their span. At the end, after we have determined j_k, we have $A_{\cdot,j} \in \operatorname{span}(A_{\cdot,j_1}, \dots, A_{\cdot,j_k})$ for every $j \leq n$.

Hence $j_1, \dots, j_k$ are uniquely determined by A. In particular k is uniquely determined by A. So in the matrix B, the positions of the columns $B_{\cdot,j_1}, \dots, B_{\cdot,j_k}$, which are equal to $e_1, \dots, e_k$, are uniquely determined by A. Now let $j \leq n$ be an index different from $j_1, \dots, j_k$. By the previous proposition we have

$$A_{\cdot,j} = \sum_{l\leq k} B_{lj} A_{\cdot,j_l}.$$

This equation implies that $B_{1j}, \dots, B_{kj}$ are the coefficients of the expansion of $A_{\cdot,j}$ as a linear combination of $A_{\cdot,j_1}, \dots, A_{\cdot,j_k}$. But these coefficients are uniquely determined by the columns of A, because $A_{\cdot,j_1}, \dots, A_{\cdot,j_k}$ is a basis for $\operatorname{span}(A_{\cdot,1}, \dots, A_{\cdot,n})$. In addition we have $B_{lj} = 0$ for $l > k$, since the rows of B below the k-th row are zero. Thus the entries of the columns $B_{\cdot,j}$ are uniquely determined by A. Hence B is uniquely determined by A. ∎

Theorem 2.47. *Suppose $w_1, \dots, w_m \in F^n$. Let $A \in F^{n\times m}$ be the matrix whose j-th column is w_j. Let $B \in F^{n\times m}$ be the reduced row echelon form of A. Suppose*

$B_{\cdot,j_1}, \ldots, B_{\cdot,j_k}$ are the columns of B that contain a leading entry. Then $w_{j_1}, \ldots, w_{j_k}$ is a basis for $\operatorname{span}(w_1, \ldots, w_m)$.

Remark. This theorem provides us an algorithm to find a basis for F^n among the vectors $w_1, \ldots, w_m$, by solving only one system of linear equations.

Proof. This theorem is a trivial consequence of Proposition 2.45. ■

Theorem 2.48. *Suppose $v_1, \ldots, v_k \in F^n$ are linearly independent, and $w_1, \ldots, w_m$ span F^n. Let $A \in F^{n\times(k+m)}$ be the matrix whose j-th column is v_j when $j \leq k$, and its $(k+i)$-th column is w_i when $i \leq m$. Let $B \in F^{n\times(k+m)}$ be the reduced row echelon form of A. Then B has exactly $n-k$ columns $B_{\cdot,j_1}, \ldots, B_{\cdot,j_{n-k}}$ that contain a leading entry, and satisfy $j_i > k$. In addition, $v_1, \ldots, v_k, w_{j_1}, \ldots, w_{j_{n-k}}$ is a basis for F^n.*

Remark. This theorem provides us an algorithm to extend the list $v_1, \ldots, v_k$ to a basis for F^n by using the elements of the list $w_1, \ldots, w_m$. Also note that in this algorithm we only need to solve one system of linear equations.

Proof. First note that $k \leq n$. Now since $w_1, \ldots, w_m \in \operatorname{span}(A_{\cdot,1}, \ldots, A_{\cdot,k+m}) \subset F^n$, and $w_1, \ldots, w_m$ generate F^n, we have

$$\dim \operatorname{span}(A_{\cdot,1}, \ldots, A_{\cdot,k+m}) = n.$$

Thus the number of leading entries of B are n, due to the Proposition 2.45. On the other hand $A_{\cdot,j} = v_j$ for $j \leq k$. So $A_{\cdot,1}, \ldots, A_{\cdot,k}$ are linearly independent, and thus they form a basis for their own span. Therefore $A_{\cdot,j}$ does not belong to the span of $A_{\cdot,1}, \ldots, A_{\cdot,j-1}$ for $j \leq k$. Hence as shown in the proof of Theorem 2.46, $B_{\cdot,1}, \ldots, B_{\cdot,k}$ must contain leading entries. Thus the other $n-k$ leading entries of B belong to the columns $B_{\cdot,j}$ for some $j > k$. Now the theorem follows trivially from Proposition 2.45. ■

2.5 Sums and Direct Sums of Subspaces

Definition 2.49. Suppose $W_1, \ldots, W_k$ are subspaces of a vector space V. Then their **sum** denoted by

$$W_1 + \cdots + W_k,$$

is the subspace generated by $\bigcup_{i=1}^{k} W_i$.

Remark. Note that we have only defined the sum of several subspaces of some given vector space, not the sum of several arbitrary and unrelated vector spaces. In fact, the above definition is meaningless when W_i's are not subspaces of a larger space V.

Remark. Also note that the sum of several subspaces does not depend on their order, since the union of several sets does not depend on the order of sets.

Theorem 2.50. *Suppose $W_1, \dots, W_k$ are subspaces of V. Then we have*

$$W_1 + \cdots + W_k = \{v_1 + \cdots + v_k : \textit{for every } v_1 \in W_1, \dots, v_k \in W_k\}.$$

Proof. Let $W := \{v_1 + \cdots + v_k : v_1 \in W_1, \dots, v_k \in W_k\}$. Then every element of W is a linear combination of some vectors in $\bigcup_{i=1}^{k} W_i$. Thus W is in the span of $\bigcup_{i=1}^{k} W_i$, i.e. $W \subset W_1 + \cdots + W_k$. To prove the equality, it suffices to show that W is a subspace, since W obviously contains each W_j, and $W_1 + \cdots + W_k$ is the smallest subspace that contains each W_j. Now let $\sum v_j, \sum u_j \in W$ where $v_j, u_j \in W_j$ for each j, and let $a \in F$. Then we have

$$\sum_{j \le k} v_j + a \sum_{j \le k} u_j = \sum_{j \le k} (v_j + a u_j) \in W,$$

since $v_j + a u_j \in W_j$ for each j. Hence W is a subspace as desired. ■

Theorem 2.51. *Suppose W_1, W_2 are finite dimensional subspaces of V. Then $W_1 + W_2$ and $W_1 \cap W_2$ are also finite dimensional subspaces, and we have*

$$\dim(W_1 + W_2) = \dim W_1 + \dim W_2 - \dim(W_1 \cap W_2).$$

Proof. First note that $W_1 \cap W_2$ is finite dimensional, since it is a subspace of W_1. Let $w_1, \dots, w_k$ be a basis for $W_1 \cap W_2$. Now $w_1, \dots, w_k$ is a linearly independent subset of W_1, so we can extend it to a basis for W_1. Let us denote this basis by $w_1, \dots, w_k, u_1, \dots, u_n$. Similarly we can extend $w_1, \dots, w_k$ to a basis for W_2, and we denote it by $w_1, \dots, w_k, v_1, \dots, v_m$. We claim that $w_1, \dots, w_k, u_1, \dots, u_n, v_1, \dots, v_m$ is a basis for $W_1 + W_2$.

Let $u + v$ be an arbitrary vector in $W_1 + W_2$, where $u \in W_1$ and $v \in W_2$. Then there are scalars $a_i, b_i, c_i, d_i \in F$ such that

$$\begin{aligned} u &= a_1 u_1 + \cdots + a_n u_n + c_1 w_1 + \cdots + c_k w_k, \\ v &= b_1 v_1 + \cdots + b_m v_m + d_1 w_1 + \cdots + d_k w_k. \end{aligned}$$

Then we have

$$\begin{aligned} u + v = a_1 u_1 + \cdots + a_n u_n + b_1 v_1 + \cdots + b_m v_m \\ + (c_1 + d_1) w_1 + \cdots + (c_k + d_k) w_k. \end{aligned}$$

Hence $w_1, \dots, w_k, u_1, \dots, u_n, v_1, \dots, v_m$ span $W_1 + W_2$. In particular $W_1 + W_2$ is finite dimensional.

Now let us show that $w_1, \ldots, w_k, u_1, \ldots, u_n, v_1, \ldots, v_m$ is linearly independent. Suppose for some $a_i, b_i, c_i \in F$ we have

$$a_1u_1 + \cdots + a_nu_n + b_1v_1 + \cdots + b_mv_m + c_1w_1 + \cdots + c_kw_k = 0.$$

Then we have

$$u := a_1u_1 + \cdots + a_nu_n = -b_1v_1 - \cdots - b_mv_m - c_1w_1 - \cdots - c_kw_k. \qquad (*)$$

But $u \in \operatorname{span}(u_1, \ldots, u_n) \subset W_1$, and $u \in \operatorname{span}(w_1, \ldots, w_k, v_1, \ldots, v_m) = W_2$. Hence $u \in W_1 \cap W_2$. Thus there are $d_i \in F$ such that $u = d_1w_1 + \cdots + d_kw_k$. Therefore

$$a_1u_1 + \cdots + a_nu_n = u = d_1w_1 + \cdots + d_kw_k.$$

Hence

$$a_1u_1 + \cdots + a_nu_n - d_1w_1 - \cdots - d_kw_k = 0.$$

So we must have $a_i = 0$ and $d_j = 0$ for each i, j, since $w_1, \ldots, w_k, u_1, \ldots, u_n$ is linearly independent. Now the equation $(*)$ implies that

$$-b_1v_1 - \cdots - b_mv_m - c_1w_1 - \cdots - c_kw_k = 0,$$

and therefore we must have $b_i = 0$ and $c_j = 0$ for each i, j, since $w_1, \ldots, w_k, v_1, \ldots, v_m$ is linearly independent too. Hence we get the desired.

Finally note that there is no repetition in the list of vectors $w_1, \ldots, w_k, u_1, \ldots, u_n$, $v_1, \ldots, v_m$, since it is a linearly independent list. Thus the number of vectors in this list is $n + m + k$, and hence we have

$$\begin{aligned}\dim(W_1 + W_2) &= n + m + k = n + k + m + k - k \\ &= \dim W_1 + \dim W_2 - \dim(W_1 \cap W_2),\end{aligned}$$

as desired. ∎

Remark. There is no simple formula similar to the above formula, for the dimension of the sum of more than two finite dimensional subspaces.

Definition 2.52. Suppose $W_1, \ldots, W_k$ are subspaces of V. Then the subspaces $W_1, \ldots, W_k$ are said to be **independent** if for every $v_1 \in W_1, \ldots, v_k \in W_k$ we have

$$v_1 + \cdots + v_k = 0 \implies v_i = 0 \text{ for every } i.$$

When $W_1, \ldots, W_k$ are independent subspaces, their sum is called their **direct sum**, and is denoted by

$$W_1 \oplus \cdots \oplus W_k.$$

Remark. Note that the direct sum of several subspaces is nothing but their sum. We just use a different notation for it, to emphasize that the subspaces are independent.

Theorem 2.53. *Suppose $W_1, \ldots, W_k$ are independent subspaces of V. Then for every $v \in W_1 \oplus \cdots \oplus W_k$ there are unique $v_1 \in W_1, \ldots, v_k \in W_k$, such that*

$$v = v_1 + \cdots + v_k.$$

Remark. The uniqueness in the above theorem means that if for some $v_i, u_i \in W_i$ we have

$$v_1 + \ldots + v_k = v = u_1 + \ldots + u_k,$$

then $u_i = v_i$ for each i.

Proof. The existence of $v_1, \ldots, v_k$ is proved in Theorem 2.50. To prove the uniqueness suppose that for some $v_i, u_i \in W_i$ we have

$$v_1 + \ldots + v_k = v = u_1 + \ldots + u_k.$$

Then we have $(v_1 - u_1) + \cdots + (v_k - u_k) = 0$. But $v_i - u_i \in W_i$ for each i. So for each i we must have $v_i - u_i = 0$, since $W_1, \ldots, W_k$ are independent subspaces. ■

Theorem 2.54. *Suppose W_1, W_2 are subspaces of V. Then W_1, W_2 are independent if and only if $W_1 \cap W_2 = \{0\}$.*

Proof. Suppose $W_1 \cap W_2 = \{0\}$, and we have $v_1 + v_2 = 0$ where $v_j \in W_j$. Then we have $v_1 = -v_2 \in W_2$. Thus $v_1 \in W_1 \cap W_2$. So $v_1 = 0$, and therefore $v_2 = 0$. Hence W_1, W_2 are independent subspaces. Conversely suppose W_1, W_2 are independent subspaces. Let $v \in W_1 \cap W_2$ be an arbitrary vector. Then we have $v + (-v) = 0$. But $v \in W_1 \cap W_2 \subset W_1$ and $-v \in W_1 \cap W_2 \subset W_2$. Hence we must have $v = -v = 0$. Thus $W_1 \cap W_2 = \{0\}$. ■

Remark. There is no simple criterion similar to the above, for the independence of more than two subspaces. For example, consider the subspaces $W_k := \{y = kx\}$ of $\mathbb{R}^2$, for $k = 1, 2, 3$. Then W_1, W_2, W_3 are not independent, since

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2 \\ -4 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and the k-th vector in the above sum belongs to W_k. But it is easy to see that $W_k \cap W_j = \{0\}$ for $k \neq j$.

Theorem 2.55. *Suppose $W_1, \ldots, W_k$ are finite dimensional subspaces of V, and $\mathcal{B}_j$ is a basis for W_j for each j. Let $\mathcal{B} := \bigcup_{j=1}^{k} \mathcal{B}_j$.*

(i) *$W_1 + \cdots + W_k$ is finite dimensional, and $\mathcal{B}$ is a set of generators for it. In addition we have*

$$\dim(W_1 + \cdots + W_k) \leq \sum_{j=1}^{k} \dim W_j.$$

(ii) *If $W_1, \ldots, W_k$ are also independent, then $\mathcal{B}$ is a basis for $W_1 \oplus \cdots \oplus W_k$, and we have*

$$\dim(W_1 \oplus \cdots \oplus W_k) = \sum_{j=1}^{k} \dim W_j.$$

Proof. **(i)** Suppose $\mathcal{B}_j = \{v_{j,1}, \ldots, v_{j,n_j}\}$. Let $W := W_1 + \cdots + W_k$, and let $v \in W$. We know that there are $v_j \in W_j$ for each j, so that $v = \sum_{j \leq k} v_j$. On the other hand, we know that there are $a_{j,i} \in F$ such that $v_j = \sum_{i \leq n_j} a_{j,i} v_{j,i}$ for each j, because $\mathcal{B}_j$ is a basis for W_j. Hence we have $v = \sum_{j \leq k} \sum_{i \leq n_j} a_{j,i} v_{j,i}$. So $\mathcal{B}$ generates W. Thus W is finitely generated, and therefore it is finite dimensional.

In addition note that by Theorem 2.42, the dimension of W is less than or equal to the number of elements of $\mathcal{B}$. Also, note that the number of elements of $\mathcal{B} = \bigcup_{j \leq k} \mathcal{B}_j$ is less than or equal to the sum of the number of elements of each $\mathcal{B}_j$. Hence we get the desired inequality for $\dim W$.

(ii) Let $W := W_1 \oplus \cdots \oplus W_k$. We know that $\mathcal{B}$ generates W. Now suppose $\mathcal{B}_j = \{v_{j,1}, \ldots, v_{j,n_j}\}$. Then $\mathcal{B} = \{v_{1,1}, \ldots, v_{1,n_1}, \ldots, v_{k,n_k}\}$. Suppose we have $\sum_{j \leq k} \sum_{i \leq n_j} a_{j,i} v_{j,i} = 0$, where $a_{j,i} \in F$. Let $v_j := \sum_{i \leq n_j} a_{j,i} v_{j,i}$. Then we have $\sum_{j \leq k} v_j = 0$, and $v_j \in W_j$ for each j. Thus we must have $v_j = 0$ for each j, since $W_1, \ldots, W_k$ are independent subspaces. Hence for every fixed j we have $\sum_{i \leq n_j} a_{j,i} v_{j,i} = 0$. But $\mathcal{B}_j$ is linearly independent, so $a_{j,i} = 0$ for every i, j. Therefore $\bigcup_{j \leq k} \mathcal{B}_j$ is linearly independent, and consequently it is a basis for W.

Finally note that $\mathcal{B}$ is linearly independent, thus its elements are all distinct. Hence the number of elements of $\mathcal{B} = \bigcup_{j \leq k} \mathcal{B}_j$ is the sum of the number of elements of each $\mathcal{B}_j$. Therefore we get the desired formula for $\dim W$. ∎

Remark. The converse of the part (ii) of the above theorem is also true, as we will see next.

Theorem 2.56. *Suppose $W_1, \ldots, W_k$ are finite dimensional subspaces of V, and $\mathcal{B}_j$ is a basis for W_j for each j. Let $\mathcal{B} := \bigcup_{j=1}^{k} \mathcal{B}_j$.*

(i) *If $\mathcal{B}$ is a basis for V, then $W_1, \ldots, W_k$ are independent subspaces, and we have $V = W_1 \oplus \cdots \oplus W_k$.*

(ii) *Let $W := W_1 + \cdots + W_k$. If $\dim W = \sum_{j=1}^{k} \dim W_j$, then $W_1, \ldots, W_k$ are independent subspaces, and we have $W = W_1 \oplus \cdots \oplus W_k$.*

Proof. **(i)** Suppose $\mathcal{B}_j = \{v_{j,1}, \dots, v_{j,n_j}\}$. Then we have

$$\mathcal{B} = \{v_{1,1}, \dots, v_{1,n_1}, \dots, v_{k,1}, \dots, v_{k,n_k}\}.$$

Let $v \in V$. We know that there are $a_{j,i} \in F$ such that $v = \sum_{j\leq k} \sum_{i\leq n_j} a_{j,i} v_{j,i}$. Set $v_j := \sum_{i\leq n_j} a_{j,i} v_{j,i}$. Then we have $v = \sum_{j\leq k} v_j$, and $v_j \in W_j$ for each j. Thus we have shown that $V = W_1 + \cdots + W_k$.

Now suppose $\sum_{j\leq k} v_j = 0$, where $v_j \in W_j$ for each j. Then there are $a_{j,i} \in F$ so that $v_j = \sum_{i\leq n_j} a_{j,i} v_{j,i}$ for each j, because $\mathcal{B}_j$ is a basis for W_j. Therefore we have $\sum_{j\leq k} \sum_{i\leq n_j} a_{j,i} v_{j,i} = 0$. But $\mathcal{B}$ is a basis, so we must have $a_{j,i} = 0$ for every i, j. Hence we have $v_j = \sum_{i\leq n_j} a_{j,i} v_{j,i} = 0$ for every j. Thus $W_1, \dots, W_k$ are independent subspaces, and therefore their sum is a direct sum.

(ii) By previous theorem, we know that $\mathcal{B}$ generates W. On the other hand, the number of elements of $\mathcal{B} = \bigcup_{j\leq k} \mathcal{B}_j$ is less than or equal to the sum of the number of elements of each $\mathcal{B}_j$. Hence the number of elements of $\mathcal{B}$ is less than or equal to the dimension of W. Therefore by Theorem 2.42, $\mathcal{B}$ is a basis for W. Thus by part (i) of the theorem we get the desired result. ■

***Exercise* 2.57.** Suppose W, U are subspaces of V, and $V = W \oplus U$. Also suppose that

$$W = W_1 \oplus \cdots \oplus W_k, \qquad U = U_1 \oplus \cdots \oplus U_l,$$

where $W_1, \dots, W_k$ and $U_1, \dots, U_l$ are subspaces of W and U respectively. Show that

$$V = W_1 \oplus \cdots \oplus W_k \oplus U_1 \oplus \cdots \oplus U_l.$$

Solution. First let us emphasize that k or l can be 1 too. Also, note that each W_i and U_j is also a subspace of V, as shown in Exercise 2.24. Let $v \in V$. Then there are $w \in W$ and $u \in U$ such that $v = w + u$. Consequently there are $w_i \in W_i$ and $u_j \in U_j$ so that

$$w = w_1 + \cdots + w_k, \qquad u = u_1 + \cdots + u_l.$$

Hence $v = w_1 + \cdots + w_k + u_1 + \cdots + u_l$. Therefore

$$V = W_1 + \cdots + W_k + U_1 + \cdots + U_l.$$

Now suppose $w_1 + \cdots + w_k + u_1 + \cdots + u_l = 0$, where $w_i \in W_i$ and $u_j \in U_j$. Let

$$w := w_1 + \cdots + w_k \in W, \qquad u := u_1 + \cdots + u_l \in U.$$

Then we have $w + u = 0$. But W, U are independent subspaces, so we must have $w = u = 0$. Thus we have $w_1 + \cdots + w_k = 0$, and $u_1 + \cdots + u_l = 0$. Therefore we obtain $w_i = 0$ for every i, and $u_j = 0$ for every j; because W and U are the direct sum of $W_1, \dots, W_k$ and $U_1, \dots, U_l$ respectively. Hence $W_1, \dots, W_k, U_1, \dots, U_l$ are independent subspaces of V, and we get the desired. ■

Chapter 3

Linear Maps

3.1 Linear Maps

Definition 3.1. Suppose V and W are two vector spaces over the same field F. A **linear map** is a function $T : V \to W$ that satisfies

(i) $T(u + v) = T(u) + T(v)$ for every $u, v \in V$.

(ii) $T(av) = aT(v)$ for every $v \in V$ and $a \in F$.

Linear maps from V to F are called **(linear) functionals** on V. Linear maps from V to itself are called **(linear) operators** on V.

Remark. We usually denote $T(v)$ by Tv.

Remark. Note that in the relation $T(u + v) = Tu + Tv$, u, v are added using the addition of V, and Tu, Tv are added using the addition of W. So in some sense, we can say that the linear map T transforms the addition of V into the addition of W. Similar remarks apply to the scalar multiplication.

Remark. When we want to emphasize the role of the field F, we will say that T is *F-linear*.

Notation. In the rest of this chapter, we assume that F is a field, V, W are vector spaces over F, and $T : V \to W$ is a linear map.

Proposition 3.2. *Suppose $T : V \to W$ is a linear map. Then*

(i) $T(0) = 0$.

(ii) *For every $v_1, \ldots, v_k \in V$ and $a_1, \ldots, a_k \in F$ we have*

$$T(a_1v_1 + \cdots + a_kv_k) = a_1T(v_1) + \cdots + a_kT(v_k).$$

Remark. The above proposition means that linear maps preserve linear combinations, and the zero vector.

Proof. **(i)** We have

$$T(0) + 0 = T(0) = T(0 + 0) = T(0) + T(0).$$

So $T(0) = 0$. Another easy way to prove this is to note that

$$T(0) = T(0 \cdot 0) = 0 \cdot T(0) = 0.$$

Notice that in the above 0 denotes both the zero scalar and the zero vector.

(ii) By an easy induction on k we can show that $T(\sum_{j\leq k} v_j) = \sum_{j\leq k} T(v_j)$. Then we have $T(\sum_{j\leq k} a_j v_j) = \sum_{j\leq k} T(a_j v_j) = \sum_{j\leq k} a_j T(v_j)$. ■

Remark. The following proposition is an easy criterion to check whether a given map is linear.

Proposition 3.3. *A function $T : V \to W$ is linear if and only if for every $u, v \in V$ and $a \in F$ we have*

$$T(u + av) = Tu + aTv.$$

Proof. When T is linear we have $T(u+av) = Tu+T(av) = Tu+aTv$. Conversely suppose that T satisfies the above property. Then by setting $a = 1$ we get $T(u+v) = Tu + Tv$ for every $u, v \in V$. So we must have $T(0) = 0$, as we have shown in the proof of the last proposition. Now by setting $u = 0$ we get

$$T(av) = T(0 + av) = T(0) + aTv = 0 + aTv = aTv.$$

Thus T is linear. ■

***Example* 3.4.** Let $A \in F^{m\times n}$. Then the function

$$\begin{array}{c} F^n \longrightarrow F^m \\ x \mapsto Ax \end{array}$$

is a linear map. Note that we consider the vectors of F^n, F^m as column vectors, so the action of A on x is just matrix multiplication. This example is the prototype example of linear maps between finite dimensional vector spaces, as we will see later. Another interesting linear map, similar to the above one, is the following

$$\begin{array}{c} F^{n\times k} \longrightarrow F^{m\times k} \\ B \mapsto AB \end{array}.$$

***Example* 3.5.** The following functions are linear maps

$$\begin{array}{c} V \to W \\ v \mapsto 0 \end{array}, \qquad \begin{array}{c} V \to V \\ v \mapsto v \end{array}.$$

The first function is called the *zero linear map*, and is usually denoted by 0. And the second function is the *identity map* of V, which is usually denoted by I_V. When V is clear from the context, we simply denote I_V by I.

***Example* 3.6.** Consider the following functions on F^∞

$$(a_1, a_2, \dots) \mapsto (a_2, a_3, \dots),$$
$$(a_1, a_2, \dots) \mapsto (0, a_1, a_2, \dots).$$

These functions are both linear. They are called the backward shift, and the forward shift, respectively.

***Example* 3.7.** Let $\mathcal{S}$ be a nonempty set, and let W be a vector space over the field F. Then the space of functions from $\mathcal{S}$ into W, i.e.

$$W^{\mathcal{S}} := \{f : f : \mathcal{S} \to W\},$$

is a vector space over F. The addition and scalar multiplication on this space are defined as follows

$$(f + g)(s) := f(s) + g(s), \qquad (af)(s) := af(s),$$

where $f, g \in W^{\mathcal{S}}$ and $a \in F$. Note that $f + g$ and af are functions, so in order to define them we have to specify their values at every $s \in \mathcal{S}$. We leave it as an exercise, to check that $W^{\mathcal{S}}$ is indeed a vector space with these operations. We only mention that the zero of this vector space is the zero function, i.e. the function that maps every $s \in \mathcal{S}$ to $0 \in W$. Also, the additive inverse of a function f is the function $(-f)(s) := -f(s)$.

Now suppose that $\mathcal{S} = V$ is also a vector space over F. Let $\mathcal{L}(V, W)$ be the set of all linear maps from V to W. Let us show that $\mathcal{L}(V, W)$ is a subspace of $W^{\mathcal{S}}$. First note that the zero function is a linear map, so it belongs to $\mathcal{L}(V, W)$. Next assume that $T, S \in \mathcal{L}(V, W)$ and $a \in F$. We know that

$$(T + S)(v) := Tv + Sv, \qquad (aT)(v) := aTv,$$

for all $v \in V$. We only need to show that $T + S, aT$ belong to $\mathcal{L}(V, W)$, i.e. they are linear maps too. Suppose that $u, v \in V$ and $b \in F$. Then we have

$$\begin{aligned}(T + S)(u + bv) &= T(u + bv) + S(u + bv) = Tu + bTv + Su + bSv \\ &= Tu + Su + b(Tv + Sv) = (T + S)(u) + b(S + T)(v), \\ (aT)(u + bv) &= a\big(T(u + bv)\big) = a(Tu + bTv) = aTu + abTv \\ &= aTu + baTv = (aT)(u) + b(aT)(v).\end{aligned}$$

Thus $T + S, aT$ are linear. Hence $\mathcal{L}(V, W)$ is a subspace of $W^{\mathcal{S}}$, and therefore $\mathcal{L}(V, W)$ is itself a vector space over the field F.

Definition 3.8. We call $\mathcal{L}(V, W)$ the **space of linear maps** from V to W. When $W = V$ we denote this space by $\mathcal{L}(V)$.

Definition 3.9. Suppose U, V, W are vector spaces over F. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, U)$. Then the **product** of S, T is their composition, i.e.

$$ST := S \circ T : V \to U.$$

Remark. We will show that ST is a linear map, so we have $ST \in \mathcal{L}(V, U)$. Also note that by definition for all $v \in V$ we have

$$(ST)(v) = S(Tv).$$

Proposition 3.10. *The composition of two linear maps is a linear map.*

Proof. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, U)$. Then for $u, v \in V$ and $a \in F$ we have

$$\begin{aligned} ST(u + av) &= S\big(T(u + av)\big) = S(Tu + aTv) \\ &= S(Tu) + aS(Tv) = STu + aSTv. \end{aligned}$$

Thus ST is linear by the previous proposition. ∎

Proposition 3.11. *Suppose U, V, W, Y are vector spaces over a field F. Let $T, T_1, T_2 \in \mathcal{L}(V, W)$, $S, S_1, S_2 \in \mathcal{L}(W, U)$, and $R \in \mathcal{L}(U, Y)$. Then we have*

(i) $R(ST) = (RS)T$.
(ii) $TI_V = T$, *and* $I_W T = T$.
(iii) $(S_1 + S_2)T = S_1T + S_2T$, *and* $S(T_1 + T_2) = ST_1 + ST_2$.
(iv) $(aS)T = a(ST) = S(aT)$, *where* $a \in F$.

Proof. It is easy to check that all the corresponding maps in the proposition have the same domain. Hence in order to show their equality we only need to check that they have the same value at every point. Let $v \in V$ be an arbitrary vector.

(i) We have

$$\big(R(ST)\big)(v) = R\big(ST(v)\big) = R\big(S(Tv)\big) = RS(Tv) = \big((RS)T\big)(v).$$

Thus we get the desired result, since v is arbitrary.

(ii) We have $TI_V(v) = T(I_V v) = T(v)$, and $I_W T(v) = I_W(Tv) = Tv$.

(iii) We have

$$\begin{aligned} \big((S_1 + S_2)T\big)(v) &= (S_1 + S_2)(Tv) = S_1(Tv) + S_2(Tv) \\ &= S_1T(v) + S_2T(v) = (S_1T + S_2T)(v). \end{aligned}$$

Note that in the above formula we only used the definition of composition of maps, and the definition of their addition. Now we have

$$\begin{aligned} \big(S(T_1 + T_2)\big)(v) &= S\big((T_1 + T_2)(v)\big) = S(T_1v + T_2v) \\ &= S(T_1v) + S(T_2v) = ST_1(v) + ST_2(v) = (ST_1 + ST_2)(v). \end{aligned}$$

This time, along with the definitions of composition and addition of maps, we also used the linearity of S.

(iv) We have

$$\big((aS)T\big)(v) = (aS)(Tv) = a\big(S(Tv)\big) = a\big(ST(v)\big) = \big(a(ST)\big)(v).$$

Note that here we only used the definitions of composition of maps and their scalar product. Now we have

$$\big(S(aT)\big)(v) = S\big((aT)(v)\big) = S\big(a(Tv)\big) = a\big(S(Tv)\big) = a\big(ST(v)\big) = \big(a(ST)\big)(v).$$

This time, in addition to the definitions of composition and scalar product of maps, we also used the linearity of S. ■

Remark. Most of the properties listed in the above proposition also hold for the composition of arbitrary functions. The exceptions are $S(T_1 + T_2) = ST_1 + ST_2$ and $S(aT) = a(ST)$. In the proof of these two properties we have used the linearity of S. These two properties make the analogy between multiplication of scalars and composition of linear maps complete. Hence we call the composition of two linear maps, their product.

Remark. Note that we can multiply any two operators in $\mathcal{L}(V)$. Thus the above proposition implies that $\mathcal{L}(V)$ is an algebra over the field F. See Section A.5 for details.

***Exercise* 3.12.** Show that if $\dim V > 1$ then the multiplication on $\mathcal{L}(V)$ is not commutative.

Definition 3.13. We say two linear operators $T, S \in \mathcal{L}(V)$ **commute** if

$$TS = ST.$$

Definition 3.14. Let $T \in \mathcal{L}(V, W)$ be a linear map. We say T is **invertible** if it is invertible as a function, i.e. if it is one-to-one and onto.

Remark. Remember that a function $f : V \to W$ is invertible if and only if there exists a function $g : W \to V$ such that $g \circ f = I_V$ and $f \circ g = I_W$. Then we say g is the *inverse* of f, and we denote it by f^{-1}. Note that the inverse of an invertible function is uniquely determined by that function.

Remark. We will show that when $T \in \mathcal{L}(V, W)$ is invertible, then the function $T^{-1} : W \to V$ is also a linear map, so we have $T^{-1} \in \mathcal{L}(W, V)$. A particular case is when $W = V$. In this case T^{-1} satisfies

$$TT^{-1} = I_V = T^{-1}T.$$

In other words, T^{-1} is the inverse of T with respect to the multiplication of $\mathcal{L}(V)$. Hence T is an invertible element of $\mathcal{L}(V)$. Conversely if $T \in \mathcal{L}(V)$ is an invertible element, i.e. if there is $S \in \mathcal{L}(V)$ such that $TS = I_V = ST$, then by the previous remark T is also invertible as a function, and we have $T^{-1} = S$.

Proposition 3.15. *The inverse of an invertible linear map is a linear map.*

Proof. Suppose $T \in \mathcal{L}(V, W)$ is invertible. Then $T^{-1} : W \to V$. Let $u, v \in W$ and $a \in F$. Then we have

$$T(T^{-1}(u + av)) = u + av = TT^{-1}u + aTT^{-1}v = T(T^{-1}u + aT^{-1}v).$$

But T is one-to-one, since it is invertible. Thus we have $T^{-1}(u + av) = T^{-1}u + aT^{-1}v$. Therefore T^{-1} is linear. ∎

***Example* 3.16.** Suppose we want to find out whether there is a linear map $T \in \mathcal{L}(\mathbb{R}^2)$ such that

$$T\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix},\ T\begin{bmatrix}2\\1\end{bmatrix} = \begin{bmatrix}1\\3\end{bmatrix},\ T\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}.$$

To answer this question, note that we have

$$\begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix},\ \text{but } T\begin{bmatrix}2\\1\end{bmatrix} - T\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1\\3\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}0\\2\end{bmatrix} \neq T\begin{bmatrix}1\\1\end{bmatrix}.$$

Therefore there is no linear map with the specified values at those given points.

Remark. The reason that the linear map T failed to exist in the above example, is that we have a nontrivial linear relation between the given points in the domain, while that linear relation does not hold between the specified values at those given points. Now we can ask the question that if the specified values also satisfied that linear relation, would there be a linear map which took those specified values? The answer to this question is positive. But formulating a general result about this problem is complicated, and its applications are few. So instead of that, we will assume that there is no nontrivial linear relation between the given points in the domain at all, i.e. we will assume that they are linearly independent. Then we can show that there is always a linear map that takes the specified values on a given linearly independent set of vectors in the domain. This is the subject of the next theorem.

Theorem 3.17. *Suppose $v_1, \ldots, v_n$ is a basis for V, and $w_1, \ldots, w_n \in W$ are arbitrary vectors. Then there is a unique linear map $T \in \mathcal{L}(V, W)$ such that for every j we have $Tv_j = w_j$.*

Remark. The above theorem implies that a linear map is uniquely determined by its values at the elements of some basis. It also provides us with a method for constructing linear maps. We just need to choose a basis for V, and assign some arbitrary values to the elements of the basis.

Proof. Let $v \in V$ be an arbitrary vector. Then there are uniquely determined scalars $a_1, \dots, a_n \in F$ such that $v = a_1 v_1 + \cdots + a_n v_n$, since $v_1, \dots, v_n$ is a basis for V. Now we define

$$Tv := a_1 w_1 + \cdots + a_n w_n \in W.$$

Then we have a function $T : V \to W$. First note that T is well defined, i.e. its value at every v is uniquely determined by v, because the scalars $a_1, \dots, a_n$ are uniquely determined by v. It is also obvious that our definition implies that $Tv_j = w_j$ for every j, since we have $v_j = 0v_1 + \cdots + 0v_{j-1} + 1v_j + 0v_{j+1} + \cdots + 0v_n$.

Let $v, u \in V$ and $a \in F$. Then we have $v = \sum_{j \le n} a_j v_j$ and $u = \sum_{j \le n} b_j v_j$, for some $a_j, b_j \in F$. Now we have

$$u + av = \sum_{j \le n} b_j v_j + a \sum_{j \le n} a_j v_j = \sum_{j \le n} (b_j + a a_j) v_j.$$

Note that $b_j + a a_j$'s are the coefficients of the expansion of $u + av$ as a linear combination of v_j's, since these coefficients are uniquely determined by every vector in V. Therefore we have

$$T(u + av) = \sum_{j \le n} (b_j + a a_j) w_j = \sum_{j \le n} b_j w_j + a \sum_{j \le n} a_j w_j = Tu + aTv.$$

Hence T is linear as desired.

Finally we have to show that T is unique. Note that the uniqueness of $a_1, \dots, a_n$ in the first paragraph of the proof does not imply that T is the only linear map that satisfies $Tv_j = w_j$. Because there might be other methods to construct such linear map, and we do not know a priori that those methods will produce the same linear map as T. So suppose that $S \in \mathcal{L}(V, W)$ satisfies $Sv_j = w_j$ for every j. Then for every $v \in V$ there are $a_1, \dots, a_n \in F$ such that $v = a_1 v_1 + \cdots + a_n v_n$. Hence we have

$$\begin{aligned} Sv &= S(a_1 v_1 + \cdots + a_n v_n) = a_1 Sv_1 + \cdots + a_n Sv_n \\ &= a_1 w_1 + \cdots + a_n w_n = a_1 Tv_1 + \cdots + a_n Tv_n = T(a_1 v_1 + \cdots + a_n v_n) = Tv. \end{aligned}$$

Therefore we must have $S = T$, and so T is unique. ∎

Remark. Suppose that in the above theorem V is finite dimensional, and $v_1, \dots, v_n$ are linearly independent, but they do not generate V. Then we can extend them

to a basis $v_1, \dots, v_n, u_1, \dots, u_k$ for V. We can also extend the list $w_1, \dots, w_n$ to $w_1, \dots, w_n, \tilde{w}_1, \dots, \tilde{w}_k$, where $\tilde{w}_i \in W$ are arbitrary vectors. Now we can apply the above theorem and conclude that there is a linear map T such that $Tv_j = w_j$ and $Tu_i = \tilde{w}_i$. Hence we have shown that a linear map T exists that satisfies $Tv_j = w_j$. But note that if W is nonzero then T is not unique, since we have freedom in choosing $\tilde{w}_1, \dots, \tilde{w}_k$. We also have freedom in choosing $u_1, \dots, u_k$ as we have seen in Exercise 2.40.

3.2 Null Spaces and Images

Definition 3.18. Let $T : V \to W$ be a linear map. The **null space**, or the **kernel**, of T is the set

$$\operatorname{null} T = \operatorname{null}(T) := \{v \in V : Tv = 0\}.$$

Also, the **image** of T is the set

$$T(V) := \{Tv : v \in V\}.$$

Remark. Let us review some terminology about functions. Consider a function $T : V \to W$. Then V is called the **domain** of T, and W is called the **codomain** of T. Note that the codomain of T is not necessarily equal to the image of T, since T might not be onto. In addition to the above terms, there is also the term "*range* of T". Depending on the text, it can mean the codomain of T, or the image of T. However, in more recent texts, it usually means the image of T. In these notes, we will not use it though, to avoid any possible confusion.

Proposition 3.19. *Let $T : V \to W$ be a linear map.*

(i) *The null space of T is a subspace of V.*

(ii) *If U is a subspace of V, then $T(U)$ is a subspace of W. In particular the image of T is a subspace of W.*

Proof. **(i)** First note that $0 \in \operatorname{null} T$, because $T0 = 0$. Now let $u, v \in \operatorname{null} T$ and $a \in F$. Then we have $Tu = 0 = Tv$. Hence $T(u + av) = Tu + aTv = 0 + a0 = 0$. Thus $u + av \in \operatorname{null} T$, and therefore $\operatorname{null} T$ is a subspace.

(ii) First note that $0 \in T(U)$, because $T0 = 0$ and $0 \in U$. Now let $w_1, w_2 \in T(U)$ and $a \in F$. Then there are $u_1, u_2 \in U$ such that $Tu_j = w_j$. Hence we have $T(u_1 + au_2) = Tu_1 + aTu_2 = w_1 + aw_2$. But $u_1 + au_2 \in U$, since U is a subspace. Therefore $w_1 + aw_2 \in T(U)$. Thus $T(U)$ is a subspace. Finally note that the image of T is $T(V)$, and V is a subspace of V. ∎

Remark. It is obvious from the definition that T is onto if and only if $T(V) = W$. Note that we do not use the linearity of T here. But we will use the linearity of T to prove the following result about $\operatorname{null} T$. It is one of the reasons that the concept of null space is useful when we work with linear maps.

Proposition 3.20. *A linear map is one-to-one if and only if its null space is* $\{0\}$.

Proof. Let $T \in \mathcal{L}(V, W)$. Suppose $\operatorname{null} T = \{0\}$. Then if $Tu = Tv$ for some $u, v \in V$, we get $T(u - v) = Tu - Tv = 0$. So $u - v \in \operatorname{null} T$. Hence $u - v = 0$. Thus $u = v$, and therefore T is one-to-one.

Conversely suppose that T is one-to-one. Let $v \in \operatorname{null} T$. Then we have $Tv = 0 = T0$. So we must have $v = 0$, since T is one-to-one. Hence $\operatorname{null} T = \{0\}$ as desired. ∎

Definition 3.21. Let $T : V \to W$ be a linear map. If $T(V)$ is finite dimensional, then the **rank** of T is $\dim T(V)$. And if $\operatorname{null}(T)$ is finite dimensional, then the **nullity** of T is $\dim \operatorname{null}(T)$.

Theorem 3.22. *Let $T : V \to W$ be a linear map, and suppose V is finite dimensional. Then $T(V), \operatorname{null}(T)$ are finite dimensional, and we have*

$$\dim T(V) + \dim \operatorname{null}(T) = \dim V.$$

Remark. In other words, rank plus nullity equals the dimension of the domain. Some authors refer to this theorem as the *rank-nullity theorem.*

Proof. We know that $\operatorname{null} T$ is finite dimensional, since it is a subspace of V. Let $u_1, \ldots, u_k$ be a basis for $\operatorname{null} T$. Now we can extend this basis of $\operatorname{null} T$ to a basis for V. So suppose $u_1, \ldots, u_k, v_1, \ldots, v_n$ is a basis for V. Now we claim that $Tv_1, \ldots, Tv_n$ is a basis for $T(V)$. Let $w \in T(V)$ be an arbitrary vector. Then there is $v \in V$ such that $w = Tv$. Also, there are $a_i, b_j \in F$ such that $v = \sum_{i \le k} a_i u_i + \sum_{j \le n} b_j v_j$. Hence we have

$$w = Tv = T\Big(\sum_{i \le k} a_i u_i + \sum_{j \le n} b_j v_j\Big) = \sum_{i \le k} a_i T u_i + \sum_{j \le n} b_j T v_j = \sum_{j \le n} b_j T v_j.$$

Note that we used the fact that $Tu_i = 0$ for every i. Thus we have shown that $Tv_1, \ldots, Tv_n$ generate $T(V)$.

Now suppose $\sum_{j \le n} b_j T v_j = 0$ for some $b_j \in F$. Then we have

$$T\Big(\sum_{j \le n} b_j v_j\Big) = \sum_{j \le n} b_j T v_j = 0.$$

Hence $\sum_{j \le n} b_j v_j \in \operatorname{null} T$. Therefore there are $a_i \in F$ so that $\sum_{j \le n} b_j v_j = \sum_{i \le k} a_i u_i$, since $u_1, \ldots, u_k$ is a basis for $\operatorname{null} T$. Thus we have

$$\sum_{j \le n} b_j v_j + \sum_{i \le k} (-a_i) u_i = 0.$$

But this implies that $b_j = 0 = a_i$ for every i, j, since $u_1, \ldots, u_k, v_1, \ldots, v_n$ is a basis for V. Therefore $Tv_1, \ldots, Tv_n$ are linearly independent too, and hence they form a basis for $T(V)$. Thus $T(V)$ is finite dimensional too. Finally note that we have

$$\dim V = n + k = \dim T(V) + \dim \operatorname{null}(T),$$

as desired. ■

Remark. Let $u \in V$, and consider the level set of T containing u, i.e. the set

$$\mathcal{S}_u := \{v \in V : Tv = Tu\}.$$

Then for $v \in \mathcal{S}_u$ we have $T(v - u) = Tv - Tu = 0$. Hence $v - u \in \operatorname{null} T$. On the other hand if $w \in \operatorname{null} T$, then $T(u + w) = Tu + Tw = Tu$. Therefore $\mathcal{S}_u$ is the set of all vectors of the form $u + w$ where $w \in \operatorname{null} T$. To express this we can write

$$\mathcal{S}_u = u + \operatorname{null} T.$$

Thus the level sets of T are translated copies of the null space of T.

Now note that the level sets of any function form a partition of its domain. So we have a partition of V into the level sets of T. Let $n := \dim V$, and $k := \dim \operatorname{null} T$. Then each $\mathcal{S}_u$ is a translated copy of $\operatorname{null} T$, so $\mathcal{S}_u$ is a "k-dimensional object". But every vector in the k-dimensional object $\mathcal{S}_u$ is mapped to Tu by T. Hence, intuitively we can say that T annihilates those k-dimensional objects $\mathcal{S}_u$, and in their place only preserves a vector whose image is Tu. Thus we can think of $T(V)$ as the image of the set of vectors that are preserved by T; and we can think of the translated copies of $\operatorname{null} T$ as the set of vectors which are annihilated by T. Therefore, intuitively, the above theorem means that the amount of vectors preserved by T plus the amount of vectors annihilated by T equals the amount of all vectors in the domain of T.

Proposition 3.23. *Suppose $T \in \mathcal{L}(V, W)$, and $v_1, \ldots, v_n \in V$. Also, suppose U is a finite dimensional subspace of V. Then we have*

(i) $T\big(\operatorname{span}(v_1, \ldots, v_n)\big) = \operatorname{span}(Tv_1, \ldots, Tv_n)$.

(ii) *If $v_1, \ldots, v_n$ are linearly independent, and T is one-to-one, then $Tv_1, \ldots, Tv_n$ are also linearly independent.*

(iii) *$T(U)$ is a finite dimensional subspace of W.*

(iv) *If T is one-to-one, then we also have*

$$\dim T(U) = \dim U.$$

In addition if $u_1, \ldots, u_k$ is a basis for U then $Tu_1, \ldots, Tu_k$ is a basis for $T(U)$.

Proof. **(i)** Let $w \in \operatorname{span}(Tv_1, \ldots, Tv_n)$. Then there are $a_j \in F$ such that

$$w = \sum_{j \le n} a_j T v_j = T\Big(\sum_{j \le n} a_j v_j\Big) \in T\big(\operatorname{span}(v_1, \ldots, v_n)\big).$$

Conversely suppose $w \in T\big(\operatorname{span}(v_1, \ldots, v_n)\big)$. Then there is $v \in \operatorname{span}(v_1, \ldots, v_n)$ such that $w = Tv$. We also have $v = \sum_{j \le n} a_j v_j$ for some $a_j \in F$. Hence we have

$$w = Tv = T\Big(\sum_{j \le n} a_j v_j\Big) = \sum_{j \le n} a_j T v_j \in \operatorname{span}(Tv_1, \ldots, Tv_n).$$

Thus the two sets are equal.

(ii) Suppose $\sum_{j \le n} a_j T v_j = 0$ for some $a_j \in F$. Then we have

$$T\Big(\sum_{j \le n} a_j v_j\Big) = \sum_{j \le n} a_j T v_j = 0.$$

Therefore we must have $\sum_{j \le n} a_j v_j = 0$, since T is one-to-one. Hence we obtain $a_j = 0$ for every j, because $v_1, \ldots, v_n$ are linearly independent. Thus $Tv_1, \ldots, Tv_n$ are linearly independent too.

(iii) Let $u_1, \ldots, u_k$ be a basis for U. Then by part (i) we have

$$T(U) = T\big(\operatorname{span}(u_1, \ldots, u_k)\big) = \operatorname{span}(Tu_1, \ldots, Tu_k).$$

Thus $T(U)$ is finitely generated, hence it is finite dimensional.

(iv) Let $u_1, \ldots, u_k$ be a basis for U. We have seen that $Tu_1, \ldots, Tu_k$ generate $T(U)$. On the other hand, $Tu_1, \ldots, Tu_k$ are linearly independent, because $u_1, \ldots, u_k$ are linearly independent, and T is one-to-one. Therefore $Tu_1, \ldots, Tu_k$ is a basis for $T(U)$. Thus we have $\dim T(U) = k = \dim U$. ∎

Remark. In the above proposition, we have shown that the image of a set of generators for a subspace is a set of generators for the image of that subspace. If in addition the linear map is one-to-one, then the image of a basis for a subspace is a basis for the image of that subspace. It is easy to see that this last assertion, and in fact the parts (ii), (iv) of the proposition, are not true without assuming that the linear map is one-to-one. For example the zero linear map between two nonzero vector spaces, maps every linearly independent set to 0, which is linearly dependent. It also maps the whole domain to the zero subspace, so it does not preserve the dimension either.

Theorem 3.24. *Suppose V, W are finite dimensional vector spaces, and* $\dim V = \dim W$. *Let $T : V \to W$ be a linear map. Then T is one-to-one if and only if it is onto.*

Remark. Remember that a map T is invertible if it is one-to-one and onto. Hence the above theorem implies that if V, W are finite dimensional and have the same dimension, then for $T \in \mathcal{L}(V, W)$ we have

$$T \text{ is invertible} \iff T \text{ is one to one} \iff T \text{ is onto.}$$

Proof. Suppose T is one-to-one. Then $\operatorname{null} T = \{0\}$. Hence $\dim \operatorname{null} T = 0$. Therefore by the previous theorem we must have $\dim T(V) = \dim V = \dim W$. So $T(V)$ is a subspace of W that has the same dimension as W. Hence we have $T(V) = W$, which means that T is onto.

Conversely suppose that T is onto. Then $T(V) = W$. Therefore we have $\dim T(V) = \dim W = \dim V$. Thus we get

$$\dim \operatorname{null} T = \dim V - \dim T(V) = 0.$$

Hence we must have $\operatorname{null} T = \{0\}$, which implies that T is one-to-one. ■

Theorem 3.25. *Suppose V, W are finite dimensional vector spaces, and $\dim V = \dim W$. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, V)$. If $ST = I_V$ then $TS = I_W$, and we have $S = T^{-1}$.*

Proof. Let $v \in \operatorname{null} T$. Then we have $Tv = 0$. On the other hand we have

$$v = I_V v = ST(v) = S(Tv) = S(0) = 0.$$

Hence $\operatorname{null} T = \{0\}$. Thus T is one-to-one, and therefore by the previous theorem T is also onto. So T is invertible, and $T^{-1} \in \mathcal{L}(W, V)$. Hence from $ST = I_V$ we get

$$S = SI_W = S(TT^{-1}) = (ST)T^{-1} = I_V T^{-1} = T^{-1}.$$

Therefore we also have $TS = TT^{-1} = I_W$. ■

Remark. The above two theorems are in particular true when V is finite dimensional, and $W = V$. This case is indeed the most important case of these results.

***Example* 3.26.** Consider the backward and forward shifts $T, S \in \mathcal{L}(F^\infty)$

$$T : (a_1, a_2, \dots) \mapsto (a_2, a_3, \dots),$$
$$S : (a_1, a_2, \dots) \mapsto (0, a_1, a_2, \dots).$$

Then T is onto, but it is not one-to-one. And S is one-to-one, but it is not onto. In addition we have $TS = I_{F^\infty}$, but $ST \neq I_{F^\infty}$ since for example

$$ST(1, 0, 0, \dots) = (0, 0, 0, \dots).$$

Therefore the above two theorems are not true in infinite dimensional vector spaces. Note that this example also shows that the multiplication of linear operators is not commutative.

3.3 Isomorphisms and Coordinates

Definition 3.27. An invertible linear map is called a **(linear) isomorphism**. Two vector spaces are called **isomorphic** if there exist an isomorphism between them.

Remark. From the viewpoint of linear algebra, isomorphic vector spaces are the same. In other words, we cannot distinguish two isomorphic vector spaces with the tools of linear algebra. But those vector spaces may have other differences, that are induced by some other structures. For example $\mathbb{R}^3$ is isomorphic to the space of polynomials with degree less than 3. So as vector spaces, they are completely the same. But on $\mathbb{R}^3$ we have the notion of cross product of vectors, which is not present in the space of polynomials. Also, we have the division algorithm for polynomials, but there is no corresponding notion for the vectors in $\mathbb{R}^3$.

Hence among isomorphic vector spaces, each particular space may have some extra features that make it suitable for some applications. But when we consider them as vector spaces, we regard them as the same space. Nevertheless, ignoring some differences and paying attention to only a few properties, is a useful idea. It helps us to study many different objects at the same time. It also helps us to understand the implications of those few properties, and to not confuse these implications with the specific properties of each particular object.

Theorem 3.28. *We have*

(i) *If two vector spaces are isomorphic, and one of them is finite dimensional, the other one is finite dimensional too.*

(ii) *Two finite dimensional vector spaces are isomorphic if and only if they have the same dimension.*

Proof. **(i)** Suppose V, W are vector spaces, and V is finite dimensional. Let $T \in \mathcal{L}(V, W)$ be an isomorphism. Then T is onto, so we have $T(V) = W$. Now Theorem 3.22 implies that W is finite dimensional, since W is the image of T.

(ii) Suppose V, W are finite dimensional vector spaces. Let $T \in \mathcal{L}(V, W)$ be an isomorphism. Then T is one-to-one and onto. Therefore we have $\operatorname{null} T = \{0\}$ and $T(V) = W$. Hence we have

$$\dim V = \dim T(V) + \dim \operatorname{null} T = \dim T(V) + 0 = \dim W.$$

Conversely suppose that $\dim V = \dim W = n$. Let $v_1, \ldots, v_n$ be a basis for V, and $w_1, \ldots, w_n$ be a basis for W. Then there is a unique $T \in \mathcal{L}(V, W)$ such that $Tv_j = w_j$ for every j. We claim that T is an isomorphism. By Theorem 3.24, it suffices to show that T is onto, since V, W have the same dimension. Let $w \in W$. Then there are $a_1, \ldots, a_n \in F$ such that $w = a_1 w_1 + \cdots + a_n w_n$. Now let $v := a_1 v_1 + \cdots + a_n v_n$. Then we have

$$Tv = T(a_1 v_1 + \cdots + a_n v_n) = a_1 T v_1 + \cdots + a_n T v_n = a_1 w_1 + \cdots + a_n w_n = w.$$

Hence T is onto, and so it is an isomorphism. Thus V, W are isomorphic as desired. ■

Remark. As a consequence of the above theorem, if V is a finite dimensional vector space over a field F, and $\dim V = n$, then V is isomorphic to F^n. So in some sense, all the finite dimensional vector spaces over F are $F, F^2, F^3, \ldots$. But despite this classification, we prefer to study abstract vector spaces instead of the concrete spaces F^n. The reason is that F^n has a standard basis that simplifies studying it, but we prefer to have an understanding of vector spaces that is independent of the choice of basis. In addition, when we study the subspaces of F^n, we do not have a standard choice of basis. Hence we have to treat those subspaces as abstract vector spaces.

On the other hand, an isomorphism between V, F^n, and the concrete nature of F^n, provide us useful tools to study V. We will construct specific isomorphisms between V and F^n later in this section, and we will see some of their applications.

Remark. The above theorem means that the dimension is an *invariant* of vector spaces that completely determines them. Let us elaborate further on this point. An invariant of a class of objects having some structure is something that we have assigned to those objects, such that if two objects are considered the same, i.e. if they are isomorphic, then their assigned invariants are the same. For example here our objects are finite dimensional vector spaces, and their assigned invariants is their dimension. The point of assigning invariants is to be able to distinguish between different objects more easily, because the invariants usually have a simpler nature than the objects themselves. For example the dimension of a vector space is a positive integer, which is much simpler than the vector space itself. And if two vector spaces have different dimensions, then we can be sure that they are not isomorphic.

Now if an invariant has the extra property that whenever two objects' invariants are the same then the two objects are isomorphic, then that invariant completely classifies those objects. The dimension of vector spaces is an invariant with this property, as we have proved above. This kind of invariants are very useful, but they are rare. Another example of an invariant that completely classifies the objects being studied, is the *genus* of orientable surfaces that are closed and connected. The genus of such a surface is the number of holes in it. For example the genus of sphere is 0, and the genus of the surface of a doughnut, which is called a *torus*, is 1. It is a deep theorem of topology that two such surfaces are homeomorphic if and only if they have the same genus. Note that in topology, isomorphisms are called homeomorphisms.

Definition 3.29. Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for the nonzero vector space V. Let $v \in V$. Then we know that there are unique scalars $a_1, \ldots, a_n \in F$ such

that

$$v = a_1 v_1 + \cdots + a_n v_n.$$

The **coordinate vector** of v with respect to $\mathcal{B}$ is the column vector

$$[v]_{\mathcal{B}} := \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \in F^n.$$

The scalars $a_1, \ldots, a_n$ are called the **coordinates** of v with respect to $\mathcal{B}$. Also, the **coordinate isomorphism** with respect to $\mathcal{B}$ is the function

$$\begin{aligned} \phi_{\mathcal{B}} : V &\longrightarrow F^n \\ v &\mapsto [v]_{\mathcal{B}} \end{aligned}.$$

Remark. Remember that we consider a basis to be a list of vectors, so a basis has an order. It is obvious that if we change the order of elements of the basis $\mathcal{B}$, then we have to change the order of coordinates of every vector. Hence the order of the basis is important here.

***Example* 3.30.** Suppose V is finite dimensional, and $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V. Then we have $[v_j]_{\mathcal{B}} = e_j$, since

$$v_j = 0v_1 + \cdots + 0v_{j-1} + 1v_j + 0v_{j+1} + \cdots + 0v_n.$$

***Example* 3.31.** Let $\mathcal{B} = \{e_1, \ldots, e_n\}$ be the standard basis for F^n. Then for any $x = [x_1, \ldots, x_n]^\intercal \in F^n$ we have $x = x_1 e_1 + \cdots + x_n e_n$. Hence $[x]_{\mathcal{B}} = x$.

Proposition 3.32. *Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V. Then the function $\phi_{\mathcal{B}} : V \to F^n$ is a linear isomorphism. Furthermore for every $x = [x_1, \ldots, x_n]^\intercal \in F^n$ we have*

$$\phi_{\mathcal{B}}^{-1}(x) = x_1 v_1 + \cdots + x_n v_n.$$

Proof. Let $u, v \in V$ and $a \in F$. Then there are uniquely determined scalars $a_j, b_j \in F$ such that $u = \sum a_j v_j$ and $v = \sum b_j v_j$. So we have $\phi_{\mathcal{B}}(u) = [a_1, \ldots, a_n]^\intercal$ and $\phi_{\mathcal{B}}(v) = [b_1, \ldots, b_n]^\intercal$. Now we know that $u + av = \sum (a_j + ab_j) v_j$. Hence

$$\begin{aligned} \phi_{\mathcal{B}}(u + av) &= [a_1 + ab_1, \ldots, a_n + ab_n]^\intercal \\ &= [a_1, \ldots, a_n]^\intercal + a[b_1, \ldots, b_n]^\intercal = \phi_{\mathcal{B}}(u) + a\phi_{\mathcal{B}}(v). \end{aligned}$$

Thus $\phi_{\mathcal{B}}$ is linear.

By Theorem 3.24, to show that $\phi_{\mathcal{B}}$ is an isomorphism, it suffices to show that it is one-to-one, because $\dim V = n = \dim F^n$. So suppose $\phi_{\mathcal{B}}(v) = 0$ for some $v \in V$. Then the definition of $\phi_{\mathcal{B}}$ implies that $v = 0v_1 + \cdots + 0v_n = 0$. Thus $\operatorname{null} \phi_{\mathcal{B}} = \{0\}$, and therefore $\phi_{\mathcal{B}}$ is one-to-one. Finally let $x = [x_1, \ldots, x_n]^\intercal \in F^n$. Then for $v := x_1 v_1 + \cdots + x_n v_n$ we have $\phi_{\mathcal{B}}(v) = [x_1, \ldots, x_n]^\intercal = x$. Hence $\phi_{\mathcal{B}}^{-1}(x) = v$ as desired. (Note that this argument also directly shows that $\phi_{\mathcal{B}}$ is onto.) ■

Remark. Suppose U is a subspace of V, and $u_1, \ldots, u_k$ generate U. Let $\mathcal{B}$ be a basis for V, and let $\phi_{\mathcal{B}} : V \to F^n$ be the coordinate isomorphism. Then by Proposition 3.23, the subspace $\phi_{\mathcal{B}}(U)$ is generated by $\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)$. Now we can apply Theorem 2.47 and find a basis $\phi_{\mathcal{B}}(u_{j_1}), \ldots, \phi_{\mathcal{B}}(u_{j_l})$ for $\phi_{\mathcal{B}}(U)$. Then Proposition 3.23 implies that $u_{j_1}, \ldots, u_{j_l}$ form a basis for $\phi_{\mathcal{B}}^{-1}\big(\phi_{\mathcal{B}}(U)\big) = U$, since $\phi_{\mathcal{B}}^{-1}$ is one-to-one as it is invertible.

Similarly, if we want to check the linear independence of $u_1, \ldots, u_k \in V$, we can apply Proposition 2.31 to $\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k) \in F^n$. If $\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)$ are linearly independent, then Proposition 3.23 implies that $u_1, \ldots, u_k$ are also linearly independent, since $\phi_{\mathcal{B}}^{-1}$ is one-to-one. And if $\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)$ are linearly dependent then we must have

$$\phi_{\mathcal{B}}(u_j) = \sum_{i \neq j} a_i \phi_{\mathcal{B}}(u_i),$$

for some $j \leq k$ and some $a_i \in F$. Therefore we have

$$u_j = \phi_{\mathcal{B}}^{-1}\big(\phi_{\mathcal{B}}(u_j)\big) = \phi_{\mathcal{B}}^{-1}\Big(\sum_{i \neq j} a_i \phi_{\mathcal{B}}(u_i)\Big) = \sum_{i \neq j} a_i \phi_{\mathcal{B}}^{-1}\big(\phi_{\mathcal{B}}(u_i)\big) = \sum_{i \neq j} a_i u_i.$$

Thus $u_1, \ldots, u_k$ are also linearly dependent.

We can also check to see if some vector $v \in V$ belongs to the $\operatorname{span}(u_1, \ldots, u_k)$. We just need to apply Proposition 2.23 to the vectors $\phi_{\mathcal{B}}(v)$ and $\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)$. If $\phi_{\mathcal{B}}(v) \in \operatorname{span}\big(\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)\big)$ then repeating the argument in the last paragraph shows that $v \in \operatorname{span}(u_1, \ldots, u_k)$. And if $\phi_{\mathcal{B}}(v) \notin \operatorname{span}\big(\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)\big)$ then we must have $v \notin \operatorname{span}(u_1, \ldots, u_k)$. Since otherwise we would have $v = \sum_{i \leq k} a_i u_i$ for some $a_i \in F$, and this implies

$$\phi_{\mathcal{B}}(v) = \phi_{\mathcal{B}}\Big(\sum_{i \leq k} a_i u_i\Big) = \sum_{i \leq k} a_i \phi_{\mathcal{B}}(u_i) \in \operatorname{span}\big(\phi_{\mathcal{B}}(u_1), \ldots, \phi_{\mathcal{B}}(u_k)\big),$$

which is a contradiction.

The above examples show the power of coordinate isomorphisms. They allow us to transfer questions about an abstract vector space V, into questions about the concrete vector space F^n. Then we can do computations inside F^n, and finally transfer our results back to V. ■

Definition 3.33. Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for the nonzero vector space V, and $\mathcal{C} = \{w_1, \ldots, w_m\}$ is a basis for the nonzero vector space W. Let $T \in \mathcal{L}(V, W)$. Then the **matrix of** T with respect to the bases $\mathcal{B}, \mathcal{C}$ is an $m \times n$ matrix whose j-th column is $[Tv_j]_{\mathcal{C}}$, i.e.

$$[T]_{\mathcal{C}}^{\mathcal{B}} := \Big[\ [Tv_1]_{\mathcal{C}} \ \Big|\ \ldots \ \Big|\ [Tv_n]_{\mathcal{C}}\ \Big] \in F^{m \times n}.$$

When $W = V$ and $\mathcal{C} = \mathcal{B}$, we use the notation $[T]_{\mathcal{B}}$ instead of $[T]_{\mathcal{B}}^{\mathcal{B}}$.

Remark. Note that in the notation $[T]_{\mathcal{C}}^{\mathcal{B}}$ we put the basis of the domain on top of the basis of the codomain. The usefulness of this choice is manifested in the next theorem.

Theorem 3.34. *Suppose that V, W are finite dimensional vector spaces, and $\mathcal{B}, \mathcal{C}$ are bases for them respectively. Let $T \in \mathcal{L}(V, W)$. Then for any $v \in V$ we have*

$$[Tv]_{\mathcal{C}} = [T]_{\mathcal{C}}^{\mathcal{B}}[v]_{\mathcal{B}}.$$

Remark. In other words, if we multiply the matrix of T and the coordinate vector of v, we get the coordinate vector of Tv.

Proof. Suppose $\mathcal{B} = \{v_1, \dots, v_n\}$ and $\mathcal{C} = \{w_1, \dots, w_m\}$. Then there are $x_j \in F$ such that $v = \sum x_j v_j$. Hence $[v]_{\mathcal{B}} = [x_1, \dots, x_n]^\intercal$. Let $A := [T]_{\mathcal{C}}^{\mathcal{B}}$ and $x := [v]_{\mathcal{B}}$. Then by the properties of matrix multiplication we get

$$\begin{aligned}[T]_{\mathcal{C}}^{\mathcal{B}}[v]_{\mathcal{B}} = Ax &= \sum_{j \le n} x_j A_{\cdot,j} = \sum_{j \le n} x_j [Tv_j]_{\mathcal{C}} = \sum_{j \le n} [x_j T v_j]_{\mathcal{C}} \\ &= \Big[\sum_{j \le n} x_j T v_j\Big]_{\mathcal{C}} = \Big[T\Big(\sum_{j \le n} x_j v_j\Big)\Big]_{\mathcal{C}} = [T(v)]_{\mathcal{C}}.\end{aligned}$$

Note that we have used the linearity of T and the linearity of the coordinate isomorphism with respect to $\mathcal{C}$. ■

Proposition 3.35. *Suppose $\mathcal{B} = \{v_1, \dots, v_n\}$ is a basis for V, and $\mathcal{C} = \{w_1, \dots, w_m\}$ is a basis for W. Then the function*

$$\begin{aligned}\phi_{\mathcal{C}}^{\mathcal{B}} : \mathcal{L}(V, W) &\longrightarrow F^{m \times n} \\ T &\longmapsto [T]_{\mathcal{C}}^{\mathcal{B}}\end{aligned}$$

is a linear isomorphism. As a result, $\mathcal{L}(V, W)$ is finite dimensional, and we have

$$\dim \mathcal{L}(V, W) = \dim V \cdot \dim W.$$

Remark. Note that the linearity of $\phi_{\mathcal{C}}^{\mathcal{B}}$ means that for $T, S \in \mathcal{L}(V, W)$ and $a \in F$ we have

$$[T + S]_{\mathcal{C}}^{\mathcal{B}} = [T]_{\mathcal{C}}^{\mathcal{B}} + [S]_{\mathcal{C}}^{\mathcal{B}}, \qquad [aT]_{\mathcal{C}}^{\mathcal{B}} = a[T]_{\mathcal{C}}^{\mathcal{B}}.$$

Also, the fact that $\phi_{\mathcal{C}}^{\mathcal{B}}$ is one-to-one implies that a linear map is uniquely determined by its matrix; and the fact that $\phi_{\mathcal{C}}^{\mathcal{B}}$ is onto implies that every matrix is the matrix of some linear map.

Proof. Let $S, T \in \mathcal{L}(V, W)$ and $a \in F$. Then for every j we have

$$\begin{aligned}\big([T + aS]_{\mathcal{C}}^{\mathcal{B}}\big)_{\cdot,j} &= [(T + aS)(v_j)]_{\mathcal{C}} = [Tv_j + aSv_j]_{\mathcal{C}} \\ &= [Tv_j]_{\mathcal{C}} + a[Sv_j]_{\mathcal{C}} = \big([T]_{\mathcal{C}}^{\mathcal{B}}\big)_{\cdot,j} + a\big([S]_{\mathcal{C}}^{\mathcal{B}}\big)_{\cdot,j} = \big([T]_{\mathcal{C}}^{\mathcal{B}} + a[S]_{\mathcal{C}}^{\mathcal{B}}\big)_{\cdot,j}.\end{aligned}$$

Note that we have used the linearity of $\phi_{\mathcal{C}}$, i.e. the coordinate isomorphism with respect to $\mathcal{C}$. Hence we have $[T + aS]_{\mathcal{C}}^{\mathcal{B}} = [T]_{\mathcal{C}}^{\mathcal{B}} + a[S]_{\mathcal{C}}^{\mathcal{B}}$. So $\phi_{\mathcal{C}}^{\mathcal{B}}$ is linear.

Now suppose $\phi_{\mathcal{C}}^{\mathcal{B}}(T) = [T]_{\mathcal{C}}^{\mathcal{B}} = 0$. Then by Theorem 3.34, for every $v \in V$ we have

$$[Tv]_{\mathcal{C}} = [T]_{\mathcal{C}}^{\mathcal{B}}[v]_{\mathcal{B}} = 0[v]_{\mathcal{B}} = 0.$$

Therefore $Tv = 0$, since $\phi_{\mathcal{C}}$ is an isomorphism. Hence $T = 0$. Thus $\phi_{\mathcal{C}}^{\mathcal{B}}$ is one-to-one.

So we only need to show that $\phi_{\mathcal{C}}^{\mathcal{B}}$ is onto. Let $A \in F^{m\times n}$. Consider the linear map $S : F^n \to F^m$ which is defined by $S(x) := Ax$ for every $x \in F^n$. Remember that $\phi_{\mathcal{B}} : V \to F^n$ and $\phi_{\mathcal{C}} : W \to F^m$ are linear isomorphisms. Let $T := \phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}} \in \mathcal{L}(V, W)$. Then for every j we have

$$\begin{aligned}\big([T]_{\mathcal{C}}^{\mathcal{B}}\big)_{\cdot,j} &= [Tv_j]_{\mathcal{C}} = \phi_{\mathcal{C}}(Tv_j) = \phi_{\mathcal{C}}\big(\phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}}(v_j)\big) \\ &= S\phi_{\mathcal{B}}(v_j) = S\big(\phi_{\mathcal{B}}(v_j)\big) = S([v_j]_{\mathcal{B}}) = A[v_j]_{\mathcal{B}} = Ae_j = A_{\cdot,j}.\end{aligned}$$

Thus $[T]_{\mathcal{C}}^{\mathcal{B}} = A$, and therefore $\phi_{\mathcal{C}}^{\mathcal{B}}$ is onto. Hence $\phi_{\mathcal{C}}^{\mathcal{B}}$ is an isomorphism as desired. As a result, $\mathcal{L}(V, W)$ is finite dimensional, and we have $\dim \mathcal{L}(V, W) = \dim F^{m\times n} = nm$. ∎

Remark. Note that in the above proof we needed to show directly that $\phi_{\mathcal{C}}^{\mathcal{B}}$ is both one-to-one and onto. Because we did not know a priori that $\mathcal{L}(V, W)$ is finite dimensional and has the same dimension as $F^{m\times n}$. Although it is not hard to explicitly construct a basis for $\mathcal{L}(V, W)$ that resembles the standard basis of $F^{m\times n}$.

***Example* 3.36.** Suppose V is finite dimensional, and $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V. It is easy to see that $[I_V]_{\mathcal{B}} = I$, i.e. the matrix of the identity map is the identity matrix. Because the j-th column of $[I_V]_{\mathcal{B}}$ is $[I_V(v_j)]_{\mathcal{B}} = [v_j]_{\mathcal{B}} = e_j = I_{\cdot,j}$. So we get the desired. Note that this result only holds when we use the same basis in both the domain and the codomain. In fact if $\mathcal{C}$ is another basis for V that is not equal to $\mathcal{B}$, then we must have $[I_V]_{\mathcal{C}}^{\mathcal{B}} \neq I$.

***Example* 3.37.** Suppose $A \in F^{m\times n}$, and $T \in \mathcal{L}(F^n, F^m)$ is defined by $T(x) = Ax$, where $x \in F^n$. Let $\mathcal{B} = \{e_1, \ldots, e_n\}$, $\mathcal{C} = \{e_1, \ldots, e_m\}$ be the standard bases for F^n, F^m, respectively. Then we have

$$[T]_{\mathcal{C}}^{\mathcal{B}} = A.$$

Because the j-th column of $[T]_{\mathcal{C}}^{\mathcal{B}}$ is

$$[Te_j]_{\mathcal{C}} = [Ae_j]_{\mathcal{C}} = [A_{\cdot,j}]_{\mathcal{C}} = A_{\cdot,j}.$$

Note that for $A_{\cdot,j} \in F^m$ we have $[A_{\cdot,j}]_{\mathcal{C}} = A_{\cdot,j}$.

***Example* 3.38.** Let $T \in \mathcal{L}(F^n, F^m)$. Then we claim that there is a unique matrix $A \in F^{m\times n}$ such that $T(x) = Ax$ for every $x \in F^n$. Let $\mathcal{B} = \{e_1, \dots, e_n\}$, $\mathcal{C} = \{e_1, \dots, e_m\}$ be the standard bases for F^n, F^m, respectively. Now set $A := [T]^{\mathcal{B}}_{\mathcal{C}}$. Then for every $x \in F^n$ we have

$$T(x) = [T(x)]_{\mathcal{C}} = [T]^{\mathcal{B}}_{\mathcal{C}}[x]_{\mathcal{B}} = Ax.$$

Note that the coordinate vector of any element y of some F^k in the standard basis is y itself. Finally note that the uniqueness of A follows from the previous example, since it shows that any matrix that satisfies our criterion must be equal to $[T]^{\mathcal{B}}_{\mathcal{C}}$.

Theorem 3.39. *Suppose that V, W, U are finite dimensional vector spaces over a field F, and $\mathcal{B}, \mathcal{C}, \mathcal{D}$ are bases for them respectively. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, U)$. Then we have*

$$[ST]^{\mathcal{B}}_{\mathcal{D}} = [S]^{\mathcal{C}}_{\mathcal{D}}[T]^{\mathcal{B}}_{\mathcal{C}}.$$

Remark. The above theorem means that the composition of linear maps corresponds to the multiplication of matrices. In fact, the multiplication of matrices is defined in a way to make this theorem valid, and this is the main reason behind its definition.

Proof. Suppose $\mathcal{B} = \{v_1, \dots, v_n\}$. Let $A := [T]^{\mathcal{B}}_{\mathcal{C}}$, $B := [S]^{\mathcal{C}}_{\mathcal{D}}$ and $C := [ST]^{\mathcal{B}}_{\mathcal{D}}$. Then by Theorem 3.34 we have

$$\begin{aligned} C_{\cdot,j} = Ce_j &= [ST]^{\mathcal{B}}_{\mathcal{D}}[v_j]_{\mathcal{B}} = [ST(v_j)]_{\mathcal{D}} = [S(Tv_j)]_{\mathcal{D}} \\ &= [S]^{\mathcal{C}}_{\mathcal{D}}[Tv_j]_{\mathcal{C}} = [S]^{\mathcal{C}}_{\mathcal{D}}\big([T]^{\mathcal{B}}_{\mathcal{C}}[v_j]_{\mathcal{B}}\big) = B(Ae_j) = (BA)e_j = (BA)_{\cdot,j}. \end{aligned}$$

So the j-th column of C is equal to the j-th column of BA for every j. Therefore $C = BA$ as desired. ∎

Remark. The above theorem enables us to give another proof for the associativity of matrix multiplication. It actually sheds some new light on this matter, and makes it clear why the multiplication of matrices, which is defined in terms of their entries in a nontrivial way, must be associative. The reason is that the multiplication of matrices corresponds to the composition of linear maps, and the composition of functions is obviously associative. Now to prove this rigorously suppose that $A \in F^{p\times m}$, $B \in F^{m\times n}$, and $C \in F^{n\times l}$. Let $T \in \mathcal{L}(F^m, F^p)$, $S \in \mathcal{L}(F^n, F^m)$, and $R \in \mathcal{L}(F^l, F^n)$ be the linear maps whose matrices with respect to the standard bases are A, B, C respectively. Then we have

$$\begin{aligned} A(BC) &= [T]([S][R]) = [T][SR] = [T(SR)] \\ &= [(TS)R] = [TS][R] = ([T][S])[R] = (AB)C. \end{aligned}$$

Theorem 3.40. *Suppose that V, W are finite dimensional vector spaces, and $\mathcal{B}, \mathcal{C}$ are bases for them respectively. Let $T : V \to W$ be an invertible linear map. Then the matrix $[T]_{\mathcal{C}}^{\mathcal{B}}$ is invertible, and we have*

$$[T^{-1}]_{\mathcal{B}}^{\mathcal{C}} = \left([T]_{\mathcal{C}}^{\mathcal{B}}\right)^{-1}.$$

Proof. First note that $\dim V = \dim W$, since T is an isomorphism. Therefore both $[T]_{\mathcal{C}}^{\mathcal{B}}, [T^{-1}]_{\mathcal{B}}^{\mathcal{C}}$ are square matrices. Now by Theorem 3.39 we have

$$[T]_{\mathcal{C}}^{\mathcal{B}}[T^{-1}]_{\mathcal{B}}^{\mathcal{C}} = [TT^{-1}]_{\mathcal{C}}^{\mathcal{C}} = [I_W]_{\mathcal{C}}^{\mathcal{C}} = I.$$

Similarly we have $[T^{-1}]_{\mathcal{B}}^{\mathcal{C}}[T]_{\mathcal{C}}^{\mathcal{B}} = [I_V]_{\mathcal{B}}^{\mathcal{B}} = I$. Hence $[T]_{\mathcal{C}}^{\mathcal{B}}$ is an invertible matrix, and its inverse is $[T^{-1}]_{\mathcal{B}}^{\mathcal{C}}$. ■

Theorem 3.41. *Suppose that V is finite dimensional, and $\mathcal{B}, \mathcal{C}$ are bases for V. Let $v \in V$, and $T \in \mathcal{L}(V)$. Then we have*

$$[v]_{\mathcal{C}} = P[v]_{\mathcal{B}}, \qquad [T]_{\mathcal{C}} = P[T]_{\mathcal{B}}P^{-1},$$

where $P = [I_V]_{\mathcal{C}}^{\mathcal{B}}$.

Remark. Note that by the previous theorem the matrix P is invertible, and we have $P^{-1} = [I_V^{-1}]_{\mathcal{B}}^{\mathcal{C}} = [I_V]_{\mathcal{B}}^{\mathcal{C}}$. We can also write the above relations as

$$[v]_{\mathcal{B}} = P^{-1}[v]_{\mathcal{C}}, \qquad [T]_{\mathcal{B}} = P^{-1}[T]_{\mathcal{C}}P.$$

The matrices P, P^{-1} are called the **change of coordinates matrices**.

Remark. Two matrices $A, B \in F^{n\times n}$ are called **similar**, if there exists an invertible matrix $C \in F^{n\times n}$ such that $B = CAC^{-1}$, or equivalently $A = C^{-1}BC$. Hence the above theorem implies that the matrices of a linear map T in two different bases, are similar matrices. Because of this theorem, many properties of similar matrices are the same, since they are actually different representations of the same linear map.

Proof. By Theorem 3.34 we have

$$[v]_{\mathcal{C}} = [I_V v]_{\mathcal{C}} = [I_V]_{\mathcal{C}}^{\mathcal{B}}[v]_{\mathcal{B}} = P[v]_{\mathcal{B}}.$$

Now by Theorem 3.39 we have

$$\begin{aligned} P[T]_{\mathcal{B}}P^{-1} &= [I_V]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}}[I_V]_{\mathcal{B}}^{\mathcal{C}} = [I_V]_{\mathcal{C}}^{\mathcal{B}}[TI_V]_{\mathcal{B}}^{\mathcal{C}} \\ &= [I_V]_{\mathcal{C}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{C}} = [I_V T]_{\mathcal{C}}^{\mathcal{C}} = [T]_{\mathcal{C}}^{\mathcal{C}} = [T]_{\mathcal{C}}. \end{aligned}$$

■

3.4 More about Matrices and Linear Systems

Theorem 3.42. *Suppose $A, B \in F^{n\times n}$. If $AB = I$ then $BA = I$, and therefore we have $B = A^{-1}$.*

Proof. Let $S, T \in \mathcal{L}(F^n)$ be defined by $T(x) = Ax$ and $S(x) = Bx$, where $x \in F^n$. Let $\mathcal{B}$ be the standard basis for F^n. Then we know that $[T]_\mathcal{B} = A$ and $[S]_\mathcal{B} = B$. Now we have

$$[TS]_\mathcal{B} = [T]_\mathcal{B}[S]_\mathcal{B} = AB = I = [I_{F^n}]_\mathcal{B}.$$

Thus $TS = I_{F^n}$. But F^n is finite dimensional, so we must have $ST = I_{F^n}$ too. Therefore we get

$$BA = [S]_\mathcal{B}[T]_\mathcal{B} = [ST]_\mathcal{B} = [I_{F^n}]_\mathcal{B} = I,$$

as desired. Hence by definition we have $B = A^{-1}$. ■

Theorem 3.43. *Let $A \in F^{m\times n}$, and consider the linear system whose coefficient matrix is A.*

(i) *If $m < n$, i.e. if the number of equations is less than the number of unknowns, then the homogeneous linear system $Ax = 0$ has at least one nontrivial solution $x \in F^n - \{0\}$.*

(ii) *If $m > n$, i.e. if the number of equations is more than the number of unknowns, then there is $b \in F^m$ such that the nonhomogeneous linear system $Ax = b$ has no solution in F^n.*

Proof. Let $T \in \mathcal{L}(F^n, F^m)$ be the linear map defined by $T(x) := Ax$ for $x \in F^n$.

(i) We know that $T(F^n) \subset F^m$, so $\dim T(F^n) \leq \dim F^m = m$. Thus when $m < n$ we have

$$\dim \operatorname{null} T = \dim F^n - \dim T(F^n) \geq \dim F^n - \dim F^m = n - m > 0.$$

Therefore $\operatorname{null} T$ contains some nonzero vector $x \in F^n$, and we have $Ax = Tx = 0$.

(ii) When $m > n$ we have

$$\dim T(F^n) = \dim F^n - \dim \operatorname{null} T \leq \dim F^n = n < m = \dim F^m.$$

Therefore $T(F^n) \neq F^m$. In other words, T is not onto. Hence there is $b \in F^m$ such that $b \notin T(F^n)$, i.e. there is no $x \in F^n$ such that $Ax = Tx = b$. ■

Definition 3.44. The **rank** of a matrix $A \in F^{m\times n}$ is the dimension of the subspace of F^m that the columns of A generate, i.e.

$$\operatorname{rank} A := \dim \operatorname{span}(A_{\cdot,1}, \ldots, A_{\cdot,n}).$$

Remark. It is obvious that $\operatorname{rank} A \leq \min\{m, n\}$, since $\operatorname{span}(A_{.,1}, \ldots, A_{.,n})$ is generated by n vectors, and is a subspace of F^m.

Theorem 3.45. *Suppose that V, W are finite dimensional vector spaces, and $\mathcal{B}, \mathcal{C}$ are bases for them respectively. Then the rank of $T \in \mathcal{L}(V, W)$ is equal to the rank of its matrix $[T]_{\mathcal{C}}^{\mathcal{B}}$, i.e.*

$$\operatorname{rank} T = \dim T(V) = \operatorname{rank} [T]_{\mathcal{C}}^{\mathcal{B}}.$$

Proof. Suppose $n = \dim V$ and $m = \dim W$. Let $A := [T]_{C}^{\mathcal{B}}$, and let $S \in \mathcal{L}(F^n, F^m)$ be the linear map defined by $S(x) := Ax$ for $x \in F^n$. By Proposition 3.23, we know that $Se_1, \ldots, Se_n$ generate $S(F^n)$, where $e_1, \ldots, e_n$ is the standard basis of F^n. But $Se_j = Ae_j = A_{.,j}$. Thus we have

$$S(F^n) = \operatorname{span}(Se_1, \ldots, Se_n) = \operatorname{span}(A_{.,1}, \ldots, A_{.,n}).$$

Hence $\operatorname{rank} S = \dim S(F^n) = \dim \operatorname{span}(A_{.,1}, \ldots, A_{.,n}) = \operatorname{rank} A$.

Now let $\phi_{\mathcal{B}} : V \to F^n$ and $\phi_{\mathcal{C}} : W \to F^m$ be the coordinate isomorphisms. We know that $[Tv]_{\mathcal{C}} = A[v]_{\mathcal{B}}$ for every $v \in V$. In other words we have $\phi_{\mathcal{C}}(Tv) = S(\phi_{\mathcal{B}}(v))$ for every $v \in V$. Hence $T = \phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}}$. Thus

$$T(V) = \phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}}(V) = \phi_{\mathcal{C}}^{-1} S(F^n),$$

since $\phi_{\mathcal{B}}$ is onto and so we have $F^n = \phi_{\mathcal{B}}(V)$. Now note that $\phi_{\mathcal{C}}^{-1}$ is invertible, so it is one-to-one. Thus by Proposition 3.23 we have $\dim \phi_{\mathcal{C}}^{-1} S(F^n) = \dim S(F^n)$. Therefore we get

$$\operatorname{rank} T = \dim T(V) = \dim \phi_{\mathcal{C}}^{-1} S(F^n) = \dim S(F^n) = \operatorname{rank} S = \operatorname{rank} [T]_{\mathcal{C}}^{\mathcal{B}},$$

as desired. ■

Theorem 3.46. *Let $A, B \in F^{m \times n}$, and suppose B is the reduced row echelon form of A. Then the rank of A is equal to the number of nonzero rows of B.*

Remark. Note that B is its own reduced row echelon form. Hence the rank of A is equal to the rank of B. In other words, the rank of a matrix is the same as the rank of its reduced row echelon form. Also note that the number of nonzero rows of B is the same as the number of leading entries of B.

Proof. The number of nonzero rows of B are the same as the number of leading entries of B. Now the theorem follows trivially from Proposition 2.45. ■

Remark. The above two theorems provide us an algorithm to compute the rank of a matrix or a linear map between finite dimensional vector spaces. First we find the matrix of the linear map with respect to some bases. Then we compute the reduced row echelon form of that matrix. And finally we count the number of nonzero rows of the matrix in reduced row echelon form. This algorithm will also enable us to compute the nullity of the linear map by using the rank-nullity theorem.

The above algorithm is suitable for computing the rank of small matrices by hand, but it is not appropriate for large-scale calculations that require computers. To see this consider for example the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & \epsilon & 0 \\ 0 & 0 & \delta \end{bmatrix},$$

where ϵ, δ are small positive real numbers. Now the rank of this matrix is obviously 3. But if ϵ, δ are so small that are considered zero by a computer, then that computer calculates the rank to be 1.

Theorem 3.47. *Let $A, B \in F^{m\times n}$, and suppose B is the reduced row echelon form of A. Let $B_{\cdot,j_1}, \ldots, B_{\cdot,j_k}$ be the columns of B that contain a leading entry, and let $B_{\cdot,l_1}, \ldots, B_{\cdot,l_{n-k}}$ be the columns of B that do not contain a leading entry. Then the set of vectors*

$$v_{l_1}, \ldots, v_{l_{n-k}},$$

is a basis for the set of solutions of the homogeneous linear system $Ax = 0$, where for $p \leq n-k$ the vector $v_{l_p} \in F^n$ is given by

$$(v_{l_p})_i = \begin{cases} -B_{q,l_p} & i = j_q, q \leq k, \\ 1 & i = l_p, \\ 0 & i = l_{\tilde{p}}, \tilde{p} \neq p, \tilde{p} \leq n-k. \end{cases}$$

Remark. Note that k is the number of leading entries of B, which by the previous theorem is equal to the rank of A. Therefore the dimension of the set of solutions of the homogeneous linear system $Ax = 0$ is

$$\dim\{x \in F^n : Ax = 0\} = n - \operatorname{rank} A.$$

This equality also follows from the rank-nullity theorem, when we apply it to the linear map $T \in \mathcal{L}(F^n, F^m)$ that is defined by $T(x) := Ax$ for $x \in F^n$.

Proof. Let $W := \{x \in F^n : Ax = 0\}$. Then Theorem 1.35 implies that $v_{l_p} \in W$ for every p. Because the system is homogeneous, so the constant term in the solution is zero. Hence if we choose the free variables to be $x_{l_p} = 1$, and $x_{l_q} = 0$ for $q \neq p$, then the general solution given in Theorem 1.35 becomes v_{l_p}.

Now let $T \in \mathcal{L}(F^n, F^m)$ be the linear map defined by $T(x) := Ax$ for $x \in F^n$. Then A is the matrix of T with respect to the standard bases of F^n, F^m. Hence we have $\operatorname{rank} T = \operatorname{rank} A$. On the other hand, it is obvious that $W = \operatorname{null} T$. Therefore we have

$$\dim W = \dim \operatorname{null} T = n - \operatorname{rank} T = n - \operatorname{rank} A = n - k,$$

since by the previous theorem we have $\operatorname{rank} A = k$. Thus in order to show that $v_{l_1}, \ldots, v_{l_{n-k}}$ form a basis for W, it suffices to show that they span W. But this is exactly what we have proved in Theorem 1.35. ■

Remark. In the above proof, it is also easy to show directly that $v_{l_1}, \ldots, v_{l_{n-k}}$ are linearly independent. Because the l_p-th component of v_{l_p} is 1, while for $q \neq p$ the l_p-th component of v_{l_q} is 0. If we use this direct approach, we can avoid using the rank-nullity theorem in the above proof.

Remark. Consider the nonhomogeneous linear system $Ax = b$, where $b \in F^m$. Suppose that this system is consistent. Let $x_0 \in F^n$ be a solution of the system. Then for any other solution x we have $A(x - x_0) = Ax - Ax_0 = b - b = 0$. Hence $x - x_0$ is a solution of the homogeneous system. Conversely if y is a solution of the homogeneous system, then $x_0 + y$ is a solution of the nonhomogeneous system, since $A(x_0 + y) = Ax_0 + Ay = b + 0 = b$. Therefore we can write

$$\{x \in F^n : Ax = b\} = \{x_0 + x : x \in F^n, Ax = 0\} =: x_0 + \{x \in F^n : Ax = 0\}.$$

And we say that the set of solutions of the nonhomogeneous system is a translated copy of the set of solutions of the corresponding homogeneous system.

Remark. When we want to concretely describe a subspace, we have to somehow describe a spanning set for that subspace. Then we can extract a basis from the given spanning set, and obtain a better understanding of the subspace. Sometimes the spanning set is given to us directly. Then we can find a basis easily, as explained in the remark after Proposition 3.32. Sometimes the subspace is described for us as the image of a given linear map, or more indirectly as the null space of a given linear map. The next theorem provides us a method to compute a basis for the subspace in these cases. We should mention that the above cases are the main ways of concretely describing a subspace.

Theorem 3.48. *Suppose that V, W are finite dimensional vector spaces, and $\mathcal{B}, \mathcal{C}$ are bases for them respectively. Let $\phi_{\mathcal{B}} : V \to F^n$ and $\phi_{\mathcal{C}} : W \to F^m$ be the coordinate isomorphisms. Let $T \in \mathcal{L}(V, W)$, and let $A = [T]_{\mathcal{C}}^{\mathcal{B}} \in F^{m \times n}$. Suppose $B \in F^{m \times n}$ is the reduced row echelon form of A. Let $B_{\cdot, j_1}, \ldots, B_{\cdot, j_k}$ be the columns of B that contain a leading entry, and let $B_{\cdot, l_1}, \ldots, B_{\cdot, l_{n-k}}$ be the columns of B that do not contain a leading entry. Then*

(i) *$\phi_{\mathcal{C}}^{-1}(A_{\cdot, j_1}), \ldots, \phi_{\mathcal{C}}^{-1}(A_{\cdot, j_k})$ is a basis for the image of T.*

(ii) $\phi_{\mathcal{B}}^{-1}(v_{l_1}), \ldots, \phi_{\mathcal{B}}^{-1}(v_{l_{n-k}})$ *is a basis for the null space of* T*, where for* $p \leq n-k$ *the vector* $v_{l_p} \in F^n$ *is given by*

$$(v_{l_p})_i = \begin{cases} -B_{q,l_p} & i = j_q, q \leq k, \\ 1 & i = l_p, \\ 0 & i = l_{\tilde{p}}, \tilde{p} \neq p, \tilde{p} \leq n-k. \end{cases}$$

Remark. Note that by Theorem 2.47, $A_{\cdot,j_1}, \ldots, A_{\cdot,j_k}$ is a basis for the subspace that the columns of A generate, so in particular k is the rank of A which is the same as the rank of T. Also note that by the previous theorem $v_{l_1}, \ldots, v_{l_{n-k}}$ is a basis for the set of solutions of the homogeneous linear system $Ax = 0$. Finally note that we can compute $\phi_{\mathcal{B}}^{-1}, \phi_{\mathcal{C}}^{-1}$ by the formula given in Proposition 3.32.

Proof. Let $S \in \mathcal{L}(F^n, F^m)$ be the linear map defined by $S(x) := Ax$ for $x \in F^n$. We know that $[Tv]_{\mathcal{C}} = A[v]_{\mathcal{B}}$ for every $v \in V$. In other words we have $\phi_{\mathcal{C}}(Tv) = S(\phi_{\mathcal{B}}(v))$ for every $v \in V$. Hence $T = \phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}}$.

(i) By Proposition 3.23, we know that $Se_1, \ldots, Se_n$ generate $S(F^n)$, where $e_1, \ldots, e_n$ is the standard basis of F^n. But $Se_j = Ae_j = A_{\cdot,j}$. Thus we have

$$S(F^n) = \operatorname{span}(Se_1, \ldots, Se_n) = \operatorname{span}(A_{\cdot,1}, \ldots, A_{\cdot,n}).$$

Now we have

$$T(V) = \phi_{\mathcal{C}}^{-1} S \phi_{\mathcal{B}}(V) = \phi_{\mathcal{C}}^{-1} S(F^n),$$

since $\phi_{\mathcal{B}}$ is onto and so we have $F^n = \phi_{\mathcal{B}}(V)$. Now note that $\phi_{\mathcal{C}}^{-1}$ is invertible, so it is one-to-one. Thus by Proposition 3.23, since $A_{\cdot,j_1}, \ldots, A_{\cdot,j_k}$ is a basis for $S(F^n)$, then $\phi_{\mathcal{C}}^{-1}(A_{\cdot,j_1}), \ldots, \phi_{\mathcal{C}}^{-1}(A_{\cdot,j_k})$ is a basis for $\phi_{\mathcal{C}}^{-1} S(F^n) = T(V)$.

(ii) Let $x \in \operatorname{null} S = \{x \in F^n : Ax = 0\}$. Then we have $T\phi_{\mathcal{B}}^{-1} x = \phi_{\mathcal{C}}^{-1} Sx = \phi_{\mathcal{C}}^{-1} 0 = 0$. Thus $\phi_{\mathcal{B}}^{-1} x \in \operatorname{null} T$. So $\phi_{\mathcal{B}}^{-1}\big(\operatorname{null} S\big) \subset \operatorname{null} T$. On the other hand, if $v \in \operatorname{null} T$ then we have $S\phi_{\mathcal{B}} v = \phi_{\mathcal{C}} T v = \phi_{\mathcal{C}} 0 = 0$. Hence $\phi_{\mathcal{B}} v \in \operatorname{null} S$, and $v = \phi_{\mathcal{B}}^{-1} \phi_{\mathcal{B}} v$. Therefore

$$\phi_{\mathcal{B}}^{-1}\big(\{x \in F^n : Ax = 0\}\big) = \phi_{\mathcal{B}}^{-1}\big(\operatorname{null} S\big) = \operatorname{null} T.$$

Now note that $\phi_{\mathcal{B}}^{-1}$ is invertible, so it is one-to-one. Thus by Proposition 3.23, since $v_{l_1}, \ldots, v_{l_{n-k}}$ is a basis for the set of solutions of the homogeneous linear system $Ax = 0$, then $\phi_{\mathcal{B}}^{-1}(v_{l_1}), \ldots, \phi_{\mathcal{B}}^{-1}(v_{l_{n-k}})$ is a basis for $\operatorname{null} T$. ■

Theorem 3.49. *Let* $A \in F^{n \times n}$*. Then the following statements are equivalent.*

(i) A *is invertible.*
(ii) $\operatorname{rank} A = n$.
(iii) *The reduced row echelon form of* A *is the identity matrix* $I \in F^{n \times n}$.
(iv) A *is the product of finitely many elementary matrices in* $F^{n \times n}$.

(v) *For every $b \in F^n$ the linear system $Ax = b$ has exactly one solution $x \in F^n$.*
(vi) *There exists $b \in F^n$ such that the linear system $Ax = b$ has exactly one solution $x \in F^n$.*

Remark. A particular case of condition (vi), which often occurs, is "the homogeneous system $Ax = 0$ does not have a nontrivial solution". Conversely, if the homogeneous system $Ax = 0$ has a nontrivial solution, then condition (v) implies that A is not invertible.

Remark. It is trivial to see that if A is invertible, then the unique solution of the linear system $Ax = b$ is $x = A^{-1}b$.

Remark. Note that the equality $\operatorname{rank} A = n$ is equivalent to the fact that the columns of A generate a subspace of F^n with dimension n, i.e. the columns of A generate F^n. But A has n columns, so if they generate F^n they must form a basis for F^n. Hence the equality $\operatorname{rank} A = n$ is equivalent to the fact that the columns of A form a basis for F^n. Similarly, by using Theorem 3.50 we can show that the equality $\operatorname{rank} A = n$ is equivalent to the fact that the rows of A form a basis for F^n.

Proof. **(i) $\Longrightarrow$ (ii):** If $Ax = 0$ then

$$x = Ix = A^{-1}Ax = A^{-1}0 = 0.$$

Thus $\{x \in F^n : Ax = 0\} = \{0\}$. Hence we have

$$n - \operatorname{rank} A = \dim\{x \in F^n : Ax = 0\} = \dim\{0\} = 0.$$

(ii) $\Longrightarrow$ (iii): Let $B \in F^{n\times n}$ be the reduced row echelon form of A. We know that the number of leading entries of B is equal to $\operatorname{rank} A = n$. Therefore every column of B must contain a leading entry, since no column can contain more than one leading entry. So Proposition 1.34 implies that every column of B is an element of the standard basis of F^n. This proposition also tells us that in a matrix which is in reduced row echelon form, the i-th column from the left that contains a leading entry is equal to e_i. Hence we must have $B_{\cdot,j} = e_j$ for every j, since $B_{\cdot,j}$ is the j-th column from the left that contains a leading entry. Thus we obtain $B = I$.

(iii) $\Longrightarrow$ (iv): We know that there is a sequence of elementary matrices $E_1, \ldots, E_k$ such that $E_1 \cdots E_k A = I$, since I is the reduced row echelon form of A. For example we can construct $E_1, \ldots, E_k$ by applying the Gaussian elimination to A. Now every elementary matrix is invertible, and its inverse is an elementary matrix too. Hence we have $A = E_k^{-1} \cdots E_1^{-1} I = E_k^{-1} \cdots E_1^{-1}$ as desired.

(iv) $\Longrightarrow$ (v): Suppose $A = E_1 \cdots E_k$, where $E_1, \ldots, E_k$ are elementary matrices. Then $Ax = b$ implies $E_1 \cdots E_k x = b$. Therefore $x = E_k^{-1} \cdots E_1^{-1} b$, since elementary matrices are invertible. On the other hand we have

$$Ax = E_1 \cdots E_k E_k^{-1} \cdots E_1^{-1} b = E_1 \cdots E_{k-1} I E_{k-1}^{-1} \cdots E_1^{-1} b = \cdots = Ib = b.$$

Thus the system $Ax = b$ has exactly one solution.

(v) $\Longrightarrow$ (vi): This is trivial.

(vi) $\Longrightarrow$ (i): Let $T \in \mathcal{L}(F^n)$ be the linear map defined by $T(x) := Ax$ for $x \in F^n$. Then A is the matrix of T with respect to the standard basis of F^n. Suppose x_0 is the unique solution of $Ax = b$. We know that

$$\{x_0\} = \{x \in F^n : Ax = b\} = x_0 + \{x \in F^n : Ax = 0\}.$$

Hence $\{x \in F^n : Ax = 0\}$ must have exactly one element, because $x_0 + x$ uniquely determines x. Thus $\{x \in F^n : Ax = 0\} = \{0\}$. Therefore

$$\dim \operatorname{null} T = \dim\{x \in F^n : Ax = 0\} = \dim\{0\} = 0.$$

Hence T is one-to-one, and therefore it is invertible. Thus the matrix of T, i.e. A, is also invertible by Theorem 3.40. ■

Remark. Let $A \in F^{n\times n}$. We know that the Gaussian elimination produces a finite sequence of elementary matrices $E_1, \ldots, E_k \in F^{n\times n}$ such that $E_k \cdots E_1 A$ is the reduced row echelon form of A. Hence when A is invertible we have

$$E_k \cdots E_1 A = I.$$

But this implies that $A^{-1} = E_k \cdots E_1 = E_k \cdots E_1 I$. The meaning of this equality is that if we apply to I the same sequence of elementary row operations which convert A to I, then we will obtain A^{-1}.

This gives us an algorithm to check whether A is invertible, and to compute its inverse if it is invertible. We apply the Gaussian elimination to A, and in each step we apply the same operation to I too. We continue until we find the matrices B, C, where B is the reduced row echelon form of A, and C is the matrix produced from I. Now if $B = I$ then A is invertible, and $A^{-1} = C$. And if $B \neq I$ then A is not invertible. ■

Theorem 3.50. *The rank of a matrix $A \in F^{m\times n}$ is equal to the dimension of the subspace of F^n that the rows of A generate, i.e.*

$$\operatorname{rank} A = \dim \operatorname{span}(A_{1,.}, \ldots, A_{m,.}).$$

As a result we have

$$\operatorname{rank} A = \operatorname{rank} A^{\intercal}.$$

Remark. This is one of the most fascinating theorems in linear algebra. It states that if we have a rectangular array of scalars, i.e. a matrix, then the maximum number of linearly independent columns is the same as the maximum number of linearly independent rows. This fact is really nontrivial, considering that our assumptions about the array of scalars are minimal.

Remark. Some authors use the name *row rank* for the dimension of the subspace of F^n that the rows of A generate, and they use the name *column rank* for the dimension of the subspace of F^m that the columns of A generate. In this terminology the above theorem says that the row rank of any matrix is equal to its column rank.

Proof. Consider the system $Ax = 0$. If we write this system more explicitly in terms of its equations, i.e. the rows of A, we get

$$\begin{cases} A_{1,.}x = 0, \\ \quad \vdots \\ A_{m,.}x = 0. \end{cases}$$

Now suppose $A_{i_1,.}, \ldots, A_{i_k,.}$ are a basis for $\operatorname{span}(A_{1,.}, \ldots, A_{m,.})$. Let $B \in F^{k\times n}$ be the matrix whose l-th row is $A_{i_l,.}$ for $l \leq k$. Then the system $Bx = 0$ is

$$\begin{cases} A_{i_1,.}x = 0, \\ \quad \vdots \\ A_{i_k,.}x = 0. \end{cases}$$

If $x \in F^n$ satisfies $Ax = 0$, then it obviously satisfies $Bx = 0$ too. Because all the equations of the system $Bx = 0$ are among the equations of $Ax = 0$. Conversely, consider the equation $A_{i,.}x = 0$ for some $i \leq m$. Then we know that there are $a_1, \ldots, a_k \in F$ such that

$$A_{i,.} = a_1 A_{i_1,.} + \cdots + a_k A_{i_k,.}.$$

Hence if $x \in F^n$ satisfies $Bx = 0$ then we have

$$\begin{aligned} A_{i,.}x &= \big(a_1 A_{i_1,.} + \cdots + a_k A_{i_k,.}\big)x \\ &= a_1 A_{i_1,.}x + \cdots + a_k A_{i_k,.}x = a_1 0 + \cdots + a_k 0 = 0. \end{aligned}$$

Thus x also satisfies $Ax = 0$. Therefore the two homogeneous systems have the same set of solutions.

Now note that $B \in F^{k\times n}$, so $\operatorname{rank} B \leq k$. Hence we have

$$\begin{aligned} n - \operatorname{rank} A &= \dim\{x \in F^n : Ax = 0\} \\ &= \dim\{x \in F^n : Bx = 0\} = n - \operatorname{rank} B \geq n - k. \end{aligned}$$

Thus $\operatorname{rank} A \leq k$. But note that the rows of A are the columns of $A^\intercal$. Therefore

$$k = \dim \operatorname{span}(A_{1,.}, \ldots, A_{m,.}) = \dim \operatorname{span}(A^\intercal_{.,1}, \ldots, A^\intercal_{.,m}) = \operatorname{rank} A^\intercal.$$

So we have proved that $\operatorname{rank} A \leq \operatorname{rank} A^{\intercal}$. If we repeat the above argument with $A^{\intercal}$ instead of A, we get

$$\operatorname{rank} A^{\intercal} \leq \operatorname{rank} (A^{\intercal})^{\intercal} = \operatorname{rank} A.$$

Hence we obtain

$$\operatorname{rank} A = \operatorname{rank} A^{\intercal} = k = \dim \operatorname{span}(A_{1,.}, \ldots, A_{m,.}),$$

as desired. ■

Chapter 4

Diagonalization

4.1 Eigenvalues and Eigenvectors

Notation. In this chapter we assume that F is a field, V is a nonzero vector space over F, and $T \in \mathcal{L}(V)$ is a linear operator.

One of the main goals in linear algebra is to understand the behavior of a linear map $T \in \mathcal{L}(V)$. In order to do this, an effective strategy is to decompose the vector space V into smaller pieces, and then study the linear map T on those smaller pieces. Now, each of those pieces must be a vector space itself, since we want to study a linear map on it. Hence those pieces must be subspaces of V. In addition we need to be able to restrict both the domain and the codomain of T to those smaller subspaces, because we want to simplify the problem by replacing V with a smaller vector space. Therefore those subspaces must be T-invariant, as we will define below.

Definition 4.1. Suppose W is a subspace of V. We say W is ***T*-invariant** if

$$T(W) \subset W.$$

In other words, if $u \in W$ then $Tu \in W$. In this case, the **restriction** of T to W is the function $T|_W : W \to W$ whose value at $u \in W$ is

$$T|_W(u) := Tu.$$

Remark. It is easy to see that $T|_W$ is a linear map, i.e. $T|_W \in \mathcal{L}(W)$.

***Exercise* 4.2.** Suppose $W = \operatorname{span}(v_1, \ldots, v_k)$. Show that if for every $j \leq k$ we have $Tv_j \in W$, then W is T-invariant.

Solution. By Propositions 3.23 and 2.22, we have

$$T(W) = \operatorname{span}(Tv_1, \ldots, Tv_k) \subset \operatorname{span}(v_1, \ldots, v_k) = W. \qquad \blacksquare$$

Remark. Suppose W is a one dimensional T-invariant subspace. Let v be a basis for W. Then v is nonzero. We also have $Tv \in W = \text{span}(v)$. Therefore $Tv = \lambda v$ for some $\lambda \in F$. This observation motivates the next definition.

Definition 4.3. A scalar $\lambda \in F$ is called an **eigenvalue** of the linear operator T if there exists a vector $v \in V$ such that $v \neq 0$, and

$$Tv = \lambda v.$$

The nonzero vector v is called an **eigenvector** of T corresponding to λ.

Similarly, a scalar $\lambda \in F$ is called an **eigenvalue** of the square matrix $A \in F^{n\times n}$ if there exists a vector $x \in F^n$ such that $x \neq 0$, and

$$Ax = \lambda x.$$

The nonzero vector x is called an **eigenvector** of A corresponding to λ.

Theorem 4.4. *Suppose V is finite dimensional, and $\mathcal{B}$ is a basis for V. Then $\lambda \in F$ is an eigenvalue of T if and only if it is an eigenvalue of $[T]_{\mathcal{B}}$.*

Proof. Suppose λ is an eigenvalue of T. Then there is a nonzero vector $v \in V$ such that $Tv = \lambda v$. Hence we have $[T]_{\mathcal{B}}[v]_{\mathcal{B}} = [Tv]_{\mathcal{B}} = [\lambda v]_{\mathcal{B}} = \lambda[v]_{\mathcal{B}}$. Also note that $[v]_{\mathcal{B}}$ is nonzero, since the coordinate isomorphism $\phi_{\mathcal{B}}$ is one-to-one. Thus λ is an eigenvalue of $[T]_{\mathcal{B}}$.

Conversely suppose that λ is an eigenvalue of $[T]_{\mathcal{B}}$. Let $n = \dim V$. Then there is a nonzero vector $x \in F^n$ such that $[T]_{\mathcal{B}}x = \lambda x$. Now let $v := \phi_{\mathcal{B}}^{-1}(x)$. Then $[v]_{\mathcal{B}} = \phi_{\mathcal{B}}\big(\phi_{\mathcal{B}}^{-1}(x)\big) = x$. Therefore

$$[Tv]_{\mathcal{B}} = [T]_{\mathcal{B}}[v]_{\mathcal{B}} = [T]_{\mathcal{B}}x = \lambda x = \lambda[v]_{\mathcal{B}} = [\lambda v]_{\mathcal{B}}.$$

Thus we must have $Tv = \lambda v$, since the coordinate isomorphism $\phi_{\mathcal{B}}$ is one-to-one. ■

Proposition 4.5. *Suppose $\lambda \in F$ and $A \in F^{n\times n}$. Then we have*

(i) *λ is an eigenvalue of T if and only if $T - \lambda I$ is not one-to-one.*

(ii) *λ is an eigenvalue of A if and only if $A - \lambda I$ is not invertible.*

Remark. Note that an $n \times n$ matrix is invertible if and only if its associated linear operator on F^n is one-to-one. But this is not true for an arbitrary linear operator on a vector space, when the vector space is infinite dimensional.

Proof. **(i)** By definition, λ is an eigenvalue of T if and only if there is a nonzero $v \in V$ such that $Tv = \lambda v$, or equivalently $(T - \lambda I)v = 0$. But this is equivalent to the fact that $\text{null}(T - \lambda I) \neq \{0\}$, which is itself equivalent to the fact that $T - \lambda I$ is not one-to-one, due to Proposition 3.20.

(ii) By definition, λ is an eigenvalue of A if and only if there is a nonzero $y \in F^n$ such that $Ay = \lambda y$, or equivalently $(A - \lambda I)y = 0$. But this is equivalent to the fact that the linear system $(A - \lambda I)x = 0$ has a nontrivial solution, which is itself equivalent to the fact that $A - \lambda I$ is not invertible, due to Theorem 3.49. ■

***Exercise* 4.6.** Suppose W is a T-invariant subspace of V, and $w \in W$ is an eigenvector of $T|_W$ corresponding to the eigenvalue λ. Show that w is also an eigenvector of T corresponding to the eigenvalue λ.

Remark. In other words, the eigenvalues and eigenvectors of a restriction of an operator are also eigenvalues and eigenvectors of the operator itself.

Solution. We have $Tw = T|_W w = \lambda w$. Note that $w \in W \subset V$, and $w \neq 0$. ■

***Exercise* 4.7.** Show that a square matrix $A \in F^{n\times n}$ is non-invertible if and only if 0 is an eigenvalue of A.

Solution. If 0 is an eigenvalue of A then there is a nonzero $y \in F^n$ such that $Ay = 0y = 0$. Hence the homogeneous system $Ax = 0$ has a nontrivial solution. Thus A is non-invertible by Theorem 3.49.

Conversely suppose that A is non-invertible. Then Theorem 3.49 implies that the homogeneous system $Ax = 0$ has a nontrivial solution $y \in F^n - \{0\}$. Hence we have $Ay = 0 = 0y$. Thus 0 is an eigenvalue of A. ■

Theorem 4.8. *Suppose $A \in F^{n\times n}$ is a diagonal matrix, and $d_1, \ldots, d_n$ are the diagonal entries of A. Then $\lambda \in F$ is an eigenvalue of A if and only if $\lambda = d_j$ for some $j \leq n$.*

Proof. We assumed that

$$A = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix}.$$

Now for every $j \leq n$ we have $Ae_j = A_{\cdot,j} = [0, \ldots, 0, d_j, 0, \ldots, 0]^\intercal = d_j e_j$. Hence d_j is an eigenvalue of A, since e_j is nonzero. Conversely suppose that λ is an eigenvalue of A. Then there is a nonzero $x \in F^n$ such that $Ax = \lambda x$. Suppose $x = [x_1, \ldots, x_n]^\intercal$. Then we have

$$\sum_{j\leq n} \lambda x_j e_j = \lambda x = Ax = A\Big(\sum_{j\leq n} x_j e_j\Big) = \sum_{j\leq n} x_j A e_j = \sum_{j\leq n} x_j d_j e_j.$$

Thus we have $\sum_{j\leq n}(\lambda - d_j)x_j e_j = 0$. But $e_1, \ldots, e_n$ are linearly independent. Therefore we must have $(\lambda - d_j)x_j = 0$ for every $j \leq n$. On the other hand $x \neq 0$, so $x_i \neq 0$ for some $i \leq n$. Hence we must have $\lambda = d_i$. ■

Theorem 4.9. *Suppose $A \in F^{n\times n}$ is a triangular matrix, and $d_1, \dots, d_n$ are the diagonal entries of A. Then A is invertible if and only if $d_j \neq 0$ for every $j \leq n$.*

Proof. In the following proof we assume that A is upper triangular, but the case of lower triangular matrices can be deduced similarly. So we have

$$A = \begin{bmatrix} d_1 & & * \\ & \ddots & \\ 0 & & d_n \end{bmatrix},$$

where $*$ denotes the entries of A that may or may not be nonzero. Suppose every d_j is nonzero. Then we can multiply the j-th row of A by $\frac{1}{d_j}$ to convert A into the matrix

$$A = \begin{bmatrix} 1 & & * \\ & \ddots & \\ 0 & & 1 \end{bmatrix}.$$

Next we can add suitable multiples of each row of A to the rows above it so that the entries of A above the diagonal become zero. Hence we can convert A by elementary row operations into the identity matrix I. Thus the reduced row echelon form of A is I, and therefore A is invertible by Theorem 3.49.

Conversely suppose that at least one of $d_1, \dots, d_n$ is zero. We have to show that A is not invertible. Let $T \in \mathcal{L}(F^n)$ be the operator that maps $x \in F^n$ to $Tx := Ax$. Then we have $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of F^n. Note that for every $i \leq n$ we have

$$Te_i = Ae_i = A_{\cdot,i} = [*, \dots, *, d_i, 0, \dots, 0]^{\intercal} = *e_1 + \cdots + *e_{i-1} + d_i e_i. \qquad (\star)$$

Therefore we have $Ae_i \in \operatorname{span}(e_1, \dots, e_i)$ for every $i \leq n$. Let j be the smallest index for which $d_j = 0$. Then equation $(\star)$ implies that

$$\begin{aligned} Te_j = Ae_j &= *e_1 + \cdots + *e_{j-1} + 0e_j \\ &= *e_1 + \cdots + *e_{j-1} \in \operatorname{span}(e_1, \dots, e_{j-1}) \subset \operatorname{span}(e_1, \dots, e_j). \end{aligned}$$

On the other hand we know that

$$Te_i = Ae_i \in \operatorname{span}(e_1, \dots, e_i) \subset \operatorname{span}(e_1, \dots, e_{j-1}) \subset \operatorname{span}(e_1, \dots, e_j),$$

for every $i \leq j-1$. Hence $W := \operatorname{span}(e_1, \dots, e_j)$ is T-invariant, as shown in Exercise 4.2. In addition, Propositions 3.23 and 2.22 imply that

$$T(W) = \operatorname{span}(Te_1, \dots, Te_j) \subset \operatorname{span}(e_1, \dots, e_{j-1}) \subsetneq W.$$

Therefore $T|_W$ is not onto. Thus it is not one-to-one either. So by Proposition 3.20 there is $w \in W \subset F^n$ such that $Tw = T|_W w = 0$; and consequently T is also not one-to-one. Hence T is not invertible, and by Theorem 3.40, A is not invertible either. ∎

Remark. Notice the new method we used in the above proof to show that a linear map T is not invertible. Instead of directly showing that T is not one-to-one, or it is not onto, we have shown that there is a T-invariant subspace W such that $T|_W$ is not onto. More generally, we can conclude that T is not invertible if there exists a subspace W such that $\dim T(W) < \dim W$. This fact is an easy consequence of the rank-nullity theorem.

Theorem 4.10. *Suppose $A \in F^{n\times n}$ is a triangular matrix, and $d_1, \ldots, d_n$ are the diagonal entries of A. Then $\lambda \in F$ is an eigenvalue of A if and only if $\lambda = d_j$ for some $j \leq n$.*

Proof. Note that similarly to the previous theorem, A can be upper triangular or lower triangular. Let us assume that A is lower triangular. Then we have

$$A = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ * & & d_n \end{bmatrix} \implies A - \lambda I = \begin{bmatrix} d_1 - \lambda & & 0 \\ & \ddots & \\ * & & d_n - \lambda \end{bmatrix},$$

where $*$ denotes the entries that may or may not be nonzero. Now, Proposition 4.5 implies that λ is an eigenvalue of A if and only if $A - \lambda I$ is non-invertible. On the other hand, by the previous theorem, $A - \lambda I$ is non-invertible if and only if one of its diagonal entries, i.e. $d_j - \lambda$ for some j, is zero. Therefore λ is an eigenvalue of A if and only if $\lambda = d_j$ for some j. ■

***Example* 4.11.** Consider the matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in F^{2\times 2}.$$

We want to find the eigenvalues of A. Suppose λ is an eigenvalue of A. Then

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

is not invertible. Hence its rank is less than 2. Thus its reduced row echelon form must have at least one zero row. If $c = 0$ then A is upper triangular, and as we have seen its eigenvalues are a, d. So suppose that $c \neq 0$. Now we can apply elementary row operations to $A - \lambda I$, to obtain its reduced row echelon form

$$\begin{bmatrix} c & d - \lambda \\ 0 & b - \frac{1}{c}(d - \lambda)(a - \lambda) \end{bmatrix}.$$

But the second row of this matrix must be zero, so we have $b - \frac{1}{c}(d-\lambda)(a-\lambda) = 0$. Therefore λ is a root of the quadratic equation

$$x^2 - (a + d)x + ad - bc = 0.$$

Note that when $c = 0$ the roots of this equation are a, d. It is easy to reverse the above chain of reasoning, and show that any root of the above equation is also an eigenvalue of A.

The above equation is called the *characteristic equation* of A. We can apply the above method to find the eigenvalues of a matrix in $F^{n\times n}$, but the calculations are cumbersome and lengthy. We will devise an easier method to compute the characteristic equation of a matrix, when we study the determinant in Chapter 7.

Theorem 4.12. *The eigenvectors of a linear operator corresponding to distinct eigenvalues are linearly independent.*

Similarly, the eigenvectors of a square matrix corresponding to distinct eigenvalues are linearly independent.

Proof. Suppose $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of T. The proof is by induction on k. Let $v_1, \ldots, v_k$ be eigenvectors of T corresponding to $\lambda_1, \ldots, \lambda_k$ respectively. If $k = 1$ then v_1 is a linearly independent list, since $v_1 \neq 0$. Now suppose the claim is true for $k-1$, i.e. $v_1, \ldots, v_{k-1}$ are linearly independent. We want to show that $v_1, \ldots, v_k$ are also linearly independent. Suppose that for some $a_1, \ldots, a_k \in F$ we have

$$a_1 v_1 + \cdots + a_k v_k = 0.$$

Let $S := T - \lambda_k I$. Then we get

$$0 = S(0) = S(a_1 v_1 + \cdots + a_k v_k) = a_1 S v_1 + \cdots + a_k S v_k.$$

But $S v_k = T v_k - \lambda_k v_k = 0$. And for $j < k$ we have

$$S v_j = T v_j - \lambda_k v_j = \lambda_j v_j - \lambda_k v_j = (\lambda_j - \lambda_k) v_j.$$

Therefore we have

$$0 = a_1 S v_1 + \cdots + a_k S v_k = a_1(\lambda_1 - \lambda_k) v_1 + \cdots + a_k(\lambda_{k-1} - \lambda_k) v_{k-1}.$$

However, $v_1, \ldots, v_{k-1}$ are linearly independent, so we must have $a_j(\lambda_j - \lambda_k) = 0$ for every $j < k$. Hence we get $a_j = 0$, because we know that $\lambda_j \neq \lambda_k$ for $j < k$. Thus we obtain

$$0 = a_1 v_1 + \cdots + a_k v_k = 0 v_1 + \cdots + 0 v_{k-1} + a_k v_k = a_k v_k.$$

So we $a_k = 0$ too, since $v_k \neq 0$. Therefore $v_1, \ldots, v_k$ are linearly independent, as desired. The case of matrices can be proved similarly. ∎

Theorem 4.13. *Suppose V is finite dimensional, and $n = \dim V$. Then every $T \in \mathcal{L}(V)$ has at most n distinct eigenvalues.*

Similarly, every matrix $A \in F^{n\times n}$ has at most n distinct eigenvalues.

Proof. Suppose $\lambda_1, \ldots, \lambda_k$ are all the distinct eigenvalues of T, and $v_1, \ldots, v_k \in V$ are eigenvectors of T corresponding to $\lambda_1, \ldots, \lambda_k$ respectively. Then we know that $v_1, \ldots, v_k$ are linearly independent. Hence we must have $k \leq \dim V = n$, as desired. The case of matrices can be proved similarly. ■

Definition 4.14. For $m \in \mathbb{N}$ we inductively define the **powers** of the linear operator $T \in \mathcal{L}(V)$ to be

$$T^0 := I_V,\ T^1 := T,\ \ldots\ T^m := T^{m-1}T.$$

Also, for every polynomial

$$p(x) = a_0 + a_1 x + \cdots + a_m x^m$$

with coefficients $a_j \in F$, we define

$$p(T) := a_0 I_V + a_1 T + \cdots + a_m T^m.$$

We say that the operator $p(T)$ is a *polynomial in* T.

Theorem 4.15. *Suppose $T, S \in \mathcal{L}(V)$. Then for all nonnegative integers m, k we have*

(i) *If T commutes with S, then T^m commutes with S^k.*

(ii) *If T is invertible, then T^m is also invertible and*

$$(T^m)^{-1} = (T^{-1})^m.$$

(iii) $T^m T^k = T^{m+k}$.

(iv) $(T^m)^k = T^{mk}$.

(v) *If T, S commute, then we have $(TS)^m = T^m S^m$.*

(vi) *For any two polynomials $p, q \in F[x]$ we have*

$$(p+q)(T) = p(T) + q(T), \qquad (pq)(T) = p(T)q(T).$$

As a result, $p(T)$ and $q(T)$ always commute.

Remark. The significance of part (vi) is that the addition and multiplication of polynomials convert to the addition and multiplication of linear operators via the map $p \mapsto p(T)$.

Proof. The proofs can be found in Sections A.1, and A.5. The proofs of parts (i) to (v) are by straightforward inductions. We only repeat the proof of part (vi) here. Suppose $p(x) = a_0 + \cdots + a_m x^m$ and $q(x) = b_0 + \cdots + b_n x^n$, where $a_i, b_j \in F$.

Suppose $n \leq m$. Let $b_j := 0$ for $n < j \leq m$. Then we have $q(x) = b_0 + \cdots + b_m x^m$. We also have

$$(p+q)(x) = q(x) = (a_0 + b_0) + \cdots + (a_m + b_m)x^m.$$

Therefore

$$\begin{aligned}(p+q)(T) &= \sum_{j\leq m}(a_j + b_j)T^j = \sum_{j\leq m}(a_jT^j + b_jT^j)\\ &= \sum_{j\leq m} a_jT^j + \sum_{j\leq m} b_jT^j = \sum_{j\leq m} a_jT^j + \sum_{j\leq n} b_jT^j = p(T) + q(T).\end{aligned}$$

Next, let us consider pq. By definition we know that $(pq)(x) = \sum_{k\leq m+n} c_k x^k$, where for $k \leq m+n$ we have $c_k := \sum_{i=\alpha}^{\beta} a_i b_{k-i}$, in which $\alpha = \max\{0, k-m\}$ and $\beta = \min\{n, k\}$. Then by the generalized distributivity and Theorem A.68 we have

$$\begin{aligned}p(T)q(T) &= \Big(\sum_{i\leq m} a_iT^i\Big)\Big(\sum_{j\leq n} b_jT^j\Big)\\ &= \sum_{i\leq m}\sum_{j\leq n}(a_iT^i)(b_jT^j) = \sum_{i\leq m}\sum_{j\leq n}(a_ib_j)(T^iT^j)\\ &= \sum_{i\leq m}\sum_{j\leq n}(a_ib_j)T^{i+j} = \sum_{k\leq m+n}\sum_{i+j=k}(a_ib_j)T^k\\ &= \sum_{k\leq m+n}\Big(\sum_{\alpha\leq i\leq\beta} a_ib_{k-i}\Big)T^k = \sum_{k\leq m+n} c_kT^k = (pq)(T).\end{aligned}$$

Finally, to prove the last statement of the theorem, note that we have

$$p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T),$$

because the multiplication of polynomials is commutative. ■

Remark. As a consequence of the above theorem, we can easily show by induction that if $p_1, \ldots, p_k \in F[x]$ then we have

$$\begin{aligned}(p_1 + \cdots + p_k)(T) &= p_1(T) + \cdots + p_k(T),\\ (p_1p_2\cdots p_k)(T) &= p_1(T)p_2(T)\cdots p_k(T).\end{aligned}$$

***Exercise* 4.16.** Suppose W is a T-invariant subspace of V. Show that for every $m \in \mathbb{N}$, W is also T^m-invariant, and we have $(T|_W)^m = T^m|_W$.

Solution. The proof is by induction on m. The case of $m = 1$ is obvious. So suppose the result holds for some m. Let $w \in W$. Then $T^m w \in W$, since by induction

hypothesis W is T^m-invariant. Thus we get $T^{m+1}w = (TT^m)w = T(T^m w) \in W$, because W is T-invariant. Hence W is also T^{m+1}-invariant.

Now set $S := T|_W$. Then by induction hypothesis we know that $S^m = (T|_W)^m = T^m|_W$. Therefore we have

$$\begin{aligned} S^{m+1}w = (SS^m)w = S(S^m w) = S(T^m|_W w) = S(T^m w) \\ = T(T^m w) = (TT^m)w = T^{m+1}w = T^{m+1}|_W w. \end{aligned}$$

Note that $T^m w \in W$, so we can apply the operator $S = T|_W$ to it. Hence we have $(T|_W)^{m+1} = T^{m+1}|_W$, since w was an arbitrary element of W. ■

Proposition 4.17. *Suppose $p \in F[x]$. Then the null space and the image of $p(T)$ are T-invariant subspaces.*

Remark. In particular, the null space and the image of T are T-invariant.

Proof. Let $v \in \operatorname{null} p(T)$. Then we have $p(T)v = 0$. Hence

$$p(T)(Tv) = \big(p(T)T\big)v = \big(Tp(T)\big)v = T\big(p(T)v\big) = T(0) = 0.$$

Thus $Tv \in \operatorname{null} p(T)$,. Therefore $\operatorname{null} p(T)$ is T-invariant.

Now suppose $w \in p(T)(V)$. Then there is $v \in V$ such that $w = p(T)v$. Hence we have

$$Tw = T\big(p(T)v\big) = \big(Tp(T)\big)v = \big(p(T)T\big)v = p(T)(Tv).$$

Thus $Tw \in p(T)(V)$. Therefore $p(T)(V)$ is T-invariant. ■

Theorem 4.18. *Suppose F is an algebraically closed field, and V is a nonzero finite dimensional vector space over F. Then every linear operator $T \in \mathcal{L}(V)$ has at least one eigenvalue.*

Similarly, every matrix $A \in F^{n\times n}$ has at least one eigenvalue.

Remark. This theorem is in particular true when $F = \mathbb{C}$, since $\mathbb{C}$ is algebraically closed.

Proof. Let $n := \dim V$. Let $v \in V$ be a nonzero vector. Then the $n+1$ vectors $v, Tv, T^2v, \ldots, T^n v$ must be linearly dependent. Therefore there are scalars $a_0, a_1, \ldots, a_n \in F$, where at least one of the a_j's is nonzero, such that

$$a_0 v + a_1 Tv + \cdots + a_n T^n v = 0.$$

Let us assume that m is the largest index for which $a_m \neq 0$. Then we have $a_0 v + \cdots + a_m T^m v = 0$. Now consider the polynomial

$$p(x) := a_0 + a_1 x + \cdots + a_m x^m \in F[x].$$

Then we have $p(T)v = 0$.

On the other hand, since F is algebraically closed, there are $c_1, \dots, c_m \in F$ and $c \in F - \{0\}$ such that

$$p(x) = c(x - c_1) \cdots (x - c_m).$$

Hence we have

$$0 = \tfrac{1}{c}0 = \tfrac{1}{c}p(T)v = (T - c_1 I) \cdots (T - c_m I)v.$$

Let j be the smallest index for which we have

$$w_j := (T - c_j I) \cdots (T - c_m I)v \neq 0.$$

Note that $1 < j \leq m$, since $w_1 = \frac{1}{c}p(T)v = 0$. Then we have

$$(T - c_{j-1}I)w_j = (T - c_{j-1}I)(T - c_j I) \cdots (T - c_m I)v = w_{j-1} = 0,$$

because $j - 1 < j$, and by our choice of j we must have $w_{j-1} = 0$. Therefore w_j is an eigenvector of T corresponding to the eigenvalue c_{j-1}, i.e. T has an eigenvalue.

The case of matrices can be proved similarly. Alternatively, for $A \in F^{n \times n}$ we can consider the operator $T \in \mathcal{L}(F^n)$ that maps $x \in F^n$ to $Tx := Ax$. Then we have $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of F^n. Now we know that T has an eigenvalue. So by Theorem 4.4 we can conclude that A has an eigenvalue too. ■

***Example* 4.19.** The above theorem is not true when the field is not algebraically closed. For example the matrix

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2 \times 2},$$

which describes a 90° rotation in $\mathbb{R}^2$, has no real eigenvalue. Because if $\lambda \in \mathbb{R}$ were an eigenvalue of A, then

$$A - \lambda I = \begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix}$$

would be non-invertible. Hence the rank of $A - \lambda I$ must be 1, since it is obviously a nonzero matrix. Thus we must have

$$[-\lambda, -1] = c\,[1, -\lambda],$$

for some $c \in \mathbb{R}$. But this implies that $-\lambda c = 1$ and $-c = -\lambda$. So we must have $-\lambda^2 = 1$ which is impossible. Note that if we consider A as a matrix in $\mathbb{C}^{2 \times 2}$, then we can easily show that $\pm i$ are the eigenvalues of A.

***Example* 4.20.** The above theorem is not true when the vector space is infinite dimensional, even if the field is algebraically closed. For example consider the forward shift $S \in \mathcal{L}(F^\infty)$

$$S : (a_1, a_2, \dots) \mapsto (0, a_1, a_2, \dots).$$

Then S does not have an eigenvalue. To show this, suppose to the contrary that λ is an eigenvalue of S. Then for some nonzero $(a_1, a_2, \dots) \in F^\infty$ we have

$$(0, a_1, a_2, \dots) = S\big((a_1, a_2, \dots)\big) = \lambda(a_1, a_2, \dots) = (\lambda a_1, \lambda a_2, \dots).$$

If $\lambda = 0$ then we must have $a_j = 0$ for every j, which is a contradiction. So suppose $\lambda \neq 0$. Then we have $\lambda a_1 = 0$, and $\lambda a_j = a_{j-1}$ for every $j \geq 2$. Hence we have $a_1 = 0$, and therefore $a_2 = 0$, and therefore $a_3 = 0$, and so forth. Thus again we must have $a_j = 0$ for every j, which is a contradiction. So S cannot have an eigenvalue.

4.2 Diagonalizable Operators

Definition 4.21. Suppose that λ is an eigenvalue of T. The set that consists of $0 \in V$ and all the eigenvectors of T corresponding to λ, is called the **eigenspace** of T corresponding to λ. We denote this set by $E_\lambda(T)$, or simply by E_λ when T is clear from the context.

Remark. It is easy to see that

$$E_\lambda(T) = \operatorname{null}(T - \lambda I),$$

where I is the identity map of V. As a consequence, we see that $E_\lambda(T)$ is a subspace.

Remark. Note that if λ is an eigenvalue of T, then $E_\lambda(T)$ is a nonzero subspace, since it contains at least one nonzero eigenvector of T corresponding to λ. Thus in particular when $E_\lambda(T)$ is finite dimensional we have $\dim E_\lambda(T) \geq 1$.

Remark. Suppose V is finite dimensional, and we know that λ is an eigenvalue of T. Then the above description of $E_\lambda(T)$ as a null space, enables us to easily find a basis for it by using Theorem 3.48.

Proposition 4.22. *The eigenspaces of T are T-invariant subspaces.*

Proof. Suppose λ is an eigenvalue of T. Let $v \in E_\lambda(T)$. Then we have

$$(T - \lambda I)(Tv) = \big((T - \lambda I)T\big)v = \big(T(T - \lambda I)\big)v = T\big((T - \lambda I)v\big) = T(0) = 0.$$

Note that $T, (T - \lambda I)$ commute, since $(T - \lambda I)$ is a polynomial in T. So we have shown that $Tv \in E_\lambda(T)$. Hence $E_\lambda(T)$ is T-invariant. ■

Remark. Note that although every eigenvector of T generates a one dimensional T-invariant subspace, the dimension of the eigenspace $E_\lambda(T)$ can be more than one. Because $E_\lambda(T)$ can contain several linearly independent eigenvectors corresponding to λ.

Remark. Let $W := E_\lambda(T)$ be an eigenspace of T. Then it is obvious that $T|_W = \lambda I_W$.

Theorem 4.23. *Suppose that $\lambda_1, \dots, \lambda_k$ are distinct eigenvalues of T. Then the eigenspaces $E_{\lambda_1}(T), \dots, E_{\lambda_k}(T)$ are independent subspaces.*

Proof. Suppose $v_j \in E_{\lambda_j}$, and $v_1 + \cdots + v_k = 0$. We have to show that $v_j = 0$ for every j. Suppose to the contrary that $v_{j_1}, \dots, v_{j_m}$ are nonzero for some $m > 0$, and the rest of v_j's are zero. Then we have

$$v_{j_1} + \cdots + v_{j_m} = 0. \qquad (*)$$

But v_{j_i} is an eigenvector of T corresponding to the eigenvalue λ_{j_i}, since $v_{j_i} \neq 0$. In addition, $\lambda_{j_1}, \dots, \lambda_{j_m}$ are distinct eigenvalues of T. Therefore $v_{j_1}, \dots, v_{j_m}$ must be linearly independent. However, this is in contradiction with equation $(*)$, because in $(*)$ the coefficient of each v_{j_i} is 1, which is nonzero. Thus m must be zero, and hence every v_j must be zero, as desired. ■

Definition 4.24. Suppose V is finite dimensional. Then the linear operator T is called **diagonalizable** if V has a basis $\mathcal{B}$ such that $[T]_\mathcal{B}$ is a diagonal matrix.

Remark. Note that by Theorems 4.4 and 4.8, the eigenvalues of the diagonalizable operator T are exactly the diagonal entries of the diagonal matrix $[T]_\mathcal{B}$.

Theorem 4.25. *Suppose V is finite dimensional, and $\mathcal{B}$ is a basis for V. Then we have*

(i) *$[T]_\mathcal{B}$ is a diagonal matrix if and only if all the elements of $\mathcal{B}$ are eigenvectors of T.*

(ii) *Suppose that $[T]_\mathcal{B}$ is a diagonal matrix. Let $\lambda_1, \dots, \lambda_k$ be all the distinct eigenvalues of T. Then the diagonal entries of $[T]_\mathcal{B}$ are $\lambda_1, \dots, \lambda_k$. Furthermore, the number of times that each λ_j appears on the diagonal of $[T]_\mathcal{B}$ is equal to $\dim E_{\lambda_j}(T)$.*

Remark. Suppose $[T]_\mathcal{B}$ is diagonal. Then the above theorem implies that every eigenvalue of T appears at least once among the diagonal entries of $[T]_\mathcal{B}$, because $\dim E_{\lambda_j}(T) \geq 1$ for every j. In addition, the following proof shows that for every j there is at least one eigenvector in $\mathcal{B}$ corresponding to λ_j.

Finally, this theorem also implies that if T is diagonalizable, then the diagonal matrix of T is uniquely determined. In other words, if $\mathcal{B}, \mathcal{C}$ are bases for V such

that $[T]_{\mathcal{B}}$ and $[T]_{\mathcal{C}}$ are diagonal, then the diagonal entires of $[T]_{\mathcal{B}}$ and $[T]_{\mathcal{C}}$ are the same, and each diagonal entry appears the same number of times on the diagonals of $[T]_{\mathcal{B}}$ and $[T]_{\mathcal{C}}$. Of course, the arrangement of the diagonal entries in $[T]_{\mathcal{B}}$ and $[T]_{\mathcal{C}}$ can be different.

Proof. Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$. We know that $[v_j]_{\mathcal{B}} = e_j$ for every j.

(i) First suppose $[T]_{\mathcal{B}}$ is diagonal, and $d_1, \ldots, d_n$ are its diagonal entries. Then we have

$$[Tv_j]_{\mathcal{B}} = \big([T]_{\mathcal{B}}\big)_{\cdot,j} = d_j e_j = d_j [v_j]_{\mathcal{B}} = [d_j v_j]_{\mathcal{B}}.$$

Hence we have $Tv_j = d_j v_j$, because the coordinate isomorphism is one-to-one. Therefore v_j is an eigenvector of T, since it is a nonzero vector. In addition note that every d_j is an eigenvalue of T.

Conversely suppose that each v_j is an eigenvector of T. Then we have $Tv_j = d_j v_j$ for some $d_j \in F$. Therefore we get

$$\big([T]_{\mathcal{B}}\big)_{\cdot,j} = [Tv_j]_{\mathcal{B}} = [d_j v_j]_{\mathcal{B}} = d_j [v_j]_{\mathcal{B}} = d_j e_j.$$

Thus $[T]_{\mathcal{B}}$ is diagonal.

(ii) Let $d_1, \ldots, d_n$ be the diagonal entries of $[T]_{\mathcal{B}}$. To simplify the notation let $A := [T]_{\mathcal{B}}$. In the previous part we have shown that every diagonal entry of A is an eigenvalue of T. Now suppose λ is an eigenvalue of T. We have to show that λ appears $\dim E_\lambda$ times on the diagonal of A. We know that λ is also an eigenvalue of $A = [T]_{\mathcal{B}}$. But the eigenvalues of a diagonal matrix are its diagonal entries. So λ appears on the diagonal of A. In particular, note that if $\lambda = d_j$ for some j, then we have $Tv_j = d_j v_j = \lambda v_j$, as shown in the previous part. Thus there is at least one eigenvector in $\mathcal{B}$ corresponding to λ.

Now suppose λ appears r times on the diagonal of A. Let $C := A - \lambda I$. Then the matrix C is diagonal, and exactly r of its diagonal entries are zero. Now if we divide each nonzero row of C by the nonzero diagonal entry in that row, and rearrange the rows so that the nonzero rows lie above the zero rows, we obtain the reduced row echelon form of C. So the rank of C equals $n - r$, since its reduced row echelon form has $n - r$ nonzero rows. On the other hand we have

$$C = A - \lambda I = [T]_{\mathcal{B}} - \lambda [I]_{\mathcal{B}} = [T - \lambda I]_{\mathcal{B}}.$$

Hence $\operatorname{rank}(T - \lambda I) = \operatorname{rank} C = n - r$. Therefore we have

$$\dim E_\lambda = \dim \operatorname{null}(T - \lambda I) = n - \operatorname{rank}(T - \lambda I) = n - (n - r) = r,$$

as desired. ∎

One of the reasons of the importance of diagonalizable operators is that we have a complete understanding of their action on a vector. To see this, suppose T is diagonalizable, and $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V such that $[T]_\mathcal{B}$ is diagonal. We know that the elements of $\mathcal{B}$ are eigenvectors of T. Now let $v \in V$, and suppose that $[v]_\mathcal{B} = [x_1, \ldots, x_n]^\intercal$. Suppose the jj-th diagonal entry of $[T]_\mathcal{B}$ is λ_j. Then as we saw in the previous proof, λ_j is an eigenvector of T, and v_j is an eigenvector corresponding to λ_j. Hence we have

$$Tv = T(x_1v_1 + \cdots + x_nv_n) = x_1Tv_1 + \cdots + x_nTv_n = x_1\lambda_1v_1 + \cdots + x_n\lambda_nv_n.$$

Equivalently, we have

$$[Tv]_\mathcal{B} = [T]_\mathcal{B}[v]_\mathcal{B} = \begin{bmatrix} \lambda_1x_1 \\ \vdots \\ \lambda_nx_n \end{bmatrix}.$$

Therefore the action of a diagonalizable operator on a vector is that it scales each coordinate of that vector, when we represent the vector in a basis consisting of the eigenvectors of the operator. In other words, every diagonalizable operator is the composition of several scalings.

Another advantage of diagonalizable operators is that calculations with diagonal matrices are much simpler than calculations with arbitrary matrices. In addition as we will see below, being diagonalizable is equivalent to having "enough" eigenvalues and eigenvectors.

Theorem 4.26. *Suppose V is finite dimensional. Then the following statements are equivalent.*

(i) *T is diagonalizable.*

(ii) *V has a basis whose elements are eigenvectors of T.*

(iii) *T has distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and*

$$V = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T).$$

(iv) *T has distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and*

$$\dim V = \sum_{j=1}^{k} \dim E_{\lambda_j}(T).$$

Remark. Note that parts (iii) and (iv) of the theorem express that $\lambda_1, \ldots, \lambda_k$ are all the eigenvalues of T, and that they are also distinct.

Remark. Also note that the sum of eigenspaces of any operator is a direct sum, since they are independent subspaces. Hence the nontrivial statement in part (iii) is that the sum of eigenspaces of T is the whole space V.

Proof. **(i) $\iff$ (ii):** By definition, T is diagonalizable if V has a basis $\mathcal{B}$ such that $[T]_{\mathcal{B}}$ is a diagonal matrix. On the other hand, Theorem 4.25 says that $[T]_{\mathcal{B}}$ is a diagonal matrix if and only if all the elements of $\mathcal{B}$ are eigenvectors of T. So T is diagonalizable if and only if V has a basis $\mathcal{B}$ whose elements are eigenvectors of T.

(ii) $\iff$ (iii): Note that the sum of eigenspaces of T is a direct sum, since they are independent subspaces. Let

$$W := E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T).$$

First suppose V has a basis $\mathcal{B}$ whose elements are eigenvectors of T. Then each element of $\mathcal{B}$ belongs to some $E_{\lambda_j}(T) \subset W$, since $\lambda_1, \ldots, \lambda_k$ are all the eigenvalues of T. Therefore we get

$$V = \operatorname{span}(\mathcal{B}) \subset W \subset V,$$

since W is a subspace. Thus $W = V$ as desired.

Conversely suppose that $W = V$. Let $\mathcal{B}_j$ be a basis for $E_{\lambda_j}(T)$ for each j. Then by Theorem 2.55 we know that $\mathcal{B} := \bigcup_{j \le k} \mathcal{B}_j$ is a basis for V. In addition it is obvious that every element of $\mathcal{B}$ is an eigenvector of T. Furthermore, as we mentioned in the remark after Theorem 4.25, every eigenvalue of T must have a corresponding eigenvector in $\mathcal{B}$. Hence $\lambda_1, \ldots, \lambda_k$ are all the eigenvalues of T.

(iii) $\iff$ (iv): Let W be as above. Then Theorem 2.55 implies that $\dim W = \sum_{j=1}^{k} \dim E_{\lambda_j}(T)$. Now suppose $W = V$. Then we get

$$\dim V = \dim W = \sum_{j=1}^{k} \dim E_{\lambda_j}(T).$$

Conversely suppose $\dim V = \sum_{j=1}^{k} \dim E_{\lambda_j}(T)$. Then we have $\dim V = \dim W$. Therefore by Theorem 2.44 we must have $W = V$, since V is finite dimensional. ■

Remark. If we know the eigenvalues of the operator T, then part (iv) of the above theorem provides us an algorithm to determine whether T is diagonalizable or not. Suppose A is the matrix of T in some basis. Let λ be one of the eigenvalues of T. Then we know that λ is also an eigenvalue of A. On the other hand, $A - \lambda I$ is the matrix of $T - \lambda I$. Therefore $A - \lambda I$ and $T - \lambda I$ have the same rank. Hence we have

$$\dim E_\lambda(T) = \dim \operatorname{null}(T - \lambda I) = n - \dim(T - \lambda I)(V) = n - \operatorname{rank}(A - \lambda I),$$

where $n = \dim V$. So in order to check the diagonalizability of T, we only need to compute $n - \operatorname{rank}(A - \lambda I)$ for each eigenvalue of T, and then check whether their sum is equal to n or not.

Finally, suppose we have shown that T is diagonalizable, and $\lambda_1, \ldots, \lambda_k$ are all the distinct eigenvalues of T. Then in order to find a basis $\mathcal{B}$ for V so that

$[T]_{\mathcal{B}}$ is diagonal, we can find a basis for each $E_{\lambda_j}(T)$ by using Theorem 3.48, and then take the union of those bases. Because that union is a basis for V, since $V = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T)$. In addition, we know that the elements of every $E_{\lambda_j}(T)$ are eigenvectors of T, so the elements of the constructed basis are eigenvectors of T too. Hence Theorem 4.25 implies that $[T]_{\mathcal{B}}$ is diagonal. Let us employ this method in the next two examples. ■

***Example* 4.27.** Let

$$A = \begin{bmatrix} 4 & 0 & 6 \\ 0 & 4 & 3 \\ 0 & 0 & 1 \end{bmatrix},$$

and consider the operator $T \in \mathcal{L}(\mathbb{R}^3)$ defined by $T(x) := Ax$ for $x \in \mathbb{R}^3$. Let us show that T is diagonalizable. We know that $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of $\mathbb{R}^3$. Thus the eigenvalues of T and A are the same. Hence $1, 4$ are the only eigenvalues of T, since A is upper triangular. In addition we have

$$A - I = \begin{bmatrix} 3 & 0 & 6 \\ 0 & 3 & 3 \\ 0 & 0 & 0 \end{bmatrix}, \qquad A - 4I = \begin{bmatrix} 0 & 0 & 6 \\ 0 & 0 & 3 \\ 0 & 0 & -3 \end{bmatrix}.$$

If we perform elementary row operations, we find the reduced row echelon form of the above matrices to be

$$A - I \to \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad A - 4I \to \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We know that the rank of a matrix is the number of nonzero rows in its reduced row echelon form. Thus we have $\operatorname{rank}(A - I) = 2$ and $\operatorname{rank}(A - 4I) = 1$. Hence as we explained in the above remark, we have

$$\begin{aligned} \dim E_1(T) &= 3 - \operatorname{rank}(A - I) = 3 - 2 = 1, \\ \dim E_4(T) &= 3 - \operatorname{rank}(A - 4I) = 3 - 1 = 2. \end{aligned}$$

Therefore we have

$$\dim E_1(T) + \dim E_4(T) = 3 = \dim \mathbb{R}^3.$$

Thus T is diagonalizable.

In addition note that the eigenvectors of T and A are also the same, because $Tx = Ax$ for every $x \in \mathbb{R}^3$. Now by using Theorem 3.47, it is easy to find a basis for the set of solutions of the homogeneous linear systems $(A - I)x = 0$ and

$(A - 4I)x = 0$, which are equal to $E_1(T)$ and $E_4(T)$ respectively. Hence we find the basis $[2, 1, -1]^\intercal$ for $E_1(T)$, and the basis e_1, e_2 for $E_4(T)$. Then

$$\mathcal{B} := \{[2, 1, -1]^\intercal, e_1, e_2\}$$

is a basis for $\mathbb{R}^3$ such that $[T]_\mathcal{B}$ is diagonal. Now we can use direct computation, or use Theorem 4.25, to deduce that

$$[T]_\mathcal{B} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

***Example* 4.28.** Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

and consider the operator $T \in \mathcal{L}(F^2)$ defined by $T(x) := Ax$ for $x \in F^2$. Let us show that T is not diagonalizable. We know that $[T]_\mathcal{B} = A$, where $\mathcal{B}$ is the standard basis of F^2. Thus the eigenvalues of T and A are the same. Hence 1 is the only eigenvalue of T, since A is upper triangular. In addition we have

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

Thus $A - I$ is a matrix in reduced row echelon form. So its rank is equal to the number of its nonzero rows, i.e. $\operatorname{rank}(A - I) = 1$. Hence as we explained in the above remark, we have

$$\dim E_1(T) = 2 - \operatorname{rank}(A - I) = 2 - 1 = 1 < 2 = \dim F^2.$$

Therefore T is not diagonalizable. Note that T does not have "enough" eigenvectors to generate its domain F^2.

Proposition 4.29. *Suppose V is finite dimensional, and $n = \dim V$. If T has n distinct eigenvalues, then T is diagonalizable.*

Proof. Suppose $\lambda_1, \ldots, \lambda_n$ are the distinct eigenvalues of T. Then we have

$$n \leq \sum_{j=1}^{n} \dim E_{\lambda_j}(T) \leq \dim V = n,$$

since $\dim E_{\lambda_j}(T) \geq 1$ for every $j \leq n$. Thus

$$\sum_{j=1}^{n} \dim E_{\lambda_j}(T) = n = \dim V.$$

Therefore T is diagonalizable due to the previous theorem. ■

Theorem 4.30. *Suppose $A, B \in F^{n\times n}$ are similar matrices. Then $\lambda \in F$ is an eigenvalue of A if and only if it is an eigenvalue of B.*

Proof. Suppose λ is an eigenvalue of A. Then there is a nonzero $x \in F^n$ such that $Ax = \lambda x$. On the other hand, A, B are similar. Therefore there is an invertible matrix $C \in F^{n\times n}$ so that $B = C^{-1}AC$. Let $y := C^{-1}x \in F^n$. Now we have

$$\begin{aligned} By = B(C^{-1}x) &= C^{-1}ACC^{-1}x = C^{-1}AIx \\ &= C^{-1}Ax = C^{-1}(\lambda x) = \lambda C^{-1}x = \lambda y. \end{aligned}$$

In addition we must have $y = C^{-1}x \neq 0$, since if $C^{-1}x = 0$ then we would have

$$x = Ix = CC^{-1}x = C0 = 0,$$

which is contrary to our assumption. Therefore λ is also an eigenvalue of B. The converse holds trivially, because we have $A = CBC^{-1} = D^{-1}BD$, where $D := C^{-1}$. ∎

Theorem 4.31. *Let $A \in F^{n\times n}$. Suppose F^n has a basis consisting of the eigenvectors of A. Let $C \in F^{n\times n}$ be the matrix whose columns are this basis of eigenvectors. Then C is invertible, and $C^{-1}AC$ is a diagonal matrix whose diagonal entries are the eigenvalues of A.*

Remark. In other words A is similar to a diagonal matrix. Also, note that the last sentence of the theorem means that all the eigenvalues of A appear on the diagonal of $C^{-1}AC$, and every diagonal entry of $C^{-1}AC$ is an eigenvalue of A.

Proof. First note that

$$\operatorname{rank} C = \dim \operatorname{span}(C_{\cdot,1}, \ldots, C_{\cdot,n}) = n,$$

since the columns of C form a basis for F^n. So C is invertible by Theorem 3.49. In addition we know that for every j there is $\lambda_j \in F$ such that $AC_{\cdot,j} = \lambda_j C_{\cdot,j}$. Hence we have

$$\begin{aligned} (C^{-1}AC)_{\cdot,j} = (C^{-1}AC)e_j &= C^{-1}A(Ce_j) = C^{-1}AC_{\cdot,j} \\ &= C^{-1}(\lambda_j C_{\cdot,j}) = \lambda_j C^{-1}C_{\cdot,j} = \lambda_j (C^{-1}C)_{\cdot,j} = \lambda_j I_{\cdot,j} = \lambda_j e_j. \end{aligned}$$

Note that we have used Theorem 1.16 in the above line. Therefore $C^{-1}AC$ is a diagonal matrix, as shown in Exercise 1.19. But by Theorem 4.8, the diagonal entries of a diagonal matrix are the same as its eigenvalues. Furthermore, the eigenvalues of similar matrices are the same, as shown in the previous theorem. Thus the diagonal entries of $C^{-1}AC$ are exactly the eigenvalues of A. ∎

Chapter 5

Inner Product Spaces

5.1 Inner Products and Norms

In this chapter, we are going to study the notions of length of a vector, and angle between two vectors. To do this, we need to equip our vector space with an inner product, as defined below. Let us also mention that we are only interested in the notion of right angle between two vectors, i.e. when the two vectors are orthogonal. So we will not assign a measure of angle between two arbitrary vectors, however it is not hard to do so with the tools we develop in this chapter.

Definition 5.1. Suppose the field F is either $\mathbb{R}$ or $\mathbb{C}$. Let V be a vector space over F. An **inner product** on V is a function

$$\langle\ ,\ \rangle : V \times V \to F$$

that satisfies

(i) $\langle\ ,\ \rangle$ is *positive definite*, which means that for all $v \in V$ we have

$$\langle v, v\rangle \geq 0,$$

i.e. $\langle v, v\rangle$ is a nonnegative real number; and in addition for all $v \in V$ we have

$$\langle v, v\rangle = 0 \implies v = 0.$$

(ii) $\langle\ ,\ \rangle$ is *conjugate symmetric*, i.e. for all $u, v \in V$ we have

$$\langle u, v\rangle = \overline{\langle v, u\rangle}.$$

(iii) $\langle\ ,\ \rangle$ is linear with respect to its first variable, i.e. for all $u, v, w \in V$ and $a \in F$ we have

$$\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle, \qquad \langle av, w\rangle = a\langle v, w\rangle.$$

A vector space equipped with an inner product is called an **inner product space**.

Remark. When $F = \mathbb{R}$, inner products are symmetric, i.e. for all $u, v \in V$ we have

$$\langle u, v\rangle = \langle v, u\rangle.$$

Because the conjugate of a real number equals itself. But when $F = \mathbb{C}$, we cannot require inner products to be symmetric, since if we do that we will lose the positive definiteness. To see this suppose $v \in V$ is a nonzero vector. Then we must have $\langle v, v\rangle > 0$. Now consider the vector iv, where $i^2 = -1$. If we assume that the inner product is symmetric we get

$$\langle iv, iv\rangle = i\langle v, iv\rangle = i\langle iv, v\rangle = i^2\langle v, v\rangle = -\langle v, v\rangle < 0,$$

which is a contraction.

Remark. We can easily show that for all $v \in V$ we have

$$\langle v, 0\rangle = 0 = \langle 0, v\rangle.$$

Because $\langle 0, v\rangle = \langle 0 + 0, v\rangle = \langle 0, v\rangle + \langle 0, v\rangle$, so $\langle 0, v\rangle = 0$. On the other hand we have $\langle v, 0\rangle = \overline{\langle 0, v\rangle} = \bar{0} = 0$, as desired. As a particular case we have $\langle 0, 0\rangle = 0$. Hence we can write

$$\langle v, v\rangle = 0 \iff v = 0.$$

Remark. In general, an inner product is not linear with respect to its second variable. But it is always *conjugate linear* with respect to its second variable, i.e. for all $u, v, w \in V$ and $a \in F$ we have

$$\langle w, u + v\rangle = \langle w, u\rangle + \langle w, v\rangle, \qquad \langle w, av\rangle = \bar{a}\langle w, v\rangle.$$

Because we have

$$\begin{aligned}\langle w, u + av\rangle &= \overline{\langle u + av, w\rangle} = \overline{\langle u, w\rangle + a\langle v, w\rangle} \\ &= \overline{\langle u, w\rangle} + \bar{a}\overline{\langle v, w\rangle} = \langle w, u\rangle + \bar{a}\langle w, v\rangle.\end{aligned}$$

Now by putting $a = 1$ or $u = 0$, we get the above identities. Note that when $F = \mathbb{R}$, the conjugate linearity is the same as linearity, since $\bar{a} = a$ for all $a \in \mathbb{R}$. Therefore inner products on real vector spaces are also linear with respect to their second variables.

Remark. It is easy to show by induction that for all $u_j, v_j \in V$ and $a_j \in F$ we have

$$\begin{aligned}\langle a_1u_1 + \cdots + a_ku_k, v\rangle &= a_1\langle u_1, v\rangle + \cdots + a_k\langle u_k, v\rangle, \\ \langle u, a_1v_1 + \cdots + a_kv_k\rangle &= \bar{a}_1\langle u, v_1\rangle + \cdots + \bar{a}_k\langle u, v_k\rangle.\end{aligned}$$

***Example* 5.2.** The *standard inner products* on $\mathbb{R}^n$ and $\mathbb{C}^n$ are the prototypes of inner products. Let $x, y \in \mathbb{R}^n$ and $z, w \in \mathbb{C}^n$. Then their inner products is defined as follows

$$\langle x, y\rangle = x_1y_1 + \cdots + x_ny_n, \qquad \langle w, z\rangle = w_1\bar{z}_1 + \cdots + w_n\bar{z}_n.$$

To prove that these are indeed inner products, note that their conjugate symmetry and positive definiteness follow easily from the definition. Now remember that we consider x, y, w, z to be column vectors, so we have the following expressions for the inner products

$$\langle x, y\rangle = y^{\mathsf{T}}x = x^{\mathsf{T}}y, \qquad \langle w, z\rangle = z^*w = w^{\mathsf{T}}\bar{z},$$

where z^* is the conjugate transpose of z as defined in Definition 1.26, and $\bar{z}$ is the column vector whose j-th component is $\bar{z}_j$. These expressions make it obvious that the standard inner products are also linear with respect to the first variable, due to the properties of matrix multiplication.

Remark. Let us mention a simple fact about $\mathbb{C}^n$, which will be useful later. Let $z \in \mathbb{C}^n$. Then for every $j \leq n$ there are $x_j, y_j \in \mathbb{R}$ such that $z_j = x_j + iy_j$. Let

$$x := [x_1, \ldots, x_n]^{\mathsf{T}},\ y := [y_1, \ldots, y_n]^{\mathsf{T}} \in \mathbb{R}^n.$$

Then it is trivial to check that $z = x + iy$.

***Example* 5.3.** Suppose V is an inner product space, and W is a subspace of V. Then W is also an inner product space with the inner product inherited from V, i.e. if we denote the inner product of V by $\langle\ ,\ \rangle$, then it is easy to see that $\langle\ ,\ \rangle|_{W\times W}$ is an inner product on W.

Notation. In the rest of this chapter, we assume that F is either $\mathbb{R}$ or $\mathbb{C}$, and V is a nonzero inner product space over F with the inner product $\langle\ ,\ \rangle$. Also, we always assume that F^n is equipped with its standard inner product, unless otherwise specified.

Definition 5.4. The **norm** of a vector $v \in V$ is the nonnegative real number

$$\|v\| := \sqrt{\langle v, v\rangle}.$$

***Example* 5.5.** Let $z \in F^n$. Then the norm of z with respect to the standard inner product is

$$\|z\| = \sqrt{|z_1|^2 + \cdots + |z_n|^2}.$$

Proposition 5.6. *For every $v \in V$ and $a \in F$ we have*

(i) $\|v\| \geq 0$, *and* $\|v\| = 0 \iff v = 0$.

(ii) $\|av\| = |a|\|v\|$.

Proof. **(i)** Since $\langle v, v\rangle \geq 0$, we have $\|v\| = \sqrt{\langle v, v\rangle} \geq 0$. In addition we have

$$\|v\| = 0 \iff \langle v, v\rangle = 0 \iff v = 0.$$

(ii) We have

$$\|av\| = \sqrt{\langle av, av\rangle} = \sqrt{a\bar{a}\langle v, v\rangle} = \sqrt{|a|^2\langle v, v\rangle} = |a|\sqrt{\langle v, v\rangle} = |a|\|v\|. \qquad \blacksquare$$

Definition 5.7. Let $u, v \in V$. We say u, v are **orthogonal** if $\langle u, v\rangle = 0$.

Remark. Note that if $\langle u, v\rangle = 0$ then we also have $\langle v, u\rangle = \overline{\langle u, v\rangle} = \bar{0} = 0$.

Remark. As we have seen before, the zero vector is orthogonal to every vector. It is easy to see that if a vector v is orthogonal to every vector, then v must be zero. Because v must be orthogonal to itself too, i.e. $\langle v, v\rangle = 0$. Therefore $v = 0$.

Pythagorean Theorem. *Suppose $u, v \in V$ are orthogonal. Then we have*

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

Proof. We know that $\langle u, v\rangle = 0 = \langle v, u\rangle$. Thus we have

$$\begin{aligned}\|u + v\|^2 = \langle u + v, u + v\rangle &= \langle u, u + v\rangle + \langle v, u + v\rangle \\ &= \langle u, u\rangle + \langle u, v\rangle + \langle v, u\rangle + \langle v, v\rangle \\ &= \langle u, u\rangle + \langle v, v\rangle = \|u\|^2 + \|v\|^2.\end{aligned} \qquad \blacksquare$$

Cauchy-Schwarz Inequality. *For every $u, v \in V$ we have*

$$|\langle u, v\rangle| \leq \|u\|\|v\|.$$

Proof. If $v = 0$ then the inequality holds trivially. So suppose that $v \neq 0$. Let

$$w := u - \frac{\langle u, v\rangle}{\|v\|^2} v.$$

Let $a := \frac{\langle u,v\rangle}{\|v\|^2}$. Then we have

$$\begin{aligned}\langle w, v\rangle &= \langle u - av, v\rangle = \langle u, v\rangle - a\langle v, v\rangle \\ &= \langle u, v\rangle - \frac{\langle u, v\rangle}{\|v\|^2}\|v\|^2 = \langle u, v\rangle - \langle u, v\rangle = 0.\end{aligned}$$

Therefore we also have $\langle w, av\rangle = \bar{a}\langle w, v\rangle = 0$. Thus the Pythagorean theorem implies that

$$\begin{aligned}\|u\|^2 = \|w + av\|^2 = \|w\|^2 + \|av\|^2 \geq \|av\|^2 = |a|^2\|v\|^2 \\ = \frac{|\langle u, v\rangle|^2}{\|v\|^4}\|v\|^2 = \frac{|\langle u, v\rangle|^2}{\|v\|^2}.\end{aligned}$$

So we get $|\langle u, v\rangle|^2 \leq \|u\|^2\|v\|^2$, which implies the desired inequality. ■

Second Proof. First note that $\langle u, v\rangle + \langle v, u\rangle = \langle u, v\rangle + \overline{\langle u, v\rangle} = 2\mathrm{Re}\big(\langle u, v\rangle\big)$. If $v = 0$ then the inequality holds trivially. So suppose that $v \neq 0$. Now let $t \in \mathbb{R}$. Then we have

$$\begin{aligned}0 \leq \|u + tv\|^2 &= \langle u + tv, u + tv\rangle \\ &= \langle u, u\rangle + \langle u, tv\rangle + \langle tv, u\rangle + \langle tv, tv\rangle \\ &= \langle u, u\rangle + t\langle u, v\rangle + t\langle v, u\rangle + t^2\langle v, v\rangle \\ &= \|u\|^2 + 2t\mathrm{Re}\big(\langle u, v\rangle\big) + t^2\|v\|^2.\end{aligned}$$

Since this inequality holds for all $t \in \mathbb{R}$, and $\|v\|^2 > 0$, the discriminant of the above quadratic function in t must be nonpositive, i.e. $\big[\mathrm{Re}\big(\langle u, v\rangle\big)\big]^2 - \|u\|^2\|v\|^2 \leq 0$. So we get

$$\big|\mathrm{Re}\big(\langle u, v\rangle\big)\big| \leq \|u\|\|v\|. \qquad (*)$$

If $\langle u, v\rangle \in \mathbb{R}$ then $\langle u, v\rangle = \mathrm{Re}\big(\langle u, v\rangle\big)$ and we have the desired inequality. Otherwise there are $r, \theta \in \mathbb{R}$ so that $\langle u, v\rangle = re^{i\theta}$. Note that we must have $r = |\langle u, v\rangle|$, since $|e^{i\theta}| = 1$. Then we have

$$\langle u, e^{i\theta}v\rangle = \overline{e^{i\theta}}\langle u, v\rangle = e^{-i\theta}\langle u, v\rangle = e^{-i\theta}re^{i\theta} = r \in \mathbb{R}.$$

Thus $r = \mathrm{Re}\big(\langle u, e^{i\theta}v\rangle\big)$. Hence the inequality $(*)$ implies that

$$|\langle u, v\rangle| = r = \big|\mathrm{Re}\big(\langle u, e^{i\theta}v\rangle\big)\big| \leq \|u\|\|e^{i\theta}v\| = \|u\||e^{i\theta}|\|v\| = \|u\|\|v\|,$$

as desired. ■

Remark. The vector $\frac{\langle u,v\rangle}{\|v\|^2}v$ in the first proof above is the orthogonal projection of the vector u on the subspace generated by v. We will study the orthogonal projections later in this chapter. See Example 5.26.

***Example* 5.8.** Let $x, y \in \mathbb{R}^n$. Then the Cauchy-Schwarz inequality for the standard inner product of $\mathbb{R}^n$ can be written as

$$|x_1y_1 + \cdots + x_ny_n|^2 \leq (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2).$$

Triangle Inequality. *For every $u, v \in V$ we have*

$$\|u+v\| \le \|u\| + \|v\|.$$

Proof. By the Cauchy-Schwarz inequality we have

$$\begin{aligned}\|u+v\|^2 &= \langle u+v, u+v\rangle = \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle \\ &\le \|u\|^2 + \|u\|\|v\| + \|v\|\|u\| + \|v\|^2 = \big(\|u\| + \|v\|\big)^2.\end{aligned}$$

Thus we get the desired inequality. ■

Remark. It is easy to show by induction that

$$\|v_1 + \cdots + v_k\| \le \|v_1\| + \cdots + \|v_k\|.$$

***Example* 5.9.** Let $x, y \in \mathbb{R}^n$. Then the triangle inequality for the standard norm of $\mathbb{R}^n$ can be written as

$$\sqrt{(x_1+y_1)^2 + \cdots + (x_n+y_n)^2} \le \sqrt{x_1^2 + \cdots + x_n^2} + \sqrt{y_1^2 + \cdots + y_n^2}.$$

Remark. The first part of the next theorem is the converse of the Pythagorean theorem when $F = \mathbb{R}$. The other two parts tell us when we have the equality in the Cauchy-Schwarz and triangle inequalities.

Theorem 5.10. *Let $u, v \in V$.*

(i) *If $F = \mathbb{R}$, and $\|u+v\|^2 = \|u\|^2 + \|v\|^2$, then $\langle u, v\rangle = 0$.*
(ii) *If $|\langle u, v\rangle| = \|u\|\|v\|$ then either $u = av$, or $v = au$, for some $a \in F$.*
(iii) *If $\|u+v\| = \|u\| + \|v\|$ then either $u = av$, or $v = au$, for some $a \in [0, \infty)$.*

Proof. **(i)** Since $F = \mathbb{R}$ we have $\langle u, v\rangle = \langle v, u\rangle$. Therefore we have

$$\begin{aligned}\|u\|^2 + \|v\|^2 = \|u+v\|^2 &= \langle u+v, u+v\rangle \\ &= \langle u,u\rangle + 2\langle u,v\rangle + \langle v,v\rangle = \|u\|^2 + 2\langle u,v\rangle + \|v\|^2.\end{aligned}$$

Thus we must have $\langle u, v\rangle = 0$.

(ii) If $v = 0$ then we have $v = 0u$. So suppose that $v \neq 0$. Let $a := \frac{\langle u,v\rangle}{\|v\|^2}$ and $w := u - av$. Now look at the first proof of the Cauchy-Schwarz inequality. In that proof, the only inequality is $\|w\|^2 + \|av\|^2 \ge \|av\|^2$. So if we have the equality instead of the inequality, we must have $\|w\| = 0$. Thus $w = 0$, and hence we get $u = av$, as desired.

(iii) Remember that $\langle u, v\rangle + \langle v, u\rangle = \langle u, v\rangle + \overline{\langle u, v\rangle} = 2\mathrm{Re}\big(\langle u, v\rangle\big)$. Now we have

$$\begin{aligned}\|u\|^2 + 2\|u\|\|v\| + \|v\|^2 &= \big(\|u\| + \|v\|\big)^2 = \|u+v\|^2 \\ &= \langle u+v, u+v\rangle = \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle \\ &= \|u\|^2 + 2\mathrm{Re}\big(\langle u,v\rangle\big) + \|v\|^2.\end{aligned}$$

Hence we get $\|u\|\|v\| = \mathrm{Re}\big(\langle u, v\rangle\big) \leq |\langle u, v\rangle| \leq \|u\|\|v\|$. Thus we have the equality in the Cauchy-Schwarz inequality. Therefore we either $u = av$, or $v = au$, for some $a \in F$. In addition we have $|\langle u, v\rangle| = \mathrm{Re}\big(\langle u, v\rangle\big)$. Thus we must have $\mathrm{Im}\big(\langle u, v\rangle\big) = 0$. So $\langle u, v\rangle$ is a real number. It is also nonnegative, since $\langle u, v\rangle = \mathrm{Re}\big(\langle u, v\rangle\big) = |\langle u, v\rangle|$. Now suppose $v = au$, the other case is similar. If $u = 0$ then $v = 0$, and we can set $a = 0$. So suppose $u \neq 0$. Then $\langle u, u\rangle = \|u\|^2 \in (0, \infty)$, and we have

$$\bar{a} = \bar{a}\frac{\langle u, u\rangle}{\langle u, u\rangle} = \frac{\langle u, au\rangle}{\langle u, u\rangle} = \frac{\langle u, v\rangle}{\langle u, u\rangle} \in [0, \infty).$$

Therefore $a \in [0, \infty)$ as desired. ■

Parallelogram Law. *For every $u, v \in V$ we have*

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2.$$

Remark. The reason that this identity is called the parallelogram law is that in a parallelogram, the sum of the squares of the lengths of the diagonals equals the sum of the squares of the lengths of the four sides.

Proof. We have

$$\begin{aligned}\|u + v\|^2 + \|u - v\|^2 &= \langle u + v, u + v\rangle + \langle u - v, u - v\rangle \\ &= \langle u, u\rangle + \langle u, v\rangle + \langle v, u\rangle + \langle v, v\rangle \\ &\qquad + \langle u, u\rangle - \langle u, v\rangle - \langle v, u\rangle + \langle v, v\rangle \\ &= 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

■

Remark. A *norm* on a real or complex vector space V is a function $\| \ \| : V \to [0, \infty)$ that satisfies the properties listed in Proposition 5.6 and the triangle inequality. Although here we only study those norms that are associated to an inner product, there are norms that are not induced by any inner product. It can be shown that a norm is induced by an inner product if and only if it satisfies the parallelogram law. The proof of this fact is a little tricky, and we do not present it here, but let us mention that the starting point is to define our candidate for the inner product as in the next theorem, and then to show that it is indeed an inner product which induces our norm.

Polarization Identities. *Let $u, v \in V$.*

(i) *When $F = \mathbb{R}$ we have*

$$\langle u, v\rangle = \frac{1}{4}\|u + v\|^2 - \frac{1}{4}\|u - v\|^2.$$

(ii) *When $F = \mathbb{C}$ we have*

$$\langle u, v\rangle = \frac{1}{4}\|u+v\|^2 - \frac{1}{4}\|u-v\|^2 + \frac{i}{4}\|u+iv\|^2 - \frac{i}{4}\|u-iv\|^2.$$

Remark. These identities are useful when we want to express the inner product in terms of the norm. For example, we will use them in Section 6.4 to show that an operator preserves the norm if and only if it preserves the inner product.

Proof. **(i)** We have

$$\begin{aligned}\|u+v\|^2 - \|u-v\|^2 &= \langle u+v, u+v\rangle - \langle u-v, u-v\rangle \\ &= \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle \\ &\qquad - \big(\langle u,u\rangle - \langle u,v\rangle - \langle v,u\rangle + \langle v,v\rangle\big) \\ &= 2\langle u,v\rangle + 2\langle v,u\rangle.\end{aligned}$$

Since $F = \mathbb{R}$ we have $\langle u,v\rangle = \langle v,u\rangle$. So we get $\|u+v\|^2 - \|u-v\|^2 = 4\langle u,v\rangle$.

(ii) We have shown that $\|u+v\|^2 - \|u-v\|^2 = 2\langle u,v\rangle + 2\langle v,u\rangle$. If we replace v by iv in this identity, we get

$$\begin{aligned}\|u+iv\|^2 - \|u-iv\|^2 &= 2\langle u, iv\rangle + 2\langle iv, u\rangle \\ &= 2\bar{i}\langle u,v\rangle + 2i\langle v,u\rangle = -2i\langle u,v\rangle + 2i\langle v,u\rangle.\end{aligned}$$

Now if we multiply the above equation by i, and then add it to the first equation, we obtain

$$\begin{aligned}&\|u+v\|^2 - \|u-v\|^2 + i\|u+iv\|^2 - i\|u-iv\|^2 \\ &\quad = 2\langle u,v\rangle + 2\langle v,u\rangle - 2i^2\langle u,v\rangle + 2i^2\langle v,u\rangle \\ &\quad = 2\langle u,v\rangle + 2\langle v,u\rangle + 2\langle u,v\rangle - 2\langle v,u\rangle = 4\langle u,v\rangle.\end{aligned}$$

■

5.2 Orthonormal Bases

Definition 5.11. A list of vectors $v_1, \ldots, v_m \in V$ is called **orthonormal** if $\|v_j\| = 1$ for every j, and $\langle v_j, v_k\rangle = 0$ for every $j \neq k$. We also consider the empty list to be orthonormal. An **orthonormal basis** for V is a basis for V which is also orthonormal.

Remark. In other words, a list of vectors $v_1, \ldots, v_m$ is orthonormal if

$$\langle v_j, v_k\rangle = \begin{cases} 1 & j = k, \\ 0 & j \neq k. \end{cases}$$

***Example* 5.12.** It is easy to see that the standard bases of $\mathbb{R}^n, \mathbb{C}^n$ are orthonormal bases.

Theorem 5.13. *An orthonormal list of vectors is linearly independent.*

Remark. This theorem provides a link between orthonormality, which is defined using the inner product, and linear independence, which is defined using the linear structure of the vector space.

Proof. We know that the empty list is linearly independent by definition. Now suppose $v_1, \dots, v_m \in V$ is an orthonormal list. Suppose $a_1 v_1 + \cdots + a_m v_m = 0$ for some scalars $a_j \in F$. Then for every j we have

$$\begin{aligned} 0 = \langle 0, v_j \rangle &= \langle a_1 v_1 + \cdots + a_m v_m, v_j \rangle = a_1 \langle v_1, v_j \rangle + \cdots + a_m \langle v_m, v_j \rangle \\ &= a_1 0 + \cdots + a_{j-1} 0 + a_j 1 + a_{j+1} 0 + \cdots + a_m 0 = a_j. \end{aligned}$$

Thus $v_1, \dots, v_m$ are linearly independent. ■

Remark. In the light of the above theorem, an orthonormal list of vectors $v_1, \dots, v_n$ is an orthonormal basis for V, if it generates V.

Theorem 5.14. *Suppose $v_1, \dots, v_n$ is an orthonormal basis for V. Let*

$$v = a_1 v_1 + \cdots + a_n v_n, \qquad w = b_1 v_1 + \cdots + b_n v_n,$$

where $a_j, b_j \in F$. Then we have

(i) $a_j = \langle v, v_j \rangle$ *for every* $j \leq n$.

(ii) $\|v\|^2 = |a_1|^2 + \cdots + |a_n|^2$.

(iii) $\langle v, w \rangle = a_1 \bar{b}_1 + \cdots + a_n \bar{b}_n$.

Remark. If we denote the orthonormal basis by $\mathcal{B}$, then we have

$$\langle v, w \rangle = a_1 \bar{b}_1 + \cdots + a_n \bar{b}_n = [w]_{\mathcal{B}}^* [v]_{\mathcal{B}}.$$

In other words, the inner product of V corresponds to the standard inner product of F^n, when we use the coordinates with respect to an orthonormal basis.

Proof. **(i)** We have

$$\begin{aligned} \langle v, v_j \rangle &= \langle a_1 v_1 + \cdots + a_n v_n, v_j \rangle = a_1 \langle v_1, v_j \rangle + \cdots + a_n \langle v_n, v_j \rangle \\ &= a_1 0 + \cdots + a_{j-1} 0 + a_j 1 + a_{j+1} 0 + \cdots + a_n 0 = a_j. \end{aligned}$$

(ii) We know that $\langle v, v_j\rangle = a_j$. Therefore we have

$$\begin{aligned}\|v\|^2 = \langle v, v\rangle &= \langle v, a_1v_1 + \cdots + a_nv_n\rangle \\ &= \bar{a}_1\langle v, v_1\rangle + \cdots + \bar{a}_n\langle v, v_n\rangle \\ &= \bar{a}_1a_1 + \cdots + \bar{a}_na_n = |a_1|^2 + \cdots + |a_n|^2.\end{aligned}$$

(iii) We know that $\langle v, v_j\rangle = a_j$. Therefore we have

$$\begin{aligned}\langle v, w\rangle = \langle v, b_1v_1 + \cdots + b_nv_n\rangle &= \bar{b}_1\langle v, v_1\rangle + \cdots + \bar{b}_n\langle v, v_n\rangle \\ = \bar{b}_1a_1 + \cdots + \bar{b}_na_n &= a_1\bar{b}_1 + \cdots + a_n\bar{b}_n.\end{aligned}$$

■

Remark. It is easy to see that the relations in the above theorem do not necessarily hold when the basis is not orthonormal.

Gram-Schmidt Process. *Suppose $v_1, \ldots, v_m \in V$ are linearly independent. We inductively define the vectors $w_1, \ldots, w_m$ as follows. Set*

$$\tilde{w}_1 := v_1, \qquad w_1 := \frac{1}{\|\tilde{w}_1\|}\tilde{w}_1 = \frac{1}{\|v_1\|}v_1,$$

and for $k \geq 1$ set

$$\tilde{w}_{k+1} := v_{k+1} - \sum_{j=1}^{k}\langle v_{k+1}, w_j\rangle w_j, \qquad w_{k+1} := \frac{1}{\|\tilde{w}_{k+1}\|}\tilde{w}_{k+1}.$$

Then $w_1, \ldots, w_m$ are orthonormal, and for $k \leq m$ we have

$$\operatorname{span}(w_1, \ldots, w_k) = \operatorname{span}(v_1, \ldots, v_k).$$

Remark. We will see that since $v_1, \ldots, v_m$ are linearly independent, all the vectors $\tilde{w}_1, \ldots, \tilde{w}_m$ are nonzero. Thus their norms are nonzero, and therefore $w_1, \ldots, w_m$ exist. However, it can be shown that if we apply the Gram-Schmidt process to a linearly dependent list $v_1, \ldots, v_m$, then for some $k \geq 1$ we will have $\tilde{w}_k = 0$, and consequently we cannot continue the process after that. Hence the Gram-Schmidt process can also detect the linear independence of a list of vectors.

Remark. Note that if we change the order of the vectors $v_1, \ldots, v_m$, then their span does not change; however, the vectors $w_1, \ldots, w_m$ will change.

Proof. We show by induction on k that $\tilde{w}_1, \ldots, \tilde{w}_k$ are nonzero, $w_1, \ldots, w_k$ are orthonormal, and $\operatorname{span}(w_1, \ldots, w_k) = \operatorname{span}(v_1, \ldots, v_k)$. For $k = 1$ we have $\tilde{w}_1 = v_1 \neq 0$, since v_1 belongs to a linearly independent set. So w_1 is defined, and we obviously have $\|w_1\| = 1$. Thus w_1 is an orthonormal list. Since w_1, v_1 are scalar multiples of each other, we have $\operatorname{span}(w_1) = \operatorname{span}(v_1)$.

Now suppose the above claims are true for some k. We need to prove them for $k+1$. We know that $\tilde{w}_1, \ldots, \tilde{w}_k$ are nonzero. Suppose to the contrary that $\tilde{w}_{k+1} = 0$. Then we have

$$v_{k+1} = \sum_{j=1}^{k} \langle v_{k+1}, w_j \rangle w_j \in \operatorname{span}(w_1, \ldots, w_k) = \operatorname{span}(v_1, \ldots, v_k),$$

which is a contradiction, because $v_1, \ldots, v_k, v_{k+1}$ are linearly independent. So $\tilde{w}_{k+1} \neq 0$ too.

We know that $w_1, \ldots, w_k$ are orthonormal, and it is obvious that $\|w_{k+1}\| = 1$. So in order to prove that $w_1, \ldots, w_k, w_{k+1}$ are orthonormal, it suffices to show that w_{k+1} is orthogonal to $w_1, \ldots, w_k$. Let $i \leq k$. Then we have

$$\begin{aligned}
\langle \tilde{w}_{k+1}, w_i \rangle &= \langle v_{k+1}, w_i \rangle - \sum_{j \leq k} \big\langle \langle v_{k+1}, w_j \rangle w_j, w_i \big\rangle \\
&= \langle v_{k+1}, w_i \rangle - \sum_{j \leq k} \langle v_{k+1}, w_j \rangle \langle w_j, w_i \rangle \\
&= \langle v_{k+1}, w_i \rangle - \langle v_{k+1}, w_i \rangle \langle w_i, w_i \rangle = \langle v_{k+1}, w_i \rangle - \langle v_{k+1}, w_i \rangle = 0.
\end{aligned}$$

Hence we have $\langle w_{k+1}, w_i \rangle = \frac{1}{\|\tilde{w}_{k+1}\|} \langle \tilde{w}_{k+1}, w_i \rangle = 0$. Thus $w_1, \ldots, w_k, w_{k+1}$ are orthonormal too.

Finally we know that

$$w_1, \ldots, w_k \in \operatorname{span}(w_1, \ldots, w_k) = \operatorname{span}(v_1, \ldots, v_k) \subset \operatorname{span}(v_1, \ldots, v_{k+1}).$$

On the other hand we have

$$\tilde{w}_{k+1} = v_{k+1} - \sum_{j \leq k} \langle v_{k+1}, w_j \rangle w_j \in \operatorname{span}(w_1, \ldots, w_k, v_{k+1}) \subset \operatorname{span}(v_1, \ldots, v_{k+1}).$$

Note that we have applied Proposition 2.22 here. Thus we have

$$w_{k+1} = \frac{1}{\|\tilde{w}_{k+1}\|} \tilde{w}_{k+1} \in \operatorname{span}(v_1, \ldots, v_{k+1}).$$

Hence Proposition 2.22 implies that

$$\operatorname{span}(w_1, \ldots, w_k, w_{k+1}) \subset \operatorname{span}(v_1, \ldots, v_{k+1}).$$

Similarly we have

$$v_1, \ldots, v_k \in \operatorname{span}(v_1, \ldots, v_k) = \operatorname{span}(w_1, \ldots, w_k) \subset \operatorname{span}(w_1, \ldots, w_{k+1}),$$

and

$$\begin{aligned} v_{k+1} &= \tilde{w}_{k+1} + \sum_{j\le k}\langle v_{k+1}, w_j\rangle w_j \\ &= \|\tilde{w}_{k+1}\| w_{k+1} + \sum_{j\le k}\langle v_{k+1}, w_j\rangle w_j \in \operatorname{span}(w_1,\dots,w_{k+1}). \end{aligned}$$

Thus we get $\operatorname{span}(v_1,\dots,v_{k+1}) \subset \operatorname{span}(w_1,\dots,w_{k+1})$, and therefore the two subspaces are equal. ∎

Remark. In the last part of the above proof we could have also argued as follows. We know that $v_1,\dots,v_{k+1}$ are linearly independent, hence their span is a $k+1$ dimensional vector space. Also, $w_1,\dots,w_{k+1}$ are orthonormal, so they are linearly independent. Therefore their span is a $k+1$ dimensional subspace, which is contained in the $k+1$ dimensional space generated by $v_1,\dots,v_{k+1}$. Thus we must have $\operatorname{span}(w_1,\dots,w_{k+1}) = \operatorname{span}(v_1,\dots,v_{k+1})$.

Theorem 5.15. *Every finite dimensional inner product space has an orthonormal basis.*

Proof. If the vector space is the zero vector space, then the empty list is an orthonormal basis for it. Now let $v_1,\dots,v_n$ be a basis for the nonzero inner product space V. Let $w_1,\dots,w_n$ be the orthonormal list constructed from $v_1,\dots,v_n$ using the Gram-Schmidt process. Then we have

$$\operatorname{span}(w_1,\dots,w_n) = \operatorname{span}(v_1,\dots,v_n) = V.$$

Therefore $w_1,\dots,w_n$ is an orthonormal basis for V. ∎

Theorem 5.16. *Every orthonormal list of vectors in a finite dimensional inner product space can be extended to an orthonormal basis.*

Remark. In the following proof we actually provide an algorithm to extend a given orthonormal list to an orthonormal basis.

Proof. Suppose $u_1,\dots,u_k$ is an orthonormal list. Then it is also linearly independent. Thus we can extend it to a basis $u_1,\dots,u_k,v_1,\dots,v_n$. Remember that we have developed several ways to extend a given linearly independent set to a basis. Now by applying the Gram-Schmidt process to the basis $u_1,\dots,u_k,v_1,\dots,v_n$, we construct an orthonormal basis $w_1,\dots,w_k,w_{k+1},\dots,w_{k+n}$. It suffices to show that $w_j = u_j$ for $j \le k$. We prove this by strong induction on j. For $j=1$ we have $w_1 = \frac{1}{\|u_1\|}u_1 = u_1$, since $\|u_1\| = 1$. Suppose the claim is true for $j = 1,2,\dots,l$. Then we have

$$\tilde{w}_{l+1} = u_{l+1} - \sum_{j\le l}\langle u_{l+1}, w_j\rangle w_j = u_{l+1} - \sum_{j\le l}\langle u_{l+1}, u_j\rangle u_j = u_{l+1},$$

since $\langle u_{l+1}, u_j \rangle = 0$ for $j \leq l$. Hence we have

$$w_{l+1} = \frac{1}{\|\tilde{w}_{l+1}\|}\tilde{w}_{l+1} = \frac{1}{\|u_{l+1}\|}u_{l+1} = u_{l+1},$$

because $\|u_{l+1}\| = 1$. Therefore we have constructed an orthonormal basis $u_1, \ldots, u_k$, $w_{k+1}, \ldots, w_{k+n}$ as desired. ■

Remark. Suppose $v_1, \ldots, v_m \in V$ are orthonormal. Then as shown in the above proof, if we apply the Gram-Schmidt process to the list $v_1, \ldots, v_m$, the resulting list is $v_1, \ldots, v_m$ itself.

5.3 Orthogonal Projections

Definition 5.17. Suppose W is a subspace of V. The **orthogonal complement** of W is

$$W^{\perp} := \{v \in V : \langle v, w \rangle = 0 \text{ for every } w \in W\}.$$

***Example* 5.18.** We have $V^{\perp} = \{0\}$. Because if a vector v belongs to $V^{\perp}$ then it is orthogonal to every vector in V. In particular v is orthogonal to itself, so we must have $v = 0$. On the other hand it is obvious that 0 is orthogonal to all vectors in V. Similarly we have $\{0\}^{\perp} = V$. Because every vector is orthogonal to the zero vector.

***Example* 5.19.** Suppose U, W are subspaces of V, and $U \subset W$. Then we have $W^{\perp} \subset U^{\perp}$. Because if $v \in W^{\perp}$ then v is orthogonal to every vector in W. In particular, v is orthogonal to every vector in U. Hence $v \in U^{\perp}$.

Remark. A simple fact that is useful when we deal with orthogonal complements is that if $v \in V$ is orthogonal to each $w_1, \ldots, w_m \in V$, then v is orthogonal to every vector in $\text{span}(w_1, \ldots, w_m)$. Because if $w \in \text{span}(w_1, \ldots, w_m)$ then there are $a_1, \ldots, a_m \in F$ such that $w = a_1 w_1 + \cdots + a_m w_m$. Hence we have

$$\begin{aligned}\langle v, w \rangle &= \langle v, a_1 w_1 + \cdots + a_m w_m \rangle \\ &= \bar{a}_1 \langle v, w_1 \rangle + \cdots + \bar{a}_m \langle v, w_m \rangle = \bar{a}_1 0 + \cdots + \bar{a}_m 0 = 0.\end{aligned}$$

Theorem 5.20. *Suppose W is a subspace of V. Then $W^{\perp}$ is also a subspace of V, and we have*

$$W \cap W^{\perp} = \{0\}.$$

Remark. This theorem implies that $W, W^{\perp}$ are independent subspaces.

Proof. First note that $0 \in W^\perp$, since 0 is orthogonal to every vector in W. Now let $u, v \in W^\perp$ and $a \in F$. Then for every $w \in W$ we have

$$\langle u + av, w\rangle = \langle u, w\rangle + a\langle v, w\rangle = 0 + a0 = 0.$$

Hence $u + av \in W^\perp$. Therefore $W^\perp$ is a subspace. Next suppose $v \in W \cap W^\perp$. Then $v \in W^\perp$, so it is orthogonal to every vector in W. In particular v is orthogonal to itself, i.e. $\langle v, v\rangle = 0$. Thus we must have $v = 0$. ■

Theorem 5.21. *Suppose W is a finite dimensional subspace of V. Then we have*
(i) $W \oplus W^\perp = V$.
(ii) $(W^\perp)^\perp = W$.

Proof. **(i)** Let $w_1, \ldots, w_m$ be an orthonormal basis for W. Then for $v \in V$ set

$$w := \langle v, w_1\rangle w_1 + \cdots + \langle v, w_m\rangle w_m, \qquad u := v - w.$$

It is obvious that $w \in W$. We claim that $u \in W^\perp$. For every $j \le m$ we have

$$\begin{aligned}\langle w, w_j\rangle &= \big\langle \langle v, w_1\rangle w_1 + \cdots + \langle v, w_m\rangle w_m, w_j\big\rangle \\ &= \langle v, w_1\rangle\langle w_1, w_j\rangle + \cdots + \langle v, w_m\rangle\langle w_m, w_j\rangle = \langle v, w_j\rangle.\end{aligned}$$

Hence $\langle u, w_j\rangle = \langle v, w_j\rangle - \langle w, w_j\rangle = 0$. Thus u is orthogonal to every vector in $\operatorname{span}(w_1, \ldots, w_m) = W$. So $u \in W^\perp$. Therefore we have

$$v = w + u \in W + W^\perp = W \oplus W^\perp.$$

Note that $W, W^\perp$ are independent subspaces, so their sum is a direct sum. Hence we have $V \subset W \oplus W^\perp$. Since the reverse inclusion is trivial, we get the desired.

(ii) Let $w \in W$. Then we know that every vector in $W^\perp$ is orthogonal to w. Hence w is also orthogonal to every vector in $W^\perp$. Thus $w \in (W^\perp)^\perp$. Therefore we have

$$W \subset (W^\perp)^\perp.$$

Note that we did not need the finite dimensionality of W for the above relation to hold. Now let $v \in (W^\perp)^\perp \subset V$. In order to prove $(W^\perp)^\perp \subset W$, it suffices to show that $v \in W$. We know that $V = W \oplus W^\perp$. Thus $v = w + u$, where $w \in W$ and $u \in W^\perp$. But w, u are orthogonal, so by the Pythagorean theorem we have

$$\|v\|^2 = \|w\|^2 + \|u\|^2. \qquad (*)$$

On the other hand, v, u are also orthogonal. Hence we have

$$0 = \langle v, u\rangle = \langle v, v - w\rangle = \langle v, v\rangle - \langle v, w\rangle.$$

Therefore $\|v\|^2 = \langle v, v\rangle = \langle v, w\rangle \leq \|v\|\|w\|$. Now if $v = 0$ then we trivially have $v \in W$ as desired. Otherwise we can cancel $\|v\|$ from both sides of the above inequality to obtain $\|v\| \leq \|w\|$. Thus the equation $(*)$ implies

$$\|w\|^2 + \|u\|^2 = \|v\|^2 \leq \|w\|^2.$$

Hence we must have $\|u\| = 0$. Therefore $u = 0$, and we get $v = w \in W$ as desired. ■

Remark. The above theorem is not true in general, when W is infinite dimensional.

Remark. Note that in the above proof, in order to prove that $(W^\perp)^\perp \subset W$, we did not use the finite dimensionality of W directly. We only used the fact that $V = W \oplus W^\perp$. So, as a result, we can conclude that if $V = W \oplus W^\perp$ then

$$W^\perp \oplus (W^\perp)^\perp = W^\perp \oplus W = W \oplus W^\perp = V.$$

Theorem 5.22. *Suppose V is finite dimensional, and W is a subspace of V.*

(i) *We have*
$$\dim W + \dim W^\perp = \dim V.$$

(ii) *Let $w_1, \ldots, w_m$ be an orthonormal basis for W, and suppose $v_{m+1}, \ldots, v_n \in V$ are such that $w_1, \ldots, w_m, v_{m+1}, \ldots, v_n$ is an orthonormal basis for V. Then $v_{m+1}, \ldots, v_n$ is an orthonormal basis for $W^\perp$.*

(iii) *Let $w_1, \ldots, w_m$ be an orthonormal basis for W, and let $v_{m+1}, \ldots, v_n$ be an orthonormal basis for $W^\perp$. Then $w_1, \ldots, w_m, v_{m+1}, \ldots, v_n$ is an orthonormal basis for V.*

Remark. The orthonormal basis of W is an orthonormal list in V. In the proof of Theorem 5.16 we have provided an algorithm to extend a given orthonormal list to an orthonormal basis for V. So if we combine that algorithm with the part (ii) of this theorem, we have an algorithm to find an orthonormal basis for $W^\perp$ by using an orthonormal basis for W.

Proof. **(i)** Since W is also finite dimensional we have $V = W \oplus W^\perp$. Thus by Theorem 2.55 we get

$$\dim V = \dim(W \oplus W^\perp) = \dim W + \dim W^\perp.$$

(ii) For every i, j we have $\langle v_i, w_j\rangle = 0$, so each v_i is orthogonal to every vector in $\operatorname{span}(w_1, \ldots, w_m) = W$. Hence $v_{m+1}, \ldots, v_n \in W^\perp$. But we know that

$$\dim W^\perp = \dim V - \dim W = n - m.$$

On the other hand, $v_{m+1}, \ldots, v_n$ is an orthonormal list, so it is linearly independent. Now, $v_{m+1}, \ldots, v_n$ is a linearly independent set of vectors in $W^\perp$ that has the same number of vectors as $\dim W$. Therefore $v_{m+1}, \ldots, v_n$ is a basis for $W^\perp$.

(iii) We know that $w_1, \ldots, w_m, v_{m+1}, \ldots, v_n$ is a basis for V, since $V = W \oplus W^\perp$. So we only need to show that this set is orthonormal. We know that every vector in this list has norm one. We also know that when $i \neq j$, w_i, w_j are orthogonal; and when $k \neq l$, v_k, v_l are orthogonal. Finally note that for every i, k, w_i, v_k are orthogonal too, because one of them is in W and the other one is in $W^\perp$. ■

Definition 5.23. Suppose W is a subspace of V such that $V = W \oplus W^\perp$. Then for every $v \in V$ there are unique vectors $w \in W$ and $u \in W^\perp$ such that $v = w + u$. The function

$$\begin{aligned} P : V &\longrightarrow V \\ v &\mapsto w \end{aligned}$$

is called the **orthogonal projection** on W. We also say that w is the orthogonal projection of v on W.

Remark. Note that P is a well defined function, since w is uniquely determined by v.

Remark. If W is an arbitrary subspace of V, then the orthogonal projection on W does not necessarily exist. But due to Theorem 5.21, when W is finite dimensional, the orthogonal projection on W exists. Also, due to the argument in a remark after the aforementioned theorem, if the orthogonal projection on W exists then the orthogonal projection on $W^\perp$ exists too.

Theorem 5.24. *Suppose W is a subspace of V, and P is the orthogonal projection on W. Then we have*

(i) *P is a linear operator, i.e. $P \in \mathcal{L}(V)$.*
(ii) *$P|_W = I_W$.*
(iii) *$P^2 = P$.*
(iv) *$P(V) = W$, and $\operatorname{null} P = W^\perp$.*
(v) *$v - Pv \in W^\perp$ for every $v \in V$, so in particular $\langle v - Pv, Pv \rangle = 0$.*
(vi) *$\|Pv\| \leq \|v\|$ for every $v \in V$.*
(vii) *$I_V - P$ is the orthogonal projection on $W^\perp$.*

Proof. Since P exists, we must have $V = W \oplus W^\perp$. Let $v \in V$. Then we have $v = w + u$, for some uniquely determined $w \in W$ and $u \in W^\perp$. Hence $Pv = w$ by definition.

(i) Let $\tilde{v} \in V$ and $a \in F$. Then we have $\tilde{v} = \tilde{w} + \tilde{u}$, for some uniquely determined $\tilde{w} \in W$ and $\tilde{u} \in W^\perp$. Hence $P\tilde{v} = \tilde{w}$. Therefore we have

$$v + a\tilde{v} = w + u + a\tilde{w} + a\tilde{u} = w + a\tilde{w} + u + a\tilde{u}.$$

But $w + a\tilde{w} \in W$ and $u + a\tilde{u} \in W^\perp$. Thus we have

$$P(v + a\tilde{v}) = w + a\tilde{w} = Pv + aP\tilde{v}.$$

Therefore P is linear.

(ii) Let $\tilde{w} \in W$. Then we have $\tilde{w} = \tilde{w} + 0$, where $\tilde{w} \in W$ and $0 \in W^\perp$. Hence by definition of P we have $P\tilde{w} = \tilde{w} = I_W\tilde{w}$.

(iii) We have $Pv = w \in W$, so $P^2v = P(Pv) = I_W(Pv) = Pv$.

(iv) Let $\tilde{w} \in W$. Then we have $\tilde{w} = P\tilde{w} \in P(V)$. So $W \subset P(V)$. On the other hand, for every $v \in V$ we have $Pv \in W$. Hence $P(V) = W$. Now let $\tilde{u} \in W^\perp$. Then $\tilde{u} = 0 + \tilde{u}$, where $0 \in W$ and $\tilde{u} \in W^\perp$. Thus $P\tilde{u} = 0$. Therefore $W^\perp \subset \operatorname{null} P$. Conversely suppose $v \in \operatorname{null} P$. Then we have $w = Pv = 0$. Hence $v = w + u = 0 + u = u \in W^\perp$. Thus $W^\perp = \operatorname{null} P$.

(v) We have $v - Pv = v - w = u \in W^\perp$. We also know that $Pv = w \in W$. Therefore we must have $\langle v - Pv, Pv\rangle = 0$.

(vi) We know that $\langle v, Pv\rangle - \langle Pv, Pv\rangle = \langle v - Pv, Pv\rangle = 0$. Hence

$$\|Pv\|^2 = \langle Pv, Pv\rangle = \langle v, Pv\rangle \leq \|v\|\|Pv\|.$$

Now if $Pv = 0$ then we trivially have $\|Pv\| = 0 \leq \|v\|$ as desired. Otherwise we can cancel $\|Pv\|$ from both sides of the above inequality to obtain $\|Pv\| \leq \|v\|$.

(vii) As we have explained in the remark before this theorem, since the orthogonal projection on W exists, the orthogonal projection on $W^\perp$ exists too. Now we have $v = u + w$, where $u \in W^\perp$ and $w \in W \subset (W^\perp)^\perp$. Hence u is the orthogonal projection of v on $W^\perp$. We also have

$$u = v - w = I_V v - Pv = (I_V - P)v.$$

Therefore $I_V - P$ is the orthogonal projection on $W^\perp$. ■

Theorem 5.25. *Suppose W is a finite dimensional subspace of V, and $w_1, \ldots, w_m$ is an orthonormal basis for W. Let P be the orthogonal projection on W. Then for every $v \in V$ we have*

$$Pv = \langle v, w_1\rangle w_1 + \cdots + \langle v, w_m\rangle w_m.$$

Proof. Set

$$w := \langle v, w_1\rangle w_1 + \cdots + \langle v, w_m\rangle w_m, \qquad u := v - w.$$

Then $v = w + u$. It is obvious that $w \in W$. Also, we have shown in the proof of Theorem 5.21 that $u \in W^\perp$. Hence $Pv = w$. ■

***Example* 5.26.** Let $w \in V - \{0\}$, and consider the one-dimensional subspace $\text{span}(w)$. Then $w_1 := \frac{1}{\|w\|} w$ is an orthonormal basis for $\text{span}(w)$. Hence the above theorem implies that the orthogonal projection of $v \in V$ on $\text{span}(w)$ is

$$Pv = \langle v, w_1 \rangle w_1 = \langle v, \frac{w}{\|w\|} \rangle \frac{w}{\|w\|} = \frac{\langle v, w \rangle}{\|w\|^2} w.$$

Theorem 5.27. *Suppose W is a subspace of V, and P is the orthogonal projection on W. Let $v \in V$. Then for every $w \in W$ we have*

$$\|v - Pv\| \leq \|v - w\|.$$

Furthermore, if $\|v - Pv\| = \|v - w\|$ for some $w \in W$, then $w = Pv$.

Remark. In other words, Pv is the unique vector in W that has the least distance to v. This result provides a geometric characterization for the orthogonal projection P.

Proof. We know that there is $u \in W^\perp$ such that $v = Pv + u$. Then for every $w \in W$ we have $v - w = Pv - w + u$. But $Pv - w \in W$, so u is orthogonal to $Pv - w$. Hence by the Pythagorean theorem we have

$$\|v - w\|^2 = \|Pv - w\|^2 + \|u\|^2 \geq \|u\|^2 = \|v - Pv\|^2, \qquad (*)$$

since $u = v - Pv$. Thus $\|v - w\| \geq \|v - Pv\|$. Now if $\|v - w\| = \|v - Pv\| = \|u\|$ then the formula $(*)$ implies that $\|Pv - w\| = 0$. Therefore we get $w = Pv$. ■

Remark. Although the above theorem seems to be a simple geometric result, it is very powerful, and it has many applications. The reason is that there are many problems that can be reformulated as the problem of minimizing the distance to a given subspace, especially a subspace of an infinite dimensional vector space. We should mention that when the subspace is infinite dimensional, the usual compactness techniques of analysis do not work, and we need new tools to establish the existence of the minimizer. But the above theorem tells us that if we find the orthogonal projection operator, we have essentially solved the minimization problem too.

Remark. If an orthonormal basis for a subspace W is given to us, then we can easily compute the orthogonal projection on W by using Theorem 5.25. But if instead we just have a basis for W, then we need to construct an orthonormal basis first. The next theorem allows us to compute the orthogonal projection on W without constructing an orthonormal basis for it. We will assume that $W \subset F^n$, since this is the most important case in applications.

Theorem 5.28. *Suppose W is a subspace of F^n, and P is the orthogonal projection on W. Let $w_1, \dots, w_m \in F^n$ be a basis for W, and let $A \in F^{n\times m}$ be the matrix whose j-th column is w_j. Then $A^*A \in F^{m\times m}$ is invertible, and for any $y \in F^n$ we have*

$$Py = A(A^*A)^{-1}A^*y.$$

Remark. Note that A^* is the conjugate transpose of the matrix A, as defined in Definition 1.26. Also note that in general $(A^*A)^{-1} \neq A^{-1}(A^*)^{-1}$, since when $m < n$, A, A^* are not square matrices, and therefore they cannot have an inverse.

Proof. First let us show that A^*A is invertible. Suppose $A^*Ax = 0$ for some $x \in F^m$. Then we have

$$\begin{aligned} 0 = \langle 0, x\rangle = \langle A^*Ax, x\rangle = x^*(A^*Ax) = (x^*A^*)(Ax) \\ = (Ax)^*(Ax) = \langle Ax, Ax\rangle = \|Ax\|^2. \end{aligned}$$

Thus $Ax = 0$. Hence we have

$$0 = Ax = \sum_{j\le m} x_j A_{\cdot,j} = \sum_{j\le m} x_j w_j.$$

Therefore $x_j = 0$ for every j, since $w_1, \dots, w_m$ are linearly independent. Hence the linear system $A^*Ax = 0$ has only one solution $x = 0$. Thus by Theorem 3.49 the matrix A^*A is invertible.

Now let $x := (A^*A)^{-1}A^*y$. Then $Ax = \sum_{j\le m} x_j A_{\cdot,j} = \sum_{j\le m} x_j w_j \in W$. Set $z := y - Ax$. We claim that $z \in W^\perp$. Note that $A^*Ax = A^*y$. Also note that for every $i \le m$ we have $w_i^* = (A_{\cdot,i})^* = A^*_{i,\cdot}$. Hence we have

$$\begin{aligned} \langle Ax, w_i\rangle = w_i^*(Ax) = A^*_{i,\cdot}(Ax) = \big(A^*(Ax)\big)_{i,\cdot} \\ = \big((A^*A)x\big)_{i,\cdot} = \big(A^*y\big)_{i,\cdot} = A^*_{i,\cdot}y = w_i^*y = \langle y, w_i\rangle. \end{aligned}$$

Therefore we have $\langle z, w_i\rangle = \langle y, w_i\rangle - \langle Ax, w_i\rangle = 0$. Thus z is orthogonal to every vector in $\operatorname{span}(w_1, \dots, w_m) = W$. Hence $z \in W^\perp$ as desired. Then we have $y = Ax + z$, where $Ax \in W$ and $z \in W^\perp$. Therefore Ax is the orthogonal projection of y on W, i.e. $Py = Ax = A(A^*A)^{-1}A^*y$. ∎

Remark. In the above theorem, it is easy to see that $W = \{Ax : x \in F^m\}$. Let

$$x_0 := (A^*A)^{-1}A^*y.$$

We know that $Ax_0 = Py$ is the closest vector in W to y. So for every $x \in F^m$ we have

$$\|Ax_0 - y\| \le \|Ax - y\|.$$

Let us present an application of the above observation. Suppose we have a collection of data points $(a_i, b_i) \in \mathbb{R}^2$ for $i = 1, \ldots, n$, and we guess that there is a linear relation between b_i and a_i, i.e. there are scalars β_0, β_1 such that $b_i = \beta_0 + \beta_1 a_i$. But in practice we cannot hope that the linear relation between b_i and a_i is exact, since there may be measurement errors, and/or other noises that we cannot measure at all. Hence instead of looking for an exact linear relation, we will look for a linear relation with the least error, i.e. we want to find β_0, β_1 such that

$$\sum_{i \leq n} |\beta_0 + \beta_1 a_i - b_i|^2$$

is as small as possible. This is called the method of **least squares**. We can solve this problem by the tools that we have developed. Let A be the $n \times 2$ matrix whose i-th row is $[1, a_i]$, and let $y = [b_1, \ldots, b_n]^\intercal$. Then we want to find $\beta = [\beta_0, \beta_1]^\intercal$ such that for every $x = [x_1, x_2]^\intercal$ we have

$$\|A\beta - y\|^2 = \sum_{i \leq n} |\beta_0 + \beta_1 a_i - b_i|^2 \leq \sum_{i \leq n} |x_1 + x_2 a_i - b_i|^2 = \|Ax - y\|^2.$$

Now the above remark tells us that if the two columns of A are linearly independent, then

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta = (A^* A)^{-1} A^* y$$

has the desired property.

Note that the linear independence of the columns of A means that $[a_1, \ldots, a_n]^\intercal$ is not a multiple of $[1, 1, \ldots, 1]^\intercal$, which is equivalent to having $a_i \neq a_j$ for some i, j. Now suppose this is the case, and let us compute a closed formula for β_0, β_1. We have

$$A^* A = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{bmatrix} = \begin{bmatrix} n & a_1 + \cdots + a_n \\ a_1 + \cdots + a_n & a_1^2 + \cdots + a_n^2 \end{bmatrix},$$

$$A^* y = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} b_1 + \cdots + b_n \\ a_1 b_1 + \cdots + a_n b_n \end{bmatrix}.$$

Let $m_a := \frac{1}{n} \sum_{i \leq n} a_i$ and $m_b := \frac{1}{n} \sum_{i \leq n} b_i$ be the *mean* of a_i's and b_i's respectively.

Let $\sigma_a^2 := \frac{1}{n}\sum_{i\leq n}(a_i - \bar{a})^2$ be the *variance* of a_i's. Then we have

$$\begin{aligned} n\sigma_a^2 = \sum_{i\leq n}(a_i - m_a)^2 &= \sum_{i\leq n} a_i^2 - 2\sum_{i\leq n} a_i m_a + nm_a^2 \\ &= \sum_{i\leq n} a_i^2 - 2m_a \sum_{i\leq n} a_i + nm_a^2 \\ &= \sum_{i\leq n} a_i^2 - 2m_a(nm_a) + nm_a^2 = \sum_{i\leq n} a_i^2 - nm_a^2. \end{aligned}$$

Therefore $\sum_{i\leq n} a_i^2 = n\sigma_a^2 + nm_a^2$. Hence we have

$$A^*A = n\begin{bmatrix} 1 & m_a \\ m_a & \sigma_a^2 + m_a^2 \end{bmatrix} \implies (A^*A)^{-1} = \frac{1}{n\sigma_a^2}\begin{bmatrix} \sigma_a^2 + m_a^2 & -m_a \\ -m_a & 1 \end{bmatrix}.$$

Now let $\sigma_{ab}^2 := \frac{1}{n}\sum_{i\leq n}(a_i - m_a)(b_i - m_b)$ be the *covariance* of a_i's and b_i's. We have

$$\begin{aligned} n\sigma_{ab}^2 &= \sum_{i\leq n}(a_i - m_a)(b_i - m_b) \\ &= \sum_{i\leq n} a_i b_i - m_a \sum_{i\leq n} b_i - m_b \sum_{i\leq n} a_i + nm_a m_b \\ &= \sum_{i\leq n} a_i b_i - m_a(nm_b) - m_b(nm_a) + nm_a m_b = \sum_{i\leq n} a_i b_i - nm_a m_b. \end{aligned}$$

To simplify the notation let $m_{ab} := \frac{1}{n}\sum_{i\leq n} a_i b_i$; so we have $\sigma_{ab}^2 = m_{ab} - m_a m_b$. Then we get

$$\begin{aligned} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (A^*A)^{-1}A^*y &= \frac{1}{n\sigma_a^2}\begin{bmatrix} \sigma_a^2 + m_a^2 & -m_a \\ -m_a & 1 \end{bmatrix}\begin{bmatrix} nm_b \\ nm_{ab} \end{bmatrix} \\ &= \frac{1}{\sigma_a^2}\begin{bmatrix} \sigma_a^2 m_b + m_a^2 m_b - m_a m_{ab} \\ -m_a m_b + m_{ab} \end{bmatrix} = \frac{1}{\sigma_a^2}\begin{bmatrix} \sigma_a^2 m_b - m_a \sigma_{ab}^2 \\ \sigma_{ab}^2 \end{bmatrix}. \end{aligned}$$

So we have

$$\beta_1 = \frac{\sigma_{ab}^2}{\sigma_a^2}, \qquad \beta_0 = -m_a\frac{\sigma_{ab}^2}{\sigma_a^2} + m_b = -m_a\beta_1 + m_b.$$

If we employ $\sigma_{ab}^2 = m_{ab} - m_a m_b$ and $\sigma_a^2 + m_a^2 = \frac{1}{n}\sum a_i^2$, we can also write

$$\beta_1 = \frac{n\sum a_i b_i - \big(\sum a_i\big)\big(\sum b_i\big)}{n\big(\sum a_i^2\big) - \big(\sum a_i\big)^2}, \qquad \beta_0 = \frac{\big(\sum b_i\big)\big(\sum a_i^2\big) - \big(\sum a_i\big)\big(\sum a_i b_i\big)}{n\big(\sum a_i^2\big) - \big(\sum a_i\big)^2}.$$

Finally let us mention that the line $t \mapsto \beta_0 + \beta_1 t$ in $\mathbb{R}^2$, which is called the **regression line**, satisfies $\beta_0 + \beta_1 m_a = m_b$. So the regression line passes through the mean of the data points. ■

Chapter 6

Operators on Inner Product Spaces

6.1 The Adjoint of an Operator

Notation. In this chapter, we assume that F is either $\mathbb{R}$ or $\mathbb{C}$, and V is a nonzero inner product space over F with the inner product $\langle\ ,\ \rangle$. We also assume that $T \in \mathcal{L}(V)$ unless otherwise specified.

Theorem 6.1. *Suppose that V is finite dimensional, and $f \in \mathcal{L}(V, F)$ is a linear functional on V. Then there exists a unique vector $u \in V$ such that*

$$f(v) = \langle v, u \rangle$$

for every $v \in V$.

Proof. Let $v_1, \ldots, v_n$ be an orthonormal basis for V. Let $b_j := \overline{f(v_j)}$ for every $j \leq n$, and let $u := \sum_{j \leq n} b_j v_j$. For every $v \in V$ there are $a_j \in F$ such that $v = \sum_{j \leq n} a_j v_j$. Hence by Theorem 5.14 we have

$$\langle v, u \rangle = \sum_{j \leq n} a_j \bar{b}_j = \sum_{j \leq n} a_j f(v_j) = f\Big(\sum_{j \leq n} a_j v_j \Big) = f(v).$$

Thus u has the desired property. Now if $w \in V$ also satisfies $\langle v, w \rangle = f(v)$ for every $v \in V$, then we have

$$\langle v, u - w \rangle = \langle v, u \rangle - \langle v, w \rangle = f(v) - f(v) = 0,$$

for every $v \in V$. Therefore we must have $u - w = 0$. So u is unique. ■

Remark. Let $A \in F^{n\times m}$, and let $A^* \in F^{m\times n}$ be the conjugate transpose of A. Then for $x \in F^m$ and $y \in F^n$ we have

$$\langle Ax, y\rangle = y^*(Ax) = (y^*A)x = (y^*(A^*)^*)x = (A^*y)^*x = \langle x, A^*y\rangle.$$

This equality motivates the next definition.

Definition 6.2. Suppose V, W are inner product spaces over F, and $T \in \mathcal{L}(V, W)$. A function $T^* : W \to V$ is called an **adjoint** of T if

$$\langle Tv, w\rangle = \langle v, T^*w\rangle$$

for every $v \in V$ and $w \in W$.

Remark. Note that in the above equality, Tv, w are multiplied using the inner product of W, and v, T^*w are multiplied using the inner product of V.

Theorem 6.3. *Suppose V, W are inner product spaces over F, and $T \in \mathcal{L}(V, W)$ has an adjoint T^*. Then the function T^* is uniquely determined by T. Furthermore, T^* is linear, i.e. $T^* \in \mathcal{L}(W, V)$.*

Proof. Suppose that $S : W \to V$ is also a function such that $\langle Tv, w\rangle = \langle v, Sw\rangle$ for every $v \in V$ and $w \in W$. Then we have

$$\langle v, T^*w - Sw\rangle = \langle v, T^*w\rangle - \langle v, Sw\rangle = \langle Tv, w\rangle - \langle Tv, w\rangle = 0.$$

Thus $T^*w - Sw$ is orthogonal to every $v \in V$, so it must be zero, i.e. $T^*w = Sw$. But w is an arbitrary element of W, hence we get $T^* = S$. Therefore T^* is uniquely determined by T.

Now let $u, w \in W$ and $a \in F$. Then for every $v \in V$ we have

$$\begin{aligned}\langle v, T^*(w + au)\rangle &= \langle Tv, w + au\rangle = \langle Tv, w\rangle + \bar{a}\langle Tv, u\rangle \\ &= \langle v, T^*w\rangle + \bar{a}\langle v, T^*u\rangle = \langle v, T^*w + aT^*u\rangle.\end{aligned}$$

Hence $T^*(w + au) - T^*w - aT^*u$ is orthogonal to every $v \in V$, so we must have $T^*(w + au) = T^*w + aT^*u$. Therefore T^* is linear. ∎

Theorem 6.4. *Suppose V, W are inner product spaces over F, and V is finite dimensional. Then every $T \in \mathcal{L}(V, W)$ has adjoint.*

Proof. Let $w \in W$, and define $f(v) := \langle Tv, w\rangle$ for $v \in V$. It is easy to see that f is linear. Let $v, \tilde{v} \in V$ and $a \in F$. Then we have

$$\begin{aligned}f(v + a\tilde{v}) &= \langle T(v + a\tilde{v}), w\rangle = \langle Tv + aT\tilde{v}, w\rangle \\ &= \langle Tv, w\rangle + a\langle T\tilde{v}, w\rangle = f(v) + af(\tilde{v}).\end{aligned}$$

Thus f is a linear functional on V. Hence by Theorem 6.1 there is a unique vector $u \in V$ such that for every $v \in V$ we have

$$\langle Tv, w\rangle = f(v) = \langle v, u\rangle.$$

Now we can define $T^*w := u$. Note that T^* is a well defined function, since for each w, the vector u is uniquely determined. ■

Remark. Suppose that in the above theorem, W is infinite dimensional. Then we know that $T(V)$ is still finite dimensional. Also for $w \in (T(V))^\perp$ we have $\langle Tv, w\rangle = 0$ for every $v \in V$. Hence we must have $T^*w = 0$. Therefore if we determine the value of T^* on the finite dimensional subspace $T(V)$, then we can easily define T^* on all of W. Because we know that $W = T(V) \oplus (T(V))^\perp$, so if $w \in W$ has the decomposition $w = w_1 + w_2$ where $w_1 \in T(V)$ and $w_2 \in (T(V))^\perp$, then we can define $T^*w := T^*w_1$. Note that we do not need this reasoning in the above proof. We only presented it here to emphasize that the finite dimensionality of W is not needed in the above theorem.

Remark. We are mainly interested in the case of operators on an inner product space, i.e. when $W = V$. So we will only consider this case from now on, although the results of this section are true in the more general setting.

Theorem 6.5. *Suppose V is finite dimensional, and $\mathcal{B} = \{v_1, \ldots, v_n\}$ is an orthonormal basis for V. Let $T \in \mathcal{L}(V)$. Then the jk-th entry of the matrix $[T]_\mathcal{B}$ is*

$$([T]_\mathcal{B})_{jk} = \langle Tv_k, v_j\rangle.$$

Remark. Note that this theorem, and the next one, are not true when the basis $\mathcal{B}$ is not orthonormal.

Proof. Note that the k-th column of the matrix $[T]_\mathcal{B}$ is $[Tv_k]_\mathcal{B}$. So the jk-th entry of $[T]_\mathcal{B}$ is the j-th component of $[Tv_k]_\mathcal{B}$. Now suppose $Tv_k = a_1v_1 + \cdots + a_nv_n$ for some $a_i \in F$. Then the j-th component of $[Tv_k]_\mathcal{B}$ is the coefficient of v_j in the above expansion, i.e. a_j. But Theorem 5.14 implies that $a_j = \langle Tv_k, v_j\rangle$. Hence we have $([T]_\mathcal{B})_{jk} = \langle Tv_k, v_j\rangle$ as desired. ■

Theorem 6.6. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Let $T \in \mathcal{L}(V)$. Then*

$$[T^*]_\mathcal{B} = [T]_\mathcal{B}^*.$$

Remark. Note that $[T]_\mathcal{B}^*$ is the conjugate transpose of the matrix $[T]_\mathcal{B}$, as defined in Definition 1.26.

Proof. Suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$. Let $A := [T]_{\mathcal{B}}$ and $B := [T^*]_{\mathcal{B}}$. Then by applying the previous theorem to T, T^* we get

$$B_{jk} = \langle T^* v_k, v_j \rangle = \overline{\langle v_j, T^* v_k \rangle} = \overline{\langle T v_j, v_k \rangle} = \overline{A_{kj}},$$

for every $j, k \leq n$. Therefore we have $B = A^*$ as desired. ■

Theorem 6.7. *Suppose $S, T \in \mathcal{L}(V)$ have adjoints. Let $a \in F$. Then we have*

(i) *$S + T$ has adjoint, and $(S + T)^* = S^* + T^*$.*
(ii) *aT has adjoint, and $(aT)^* = \bar{a}T^*$.*
(iii) *ST has adjoint, and $(ST)^* = T^* S^*$.*
(iv) *I_V has adjoint, and $I_V^* = I_V$.*
(v) *T^* has adjoint, and $(T^*)^* = T$.*

Proof. Note that in all of the following parts, we are tacitly using the fact that the adjoint of an operator is unique, when it exists.

(i) We have

$$\begin{aligned}\langle (S+T)v, w \rangle &= \langle Sv + Tv, w \rangle = \langle Sv, w \rangle + \langle Tv, w \rangle \\ &= \langle v, S^* w \rangle + \langle v, T^* w \rangle = \langle v, S^* w + T^* w \rangle = \langle v, (S^* + T^*) w \rangle,\end{aligned}$$

for every $v, w \in V$. Therefore by definition $S+T$ has adjoint, and we have $(S+T)^* = S^* + T^*$.

(ii) We have

$$\begin{aligned}\langle (aT)v, w \rangle &= \langle aTv, w \rangle = a\langle Tv, w \rangle \\ &= a\langle v, T^* w \rangle = \langle v, \bar{a} T^* w \rangle = \langle v, (\bar{a} T^*) w \rangle,\end{aligned}$$

for every $v, w \in V$. Hence by definition aT has adjoint, and we have $(aT)^* = \bar{a}T^*$.

(iii) We have

$$\begin{aligned}\langle (ST)v, w \rangle &= \langle S(Tv), w \rangle = \langle Tv, S^* w \rangle \\ &= \langle v, T^*(S^* w) \rangle = \langle v, (T^* S^*) w \rangle,\end{aligned}$$

for every $v, w \in V$. Hence by definition ST has adjoint, and we have $(ST)^* = T^* S^*$.

(iv) We have $\langle I_V v, w \rangle = \langle v, w \rangle = \langle v, I_V w \rangle$ for every $v, w \in V$. Therefore by definition I_V has adjoint, and we have $I_V^* = I_V$.

(v) We have

$$\langle T^* v, w \rangle = \overline{\langle w, T^* v \rangle} = \overline{\langle T w, v \rangle} = \langle v, T w \rangle,$$

for every $v, w \in V$. Hence by definition T^* has adjoint, and we have $(T^*)^* = T$. ■

Remark. If T has adjoint then by definition we have $\langle Tv, w\rangle = \langle v, T^*w\rangle$ for every $v, w \in V$. Now the above theorem implies that we also have

$$\langle v, Tw\rangle = \langle v, (T^*)^*w\rangle = \langle T^*v, w\rangle.$$

Theorem 6.8. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is invertible. Then T^* is invertible, and we have*

$$(T^*)^{-1} = (T^{-1})^*.$$

Proof. First note that T and T^{-1} have adjoints, since V is finite dimensional. Then we have

$$(T^{-1})^*T^* = (TT^{-1})^* = I_V^* = I_V,$$
$$T^*(T^{-1})^* = (T^{-1}T)^* = I_V^* = I_V.$$

Thus T^* is invertible, and its inverse is $(T^{-1})^*$. ■

Theorem 6.9. *Suppose $T \in \mathcal{L}(V)$ has adjoint. Then we have*

(i) $\operatorname{null} T^* = (T(V))^\perp$.

(ii) *If V is finite dimensional, then $T^*(V) = (\operatorname{null} T)^\perp$.*

Remark. We can obviously change the role of T, T^* in the above relations, since $(T^*)^* = T$. Hence we can also write

(iii) $\operatorname{null} T = \operatorname{null} (T^*)^* = (T^*(V))^\perp$.

(iv) If V is finite dimensional, then $T(V) = (T^*)^*(V) = (\operatorname{null} T^*)^\perp$.

Proof. **(i)** Let $v \in \operatorname{null} T^*$, and $w \in T(V)$. Then there is $u \in V$ such that $w = Tu$. Thus we have

$$\langle v, w\rangle = \langle v, Tu\rangle = \langle T^*v, u\rangle = \langle 0, u\rangle = 0.$$

Hence $v \in (T(V))^\perp$. On the other hand if $v \in (T(V))^\perp$ then for every $u \in V$ we have

$$\langle T^*v, u\rangle = \langle v, Tu\rangle = 0,$$

since $Tu \in T(V)$. Therefore we must have $T^*v = 0$ as desired.

(ii) We know that $\operatorname{null} T = (T^*(V))^\perp$. We also know that $T^*(V) \subset V$ is finite dimensional. Hence by Theorem 5.21 we get

$$(\operatorname{null} T)^\perp = \big((T^*(V))^\perp\big)^\perp = T^*(V).$$ ■

Remark. The above theorem provides a new tool for computing the orthogonal complement of a subspace, when that subspace is given as the image or the null space of an operator.

Theorem 6.10. *Suppose $T \in \mathcal{L}(V)$ has adjoint, and W is a T-invariant subspace of V. Then $W^\perp$ is T^*-invariant.*

Proof. Let $v \in W^\perp$ and $w \in W$. Then we have

$$\langle T^*v, w\rangle = \langle v, Tw\rangle = 0,$$

since $Tw \in W$ and $v \in W^\perp$. Hence T^*v is orthogonal to every $w \in W$. Thus $T^*v \in W^\perp$, and therefore $W^\perp$ is T^*-invariant. ■

6.2 Self-Adjoint Operators

Definition 6.11. Suppose the operator $T \in \mathcal{L}(V)$ has adjoint. Then T is called **self-adjoint** if $T^* = T$. In other words, T is self-adjoint if

$$\langle Tv, w\rangle = \langle v, Tw\rangle$$

for every $v, w \in V$.

Also, a square matrix $A \in F^{n\times n}$ is called **self-adjoint**, or **Hermitian**, if $A^* = A$; and it is called **symmetric** if $A^\intercal = A$.

Remark. It is obvious that for a square matrix $A \in \mathbb{R}^{n\times n}$, being self-adjoint is the same as being symmetric.

Theorem 6.12. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Then $T \in \mathcal{L}(V)$ is self-adjoint if and only if $[T]_\mathcal{B}$ is self-adjoint.*

Remark. This theorem is not true if the basis $\mathcal{B}$ is not orthonormal.

Proof. Let $A := [T]_\mathcal{B}$. Then $A^* = [T^*]_\mathcal{B}$, since $\mathcal{B}$ is orthonormal. Hence we have

$$T = T^* \iff [T]_\mathcal{B} = [T^*]_\mathcal{B} \iff A = A^*.$$

Note that we have used the fact that an operator is uniquely determined by its matrix. ■

***Exercise* 6.13.** Suppose W is a subspace of V, and P is the orthogonal projection on W. Show that P is self-adjoint.

Solution. We know that $V = W \oplus W^\perp$. Let $v, \tilde{v} \in V$. Then there are uniquely determined vectors $w, \tilde{w} \in W$ and $u, \tilde{u} \in W^\perp$, such that $v = w + u$ and $\tilde{v} = \tilde{w} + \tilde{u}$. By definition we have $Pv = w$, $P\tilde{v} = \tilde{w}$. Now we have

$$\begin{aligned}\langle Pv, \tilde{v}\rangle = \langle w, \tilde{w} + \tilde{u}\rangle &= \langle w, \tilde{w}\rangle + \langle w, \tilde{u}\rangle = \langle w, \tilde{w}\rangle + 0 \\ &= \langle w, \tilde{w}\rangle + \langle u, \tilde{w}\rangle = \langle w + u, \tilde{w}\rangle = \langle v, \tilde{w}\rangle = \langle v, P\tilde{v}\rangle.\end{aligned}$$

Hence P is self-adjoint. ■

Theorem 6.14. *The eigenvalues of a self-adjoint operator are real.*

Similarly, the eigenvalues of a self-adjoint matrix $A \in F^{n\times n}$ are real.

Proof. Suppose T is a self-adjoint operator, and v is an eigenvector of T corresponding to the eigenvalue λ. If $F = \mathbb{R}$ then λ is a real number by definition. If $F = \mathbb{C}$ then we have

$$\bar{\lambda}\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Tv\rangle = \langle Tv, v\rangle = \langle \lambda v, v\rangle = \lambda\langle v, v\rangle.$$

But $v \neq 0$ so $\langle v, v\rangle \neq 0$. Hence we must have $\bar{\lambda} = \lambda$, and therefore λ is real. The case of self-adjoint matrices can be proved similarly, by using the standard inner product of F^n. ■

Theorem 6.15. *The eigenvectors of a self-adjoint operator corresponding to distinct eigenvalues are orthogonal.*

Similarly, the eigenvectors of a self-adjoint matrix $A \in F^{n\times n}$ corresponding to distinct eigenvalues are orthogonal.

Proof. Suppose T is a self-adjoint operator, and v, w are eigenvectors of T corresponding to distinct eigenvalues λ, μ respectively. Then we have

$$\lambda\langle v, w\rangle = \langle \lambda v, w\rangle = \langle Tv, w\rangle = \langle v, Tw\rangle = \langle v, \mu w\rangle = \bar{\mu}\langle v, w\rangle = \mu\langle v, w\rangle,$$

where in the last equality we used the fact that μ is real. But $\lambda \neq \mu$, so we must have $\langle v, w\rangle = 0$ as desired. The case of matrices can be proved similarly, by using the standard inner product of F^n. ■

Proposition 6.16. *Suppose $T \in \mathcal{L}(V)$ is self-adjoint, and $W \subset V$ is a T-invariant subspace. Then $T|_W \in \mathcal{L}(W)$ is also self-adjoint.*

Proof. Let $S := T|_W$. Then for $w \in W$ we have $Sw = Tw$. Now remember that the inner product on W is just the restriction of the inner product of V. Hence for $u, w \in W$ we have

$$\langle Sw, u\rangle = \langle Tw, u\rangle = \langle w, Tu\rangle = \langle w, Su\rangle.$$ ■

Theorem 6.17. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is self-adjoint. Then T has at least one eigenvalue.*

Similarly, every symmetric matrix $A \in \mathbb{R}^{n\times n}$ has at least one real eigenvalue.

Remark. We know that every matrix in $\mathbb{C}^{n\times n}$ has at least one eigenvalue, since $\mathbb{C}$ is algebraically closed. So we only included real self-adjoint matrices, i.e. real symmetric matrices, in the above theorem.

Proof. We know that every operator on a nonzero finite dimensional complex vector space has at least one eigenvalue, since $\mathbb{C}$ is algebraically closed. So we only need to prove the theorem when $F = \mathbb{R}$. Suppose $\mathcal{B}$ is an orthonormal basis for V, and $A := [T]_{\mathcal{B}} \in \mathbb{R}^{n\times n}$, where $n = \dim V$. We know that A is a self-adjoint matrix. So A is a real symmetric matrix. Now consider the linear map $S : \mathbb{C}^n \to \mathbb{C}^n$ which maps $z \in \mathbb{C}^n$ to $Sz := Az$. Then we have $[S]_{\mathcal{C}} = A$, where $\mathcal{C}$ is the standard basis of $\mathbb{C}^n$. Therefore S is a self-adjoint operator, since A is self-adjoint, and $\mathcal{C}$ is an orthonormal basis.

Now we know that S has at least one eigenvalue $\lambda \in \mathbb{C}$. But S is self-adjoint, so $\lambda \in \mathbb{R}$. Let $z \in \mathbb{C}^n$ be an eigenvector of S corresponding to λ. We know that $z = x + iy$, for some $x, y \in \mathbb{R}^n$. Then we have

$$Ax + iAy = Az = \lambda z = \lambda x + i\lambda y.$$

Note that the entries of A are real, therefore $Ax, Ay \in \mathbb{R}^n$. Now in the above equality, each component of both sides must be equal. Hence the real part and the imaginary part of each component of both sides are equal too. Thus we have

$$Ax = \lambda x, \qquad Ay = \lambda y.$$

But $z \neq 0$, since z is an eigenvector. So at least one of x, y is nonzero. Suppose $x \neq 0$. Let $v \in V$ be the vector that satisfies $[v]_{\mathcal{B}} = x$. Then we have

$$[Tv]_{\mathcal{B}} = [T]_{\mathcal{B}}[v]_{\mathcal{B}} = Ax = \lambda x = \lambda[v]_{\mathcal{B}} = [\lambda v]_{\mathcal{B}}.$$

Hence $Tv = \lambda v$, because the coordinate isomorphism is one-to-one. Note that in the above, we have also used the linearity of the coordinate isomorphism. Finally note that $v \neq 0$, since $x \neq 0$. Therefore v is an eigenvector of T, and λ is an eigenvalue of T. Finally note that the result for real symmetric matrices is actually proved here too. ■

Theorem 6.18. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is self-adjoint. Then T is diagonalizable. Furthermore, V has an orthonormal basis consisting of the eigenvectors of T.*

Remark. In other words, V has an orthonormal basis $\mathcal{B}$ such that $[T]_{\mathcal{B}}$ is a diagonal matrix. Note that the diagonal entries of $[T]_{\mathcal{B}}$ are all real numbers, since they are the eigenvalues of T. The converse of this result is also true as we will show in the next theorem.

Proof. The proof is by induction on $\dim V$. When $\dim V = 1$ the result holds trivially, because every operator on a one-dimensional space is just multiplication by some scalar. So any vector in the space with norm one is an orthonormal basis

for the space, and an eigenvector of the operator T. Now suppose the theorem holds for every self-adjoint operator on a nonzero inner product space whose dimension is less than $\dim V$. We know that T has at least one eigenvalue, since it is self-adjoint. Let λ be an eigenvalue of T. Set $W := E_\lambda(T)$. Note that $W \neq \{0\}$, since it is an eigenspace. If $W = V$ then every vector in V is an eigenvector of T. Hence if $\mathcal{B}$ is an arbitrary orthonormal basis for V, then its elements are eigenvectors of T. Thus by Theorem 4.26, T is diagonalizable too.

So let us assume that $W \neq V$. Remember that we have $V = W \oplus W^\perp$. Therefore we get

$$0 < \dim W^\perp = \dim V - \dim W < \dim V,$$

since $0 < \dim W < \dim V$. On the other hand note that W is T-invariant. So by Theorem 6.10, $W^\perp$ is invariant under $T^* = T$, i.e. $W^\perp$ is also T-invariant. Let $S := T|_{W^\perp}$. Then by Proposition 6.16, S is a self-adjoint operator on the nonzero space $W^\perp$. Hence by the induction hypothesis, $W^\perp$ has an orthonormal basis consisting of the eigenvectors of S. Let us denote this basis by $\mathcal{B}$. Let $\mathcal{C}$ be an orthonormal basis for W. Note that every vector in W is an eigenvector of T. Now $\mathcal{B} \cup \mathcal{C}$ is a basis for V whose elements are eigenvectors of T, because the eigenvectors of S are also eigenvectors of T, as shown in Exercise 4.6. Hence by Theorem 4.26, T is diagonalizable. In addition, $\mathcal{B} \cup \mathcal{C}$ is orthonormal by Theorem 5.22. Therefore $\mathcal{B} \cup \mathcal{C}$ is an orthonormal basis for V whose elements are eigenvectors of T. ■

Theorem 6.19. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$. If V has an orthonormal basis $\mathcal{B}$ such that $[T]_\mathcal{B}$ is a diagonal matrix whose diagonal entries are real, then T is self-adjoint.*

Proof. The matrix $[T]_\mathcal{B}$ is a real symmetric matrix, because it is diagonal, and its diagonal entries are real numbers. Thus $[T]_\mathcal{B}$ is a self-adjoint matrix. Hence T is also self-adjoint, since $\mathcal{B}$ is orthonormal. ■

Definition 6.20. Suppose V is finite dimensional, and $S, T \in \mathcal{L}(V)$. We say S, T are **simultaneously diagonalizable** if there exists a basis $\mathcal{B}$ for V such that the matrices $[S]_\mathcal{B}, [T]_\mathcal{B}$ are both diagonal.

Remark. In other words, S, T are simultaneously diagonalizable if V has a basis whose elements are eigenvectors of both S, T. Although the corresponding eigenvalues of S, T are not necessarily the same.

Theorem 6.21. *Suppose V is finite dimensional, and $S, T \in \mathcal{L}(V)$ are self-adjoint. Then S, T are simultaneously diagonalizable if and only if $ST = TS$.*

Furthermore if S, T commute, then V has an orthonormal basis whose elements are eigenvectors of both S, T, i.e. V has an orthonormal basis $\mathcal{B}$ such that the matrices $[S]_\mathcal{B}, [T]_\mathcal{B}$ are both diagonal.

Remark. The first part of this theorem is true without assuming that the operators are self-adjoint, but the proof is harder in the general case. However this special case has interesting applications. For example in quantum mechanics the observables correspond to self-adjoint operators, and the observed values of an observable are the eigenvalues of its operator. In this setting, the simultaneous diagonalizability of two self-adjoint operators means that we can measure their corresponding observables with any precision at the same time. Equivalently, if two self-adjoint operators do not commute then we cannot measure their corresponding observables with any desired precision at the same time. This interesting phenomenon is related to the *Heisenberg's uncertainty principle*.

Proof. Suppose $\mathcal{B}$ is a basis for V such that the matrices $[S]_\mathcal{B}, [T]_\mathcal{B}$ are both diagonal. Then we have

$$[ST]_\mathcal{B} = [S]_\mathcal{B}[T]_\mathcal{B} = [T]_\mathcal{B}[S]_\mathcal{B} = [TS]_\mathcal{B},$$

since diagonal matrices commute. Hence we have $ST = TS$, because the matrix of an operator uniquely determines that operator. Conversely suppose that $ST = TS$. Let $\lambda_1, \dots, \lambda_k$ be all the distinct eigenvalues of T. We know that T is diagonalizable, so by Theorem 4.26 we have

$$V = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T).$$

Note that each $E_{\lambda_j}(T)$ is nonzero, since it contains at least one nonzero eigenvector. Now note that for $v \in E_{\lambda_j}(T)$ we have $Tv = \lambda_j v$. So

$$TSv = STv = S(\lambda_j v) = \lambda_j Sv.$$

Thus $Sv \in E_{\lambda_j}(T)$. Hence each $E_{\lambda_j}(T)$ is S-invariant.

Let $S_j := S|_{E_{\lambda_j}(T)}$. Then by Proposition 6.16, S_j is a self-adjoint operator on the nonzero space $E_{\lambda_j}(T)$. Therefore $E_{\lambda_j}(T)$ has an orthonormal basis whose elements are eigenvectors of S. Let us denote this basis by $\mathcal{B}_j$. Note that the elements of $\mathcal{B}_j$ are also eigenvectors of T, corresponding to the eigenvalue λ_j. Let $\mathcal{B} := \bigcup_{j \le k} \mathcal{B}_j$. Then by Theorem 2.55, $\mathcal{B}$ is a basis for V. Furthermore, the elements of $\mathcal{B}$ are eigenvectors of both S, T, because the eigenvectors of the restrictions of an operator are also eigenvectors of the operator itself, as shown in Exercise 4.6. Hence by Theorem 4.25, the matrices of both S, T in the basis $\mathcal{B}$ are diagonal, i.e. S, T are simultaneously diagonalizable.

In addition, $\mathcal{B}$ is orthonormal. Because all its elements have norm one. And any two vectors in $\mathcal{B}$ are orthogonal, since if they both belong to some $\mathcal{B}_j$, then they are orthogonal as $\mathcal{B}_j$ is orthonormal; otherwise one of the vectors is in some $\mathcal{B}_i$ and the other one is in some $\mathcal{B}_j$ for $j \neq i$, but then the two vectors must be orthogonal, because they are eigenvectors of the self-adjoint operator T corresponding to distinct

eigenvalues λ_i, λ_j. Therefore $\mathcal{B}$ is an orthonormal basis for V whose elements are eigenvectors of both S, T. ■

Second Proof. Suppose $ST = TS$. We prove the desired result by induction on $\dim V$. When $\dim V = 1$ the result holds trivially, because every operator on a one-dimensional space is just multiplication by some scalar. So any vector in the space with norm one is an orthonormal basis for the space, and an eigenvector of both S, T. Now suppose the result holds for every pair of commuting self-adjoint operators on a nonzero inner product space whose dimension is less than $\dim V$.

We know that T has at least one eigenvalue, since it is self-adjoint. Let λ be an eigenvalue of T. Set $W := E_\lambda(T)$. Note that $W \neq \{0\}$, since it is an eigenspace. If $W = V$ then every vector in V is an eigenvector of T. Let $\mathcal{B}$ be an orthonormal basis for V consisting of the eigenvectors of S. We know that such a basis exists, since S is self-adjoint. Then the elements of $\mathcal{B}$ are eigenvectors of both S, T. Thus by Theorem 4.25, the matrices $[S]_\mathcal{B}, [T]_\mathcal{B}$ are both diagonal, i.e. S, T are simultaneously diagonalizable.

So let us assume that $W \neq V$. Remember that we have $V = W \oplus W^\perp$. Therefore we get

$$0 < \dim W^\perp = \dim V - \dim W < \dim V,$$

since $0 < \dim W < \dim V$. On the other hand note that W is T-invariant. So by Theorem 6.10, $W^\perp$ is invariant under $T^* = T$, i.e. $W^\perp$ is also T-invariant. In addition note that for $v \in W$ we have $Tv = \lambda v$. So

$$TSv = STv = S(\lambda v) = \lambda Sv.$$

Thus $Sv \in E_\lambda(T) = W$. Hence W is S-invariant too. Furthermore, by Theorem 6.10, $W^\perp$ is invariant under $S^* = S$, i.e. $W^\perp$ is also S-invariant.

Let $T_1 := T|_{W^\perp}$ and $S_1 := S|_{W^\perp}$. Then by Proposition 6.16, S_1, T_1 are self-adjoint operators on the nonzero space $W^\perp$. Also note that for every $v \in W^\perp$ we have

$$S_1 T_1 v = S|_{W^\perp} T|_{W^\perp} v = STv = TSv = T|_{W^\perp} S|_{W^\perp} v = T_1 S_1 v.$$

Note that $S|_{W^\perp} v = Sv \in W^\perp$, so we can apply the operator $T|_{W^\perp}$ to it. Thus we have $S_1 T_1 = T_1 S_1$. Hence by the induction hypothesis, $W^\perp$ has an orthonormal basis whose elements are eigenvectors of both S_1, T_1. Let us denote this basis by $\mathcal{B}$. Similarly we can find an orthonormal basis $\mathcal{C}$ for W, whose elements are eigenvectors of both $S|_W, T|_W$.

Then $\mathcal{B} \cup \mathcal{C}$ is a basis for V whose elements are eigenvectors of both S, T. Because the eigenvectors of the restrictions of an operator are also eigenvectors of the operator itself, as shown in Exercise 4.6. Hence by Theorem 4.25, the matrices of both S, T in the basis $\mathcal{B} \cup \mathcal{C}$ are diagonal, i.e. S, T are simultaneously diagonalizable. In addition, $\mathcal{B} \cup \mathcal{C}$ is orthonormal by Theorem 5.22. Therefore $\mathcal{B} \cup \mathcal{C}$ is an orthonormal basis for V whose elements are eigenvectors of both S, T. ■

Proposition 6.22. *Suppose $T \in \mathcal{L}(V)$ is self-adjoint. Then for all $v \in V$, $\langle Tv, v\rangle$ is a real number.*

Proof. We have $\overline{\langle Tv, v\rangle} = \langle v, Tv\rangle = \langle Tv, v\rangle$, so $\langle Tv, v\rangle$ is real. Note that the first equality holds due to the conjugate symmetry of the inner product, and the second equality holds since T is self-adjoint. ∎

Theorem 6.23. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is self-adjoint. Let $\lambda_{\max}, \lambda_{\min}$ be the largest and the smallest eigenvalues of T respectively. Then we have*

$$\lambda_{\max} = \max_{v \in V - \{0\}} \frac{\langle Tv, v\rangle}{\|v\|^2}, \qquad \lambda_{\min} = \min_{v \in V - \{0\}} \frac{\langle Tv, v\rangle}{\|v\|^2}.$$

Remark. The expression $\langle Tv, v\rangle / \|v\|^2$ is called the *Rayleigh quotient*. Note that the above theorem also states that the Rayleigh quotient attains its maximum and minimum on $V - \{0\}$. The significance of this theorem is that it provides us a new tool for computing eigenvalues, besides the usual method of finding the roots of some polynomial associated to T. Finally we should mention that this result is not true if T is not self-adjoint.

Proof. Suppose $v_1, \ldots, v_n$ is an orthonormal basis for V consisting of the eigenvectors of T. Hence we have $Tv_j = \lambda_j v_j$ for some $\lambda_j \in \mathbb{R}$. Suppose we have arranged $v_1, \ldots, v_n$ so that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then $\lambda_{\min} = \lambda_1$ and $\lambda_{\max} = \lambda_n$. Now suppose for $v \in V$ we have $v = a_1 v_1 + \cdots + a_n v_n$. Then $Tv = a_1\lambda_1 v_1 + \cdots + a_n \lambda_n v_n$. Hence by Theorem 5.14 we have

$$\langle Tv, v\rangle = a_1\lambda_1\bar{a}_1 + \cdots + a_n\lambda_n\bar{a}_n = \lambda_1|a_1|^2 + \cdots + \lambda_n|a_n|^2.$$

We also have $\|v\|^2 = |a_1|^2 + \cdots + |a_n|^2$. Therefore if $v \neq 0$ we have

$$\lambda_1 = \frac{\lambda_1 \sum_{j\leq n} |a_j|^2}{\sum_{j\leq n} |a_j|^2} \leq \frac{\langle Tv, v\rangle}{\|v\|^2} = \frac{\sum_{j\leq n} \lambda_j |a_j|^2}{\sum_{j\leq n} |a_j|^2} \leq \frac{\lambda_n \sum_{j\leq n} |a_j|^2}{\sum_{j\leq n} |a_j|^2} = \lambda_n.$$

So the Rayleigh quotient is bounded by λ_1, λ_n. On the other hand

$$\frac{\langle Tv_1, v_1\rangle}{\|v_1\|^2} = \lambda_1, \qquad \frac{\langle Tv_n, v_n\rangle}{\|v_n\|^2} = \lambda_n.$$

Thus λ_1, λ_n are the minimum and the maximum of the Rayleigh quotient respectively. ∎

Remark. With a little extra effort in the above proof, we can show that for every $j \leq n$ we have

$$\lambda_j = \min_{\substack{U \text{ is a subspace} \\ \text{of } V,\, \dim U = j}} \max_{v \in U - \{0\}} \frac{\langle Tv, v\rangle}{\|v\|^2}.$$

This is a case of the so-called *min-max principle*.

6.3 Normal Operators

Definition 6.24. Suppose the operator $T \in \mathcal{L}(V)$ has adjoint. Then T is called **normal** if

$$T^*T = TT^*.$$

Also, a square matrix $A \in F^{n\times n}$ is called **normal** if $A^*A = AA^*$.

Remark. Note that if T is normal then T^* is also normal, since $(T^*)^* = T$. The same comment is true about normal matrices.

***Example* 6.25.** Self-adjoint operators or matrices are also normal.

Theorem 6.26. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Then $T \in \mathcal{L}(V)$ is normal if and only if $[T]_\mathcal{B}$ is normal.*

Remark. This theorem is not true if the basis $\mathcal{B}$ is not orthonormal.

Proof. Let $A := [T]_\mathcal{B}$. Then $A^* = [T^*]_\mathcal{B}$, since $\mathcal{B}$ is orthonormal. Hence we have $AA^* = [T]_\mathcal{B}[T^*]_\mathcal{B} = [TT^*]_\mathcal{B}$. Similarly we have $A^*A = [T^*T]_\mathcal{B}$. Therefore

$$TT^* = T^*T \iff [TT^*]_\mathcal{B} = [T^*T]_\mathcal{B} \iff AA^* = A^*A.$$

Note that we have used the fact that an operator is uniquely determined by its matrix. ■

Theorem 6.27. *Suppose $T \in \mathcal{L}(V)$ is normal. Then for every $v \in V$ we have*

$$\|T^*v\| = \|Tv\|.$$

Proof. We have

$$\|Tv\|^2 = \langle Tv, Tv\rangle = \langle v, T^*Tv\rangle = \langle v, TT^*v\rangle = \langle T^*v, T^*v\rangle = \|T^*v\|^2.$$ ■

Theorem 6.28. *Suppose $T \in \mathcal{L}(V)$ is normal, and v is an eigenvector of T corresponding to the eigenvalue λ. Then v is also an eigenvector of T^* corresponding to the eigenvalue $\bar{\lambda}$.*

Proof. Let $S := T - \lambda I$, where I is the identity map of V. Then $S^* = T^* - \bar{\lambda}I$, and we have

$$\begin{aligned} S^*S &= (T^* - \bar{\lambda}I)(T - \lambda I) = T^*T - \bar{\lambda}T - \lambda T^* + |\lambda|^2 I \\ &= TT^* - \lambda T^* - \bar{\lambda}T + |\lambda|^2 I = (T - \lambda I)(T^* - \bar{\lambda}I) = SS^*. \end{aligned}$$

Hence S is normal too. Now note that $Sv = Tv - \lambda v = 0$. Therefore by the previous theorem we have $\|S^*v\| = \|Sv\| = 0$. Thus $T^*v - \bar{\lambda}v = S^*v = 0$, as desired. ■

Theorem 6.29. *The eigenvectors of a normal operator corresponding to distinct eigenvalues are orthogonal.*

Proof. Suppose T is a normal operator, and v, w are eigenvectors of T corresponding to distinct eigenvalues λ, μ respectively. Then we have

$$\lambda\langle v, w\rangle = \langle \lambda v, w\rangle = \langle Tv, w\rangle = \langle v, T^*w\rangle = \langle v, \bar{\mu}w\rangle = \bar{\bar{\mu}}\langle v, w\rangle = \mu\langle v, w\rangle.$$

Note that we used the fact that $T^*w = \bar{\mu}w$. Now since $\lambda \neq \mu$, we must have $\langle v, w\rangle = 0$ as desired. ∎

Theorem 6.30. *Suppose $T \in \mathcal{L}(V)$ is normal, and $W \subset V$ is a finite dimensional T-invariant subspace. Then*

(i) *$W, W^\perp$ are both T-invariant and T^*-invariant.*
(ii) *$T|_W \in \mathcal{L}(W)$ is normal, and we have $\left(T|_W\right)^* = T^*|_W$.*
(iii) *$T|_{W^\perp} \in \mathcal{L}(W^\perp)$ is normal, and we have $\left(T|_{W^\perp}\right)^* = T^*|_{W^\perp}$.*

Remark. Note that V can be infinite dimensional. Thus $W^\perp$ can be infinite dimensional too.

Proof. **(i)** Let $w_1, \ldots, w_m$ be an orthonormal basis for W. Then for every $j \leq m$ we have $Tw_j \in W$. So there are $a_{ij} \in F$ such that $Tw_j = \sum_{i\leq m} a_{ij}w_i$. On the other hand, we know that $V = W \oplus W^\perp$. Thus there are $u_j \in W$ and $v_j \in W^\perp$ such that $T^*w_j = u_j + v_j$. Hence there are $b_{ij} \in F$ such that $u_j = \sum_{i\leq m} b_{ij}w_i$. Therefore

$$T^*w_j = \sum_{i\leq m} b_{ij}w_i + v_j.$$

Now for every k, j we have

$$\begin{aligned} a_{kj} &= \sum_{i\leq m} a_{ij}\langle w_i, w_k\rangle = \left\langle \sum_{i\leq m} a_{ij}w_i, w_k \right\rangle = \langle Tw_j, w_k\rangle \\ &= \langle w_j, T^*w_k\rangle = \left\langle w_j, \sum_{i\leq m} b_{ik}w_i + v_k \right\rangle = \sum_{i\leq m} \bar{b}_{ik}\langle w_j, w_i\rangle + \langle w_j, v_k\rangle = \bar{b}_{jk}. \end{aligned}$$

Hence we have $|a_{kj}| = |\bar{b}_{jk}| = |b_{jk}|$. In addition note that by Theorem 5.14 we have

$$\|Tw_j\|^2 = \sum_{k\leq m} |a_{kj}|^2, \qquad \|u_j\|^2 = \sum_{k\leq m} |b_{kj}|^2.$$

Also, by Pythagorean theorem we have $\|T^*w_j\|^2 = \|u_j\|^2 + \|v_j\|^2$. Therefore by

Theorem 6.27 we get

$$\begin{aligned}\sum_{j\le m}\sum_{k\le m}|b_{kj}|^2+\sum_{j\le m}\|v_j\|^2 &= \sum_{j\le m}\|T^*w_j\|^2=\sum_{j\le m}\|Tw_j\|^2\\ &=\sum_{j\le m}\sum_{k\le m}|a_{kj}|^2=\sum_{j\le m}\sum_{k\le m}|b_{jk}|^2\\ &=\sum_{k\le m}\sum_{j\le m}|b_{jk}|^2=\sum_{j\le m}\sum_{k\le m}|b_{kj}|^2.\end{aligned}$$

Note that in the last line we first changed the order of summation as in Theorem A.68, then we changed the name of the index j to k, and the index k to j. Now the above equality implies that $\sum_{j\le m}\|v_j\|^2=0$. Hence $\|v_j\|=0$ for every j. Thus we must have $T^*w_j=u_j+0=u_j\in W$. Therefore as shown in Exercise 4.2, W is T^*-invariant. So by Theorem 6.10, $W^\perp$ is invariant under $(T^*)^*=T$, i.e. $W^\perp$ is also T-invariant. Similarly, $W^\perp$ is T^*-invariant too, since W is T-invariant.

(ii) Let $S:=T|_W$. First note that $S^*=T^*|_W$, since for every $v,w\in W$ we have

$$\langle Sv,w\rangle=\langle T|_Wv,w\rangle=\langle Tv,w\rangle=\langle v,T^*w\rangle=\langle v,T^*|_Ww\rangle.$$

Note that this argument would not work if W were not invariant under T^*, because the adjoint of $S\in\mathcal{L}(W)$ must be an operator in $\mathcal{L}(W)$, not in $\mathcal{L}(V)$. Now note that for every $v\in W$ we have

$$SS^*v=T|_WT^*|_Wv=TT^*v=T^*Tv=T^*|_WT|_Wv=S^*Sv.$$

Note that $T|_Wv=Tv\in W$, so we can apply the operator $T^*|_W$ to it. Thus we have $SS^*=S^*S$, and therefore S is normal.

(iii) The proof is the same as in part (ii). Note that in the proof of part (ii) we did not use the fact that W is finite dimensional, we only used the fact that W is invariant under both T,T^*. ■

Theorem 6.31. *Suppose $F=\mathbb{C}$, and V is finite dimensional. Let $T\in\mathcal{L}(V)$ be a normal operator. Then T is diagonalizable. Furthermore, V has an orthonormal basis consisting of the eigenvectors of T.*

Remark. In other words, V has an orthonormal basis $\mathcal{B}$ such that $[T]_\mathcal{B}$ is a diagonal matrix. The converse of this result is also true as we will show in the next theorem.

Remark. The above theorem is not true when $F=\mathbb{R}$, i.e. normal operators on real inner product spaces are not necessarily diagonalizable.

Proof. The proof is by induction on $\dim V$. When $\dim V=1$ the result holds trivially, because every operator on a one-dimensional space is just multiplication by some scalar. So any vector in the space with norm one is an orthonormal basis

for the space, and an eigenvector of the operator T. Now suppose the theorem holds for every normal operator on a nonzero complex inner product space whose dimension is less than $\dim V$. We know that T has at least one eigenvalue, since $F = \mathbb{C}$. Let λ be an eigenvalue of T. Set $W := E_\lambda(T)$. Note that $W \neq \{0\}$, since it is an eigenspace. If $W = V$ then every vector in V is an eigenvector of T. Hence if $\mathcal{B}$ is an arbitrary orthonormal basis for V, then its elements are eigenvectors of T. Thus by Theorem 4.26, T is diagonalizable too.

So let us assume that $W \neq V$. Remember that we have $V = W \oplus W^\perp$. Therefore we get

$$0 < \dim W^\perp = \dim V - \dim W < \dim V,$$

since $0 < \dim W < \dim V$. On the other hand note that for every $v \in W$ we have $Tv = \lambda v$. Thus by Theorem 6.28 we have

$$T^*v = \bar{\lambda} v \in W.$$

Hence W is T^*-invariant. So by Theorem 6.10, $W^\perp$ is invariant under $(T^*)^* = T$, i.e. $W^\perp$ is T-invariant too. Let $S := T|_{W^\perp}$. By Theorem 6.30, S is a normal operator on the nonzero space $W^\perp$. Note that since W is T-invariant, Theorem 6.30 also implies that $W^\perp$ is T-invariant. But the above reasoning for the T-invariance of $W^\perp$ is much simpler than the one presented in the proof of Theorem 6.30, due to the simple description of W.

Hence by the induction hypothesis, $W^\perp$ has an orthonormal basis consisting of the eigenvectors of S. Let us denote this basis by $\mathcal{B}$. Let $\mathcal{C}$ be an orthonormal basis for W. Note that every vector in W is an eigenvector of T. Now $\mathcal{B} \cup \mathcal{C}$ is a basis for V whose elements are eigenvectors of T. Because the eigenvectors of S are also eigenvectors of T, as shown in Exercise 4.6. Hence by Theorem 4.26, T is diagonalizable. In addition, $\mathcal{B} \cup \mathcal{C}$ is orthonormal by Theorem 5.22. Therefore $\mathcal{B} \cup \mathcal{C}$ is an orthonormal basis for V whose elements are eigenvectors of T. ■

Theorem 6.32. *Suppose $F = \mathbb{C}$, and V is finite dimensional. Let $T \in \mathcal{L}(V)$, and suppose that V has an orthonormal basis $\mathcal{B}$ such that $[T]_\mathcal{B}$ is diagonal. Then T is normal.*

Proof. Note that $[T^*]_\mathcal{B} = [T]_\mathcal{B}^*$, since $\mathcal{B}$ is orthonormal. But $[T]_\mathcal{B}$ is diagonal, so $[T]_\mathcal{B}^*$ is also diagonal. Therefore $[T]_\mathcal{B}, [T]_\mathcal{B}^*$ commute, since diagonal matrices commute. Hence $[T]_\mathcal{B}$ is normal. Thus T is normal too, because $\mathcal{B}$ is an orthonormal basis. ■

Theorem 6.33. *Suppose V is a nonzero finite dimensional vector space over $\mathbb{R}$, and $T \in \mathcal{L}(V)$. Then there exists a T-invariant subspace $W \subset V$ such that* $\dim W$ *is either* 1 *or* 2.

Remark. We know that every operator on a finite dimensional complex vector space has at least one eigenvector, or equivalently it has a one dimensional invariant subspace. But operators on finite dimensional real vector spaces do not have this property. However, the above theorem says that a weaker version of this property holds for operators on real vector spaces. This theorem holds because of the special relationship between $\mathbb{R}$ and $\mathbb{C}$. There is no similar result about operators on vector spaces over an arbitrary field.

Proof. Suppose $\mathcal{B}$ is a basis for V, and $A := [T]_{\mathcal{B}} \in \mathbb{R}^{n\times n}$, where $n = \dim V$. Consider the linear map $S : \mathbb{C}^n \to \mathbb{C}^n$ which maps $z \in \mathbb{C}^n$ to $Sz := Az$. Then we have $[S]_{\mathcal{C}} = A$, where $\mathcal{C}$ is the standard basis of $\mathbb{C}^n$. Now we know that S has at least one eigenvalue $\lambda \in \mathbb{C}$. Suppose $\lambda = a + ib$, where $a, b \in \mathbb{R}$. Let $z \in \mathbb{C}^n$ be an eigenvector of S corresponding to λ. We know that $z = x + iy$, for some $x, y \in \mathbb{R}^n$. Then we have

$$Ax + iAy = Az = \lambda z = \lambda x + i\lambda y = ax + ibx + iay - by.$$

Note that the entries of A are real, therefore $Ax, Ay \in \mathbb{R}^n$. Also, in the above equality, each component of both sides must be equal. Hence the real part and the imaginary part of each component of both sides are equal too. Thus we have

$$Ax = ax - by, \qquad Ay = bx + ay.$$

Now let $v, w \in V$ be the vectors that satisfy $[v]_{\mathcal{B}} = x$ and $[w]_{\mathcal{B}} = y$. Then we have

$$[Tv]_{\mathcal{B}} = [T]_{\mathcal{B}}[v]_{\mathcal{B}} = Ax = ax - by = a[v]_{\mathcal{B}} - b[w]_{\mathcal{B}} = [av - bw]_{\mathcal{B}},$$
$$[Tw]_{\mathcal{B}} = [T]_{\mathcal{B}}[w]_{\mathcal{B}} = Ay = bx + ay = b[v]_{\mathcal{B}} + a[w]_{\mathcal{B}} = [bv + aw]_{\mathcal{B}}.$$

Hence $Tv = av - bw$ and $Tw = bv + aw$, because the coordinate isomorphism is one-to-one. Note that in the above, we have also used the linearity of the coordinate isomorphism. Now let $W := \operatorname{span}(v, w)$. Note that $z \neq 0$, since z is an eigenvector. So at least one of x, y is nonzero. Therefore at least one of v, w is nonzero. Thus $\dim W$ is either 1 or 2. Also, we have $Tv, Tw \in \operatorname{span}(v, w)$. Hence as shown in Exercise 4.2, W is T-invariant, as desired. ∎

Remark. The above theorem can be used to characterize normal operators on a finite dimensional real inner product space, similarly to the Theorem 6.41.

6.4 Unitary Operators

Definition 6.34. Suppose V is an inner product space. For two vectors $u, v \in V$ we define their **distance** to be

$$d(u, v) := \|u - v\|.$$

The function $d : V \times V \to \mathbb{R}$ is called the **metric** on V induced by its norm.

Remark. Let $u, v, w \in V$. It is easy to see that the metric d on V has the following properties. Thus V equipped with d is a so-called *metric space*.

(i) $d(u, v) \geq 0$, and $d(u, v) = 0 \iff u = v$.
(ii) $d(u, v) = d(v, u)$.
(iii) $d(u, v) \leq d(u, w) + d(w, v)$.

To prove the first property note that $d(u, v) = \|u - v\| \geq 0$, and $d(u, u) = \|u - u\| = \|0\| = 0$. Now if $d(u, v) = 0$ then $u - v = 0$, hence $u = v$. To prove the second property we have

$$d(u, v) = \|u - v\| = \|(-1)(v - u)\| = |-1|\|v - u\| = d(v, u).$$

And for the third property we have

$$\begin{aligned} d(u, v) = \|u - v\| &= \|u - w + w - v\| \\ &\leq \|u - w\| + \|w - v\| = d(u, w) + d(w, v). \end{aligned}$$

Finally let us mention that the third property is also called the *triangle inequality*.

Remark. We say a sequence $(v_j)_{j \in \mathbb{N}}$ of vectors in V *converge* to the *limit* $v \in V$ if

$$\lim_{j \to \infty} d(v_j, v) = \lim_{j \to \infty} \|v_j - v\| = 0.$$

In this case we write $\lim_{j \to \infty} v_j = v$, or $v_j \to v$. It can be shown that the limit of a sequence in a metric space is unique, if it exists.

Definition 6.35. A function $f : V \to V$ is called an **isometry** if for every $u, v \in V$ we have

$$\|f(u) - f(v)\| = \|u - v\|.$$

Remark. In other words, an isometry is a function that preserves the distance between any two points.

Theorem 6.36. *Suppose $F = \mathbb{R}$, and $f : V \to V$ is an isometry. Then there exists a linear map $T \in \mathcal{L}(V)$ which is also an isometry, such that for every $v \in V$ we have*

$$f(v) = Tv + f(0).$$

Remark. In other words, every isometry is the composition of a linear isometry and a translation. Also, note that V can be infinite dimensional too.

Proof. Let $Tv := f(v) - f(0)$ for every $v \in V$. Note that $T(0) = 0$. It is also easy to see that T is an isometry. Because for every $u, v \in V$ we have

$$\|Tu - Tv\| = \big\|f(u) - f(0) - \big(f(v) - f(0)\big)\big\| = \|f(u) - f(v)\| = \|u - v\|.$$

So we only need to show that T is linear. Suppose u, v are distinct points of V. Let $w := \frac{1}{2}u + \frac{1}{2}v$ be the midpoint between u, v. Let $d := \frac{1}{2}\|u - v\|$. Then $d \neq 0$ since $u \neq v$. We also have

$$\|u - w\| = \|u - (\tfrac{1}{2}u + \tfrac{1}{2}v)\| = \|\tfrac{1}{2}u - \tfrac{1}{2}v\| = \tfrac{1}{2}\|u - v\| = d.$$

Similarly we have $\|w - v\| = d$. Hence we get

$$\|Tu - Tv\| = 2d, \qquad \|Tu - Tw\| = d, \qquad \|Tw - Tv\| = d,$$

since T is an isometry. Therefore we have

$$\begin{aligned}\|Tu - Tw + Tw - Tv\| = \|Tu - Tv\| &= 2d \\ &= d + d = \|Tu - Tw\| + \|Tw - Tv\|.\end{aligned}$$

Hence by Theorem 5.10 we must have

$$Tu - Tw = a(Tw - Tv), \qquad \text{or} \qquad Tw - Tv = a(Tu - Tw),$$

for some $a \in [0, \infty)$. Now note that $\|Tu - Tw\| = \|Tw - Tv\| = d \neq 0$, so in either case we must have $|a| = 1$, and consequently $a = \pm 1$. Thus in either case we obtain $Tu - Tw = \pm(Tw - Tv)$. But if we have $Tu - Tw = -(Tw - Tv)$ then we get $Tu = Tv$. This implies $2d = \|Tu - Tv\| = 0$, which is a contradiction.

Hence we must have $Tu - Tw = Tw - Tv$. Therefore $2Tw = Tu + Tv$, i.e.

$$T(\tfrac{1}{2}u + \tfrac{1}{2}v) = Tw = \tfrac{1}{2}Tu + \tfrac{1}{2}Tv.$$

Note that we obtained the above equation under the assumption that $u \neq v$, but we can easily check that it also holds when $u = v$. If we set $v = 0$ in the above equation we get $T(\frac{1}{2}u) = \frac{1}{2}Tu + \frac{1}{2}T(0) = \frac{1}{2}Tu$. So the above equation becomes

$$T(\tfrac{1}{2}u + \tfrac{1}{2}v) = \tfrac{1}{2}Tu + \tfrac{1}{2}Tv = T(\tfrac{1}{2}u) + T(\tfrac{1}{2}v).$$

This equation holds for every $u, v \in V$. Thus instead of u, v, we can insert any other vectors in this identity. Let us replace u with $2u$, and v with $2v$. Then we get

$$T(u + v) = Tu + Tv. \qquad (*)$$

Hence in order to show that T is linear, we only need to prove that $T(av) = aTv$ for every $a \in \mathbb{R}$ and $v \in V$.

Now if we set $u = v$ in $(*)$ we get $T(2v) = 2Tv$. Suppose we have shown that $T(kv) = kTv$ for some $k \in \mathbb{N}$. Then we have

$$T\big((k+1)v\big) = T(kv + v) = T(kv) + Tv = kTv + Tv = (k+1)Tv.$$

Therefore by induction we have shown that $T(nv) = nTv$ for every $n \in \mathbb{N}$. Note that this relation holds trivially for $n = 1$, since $T(1v) = Tv = 1Tv$. We also have $T(0v) = T(0) = 0 = 0Tv$. In addition, we have

$$\begin{aligned} 0 = T(0v) = T\big((-n+n)v\big) &= T(-nv + nv) \\ &= T(-nv) + T(nv) = T(-nv) + nTv. \end{aligned}$$

So we get

$$T(-nv) = -(nTv) = (-1)(nTv) = \big((-1)n\big)Tv = (-n)Tv.$$

Thus for every $m \in \mathbb{Z}$ we have $T(mv) = mTv$. Next note that for $n \in \mathbb{N}$ we have $Tv = T\big(n(\frac{1}{n}v)\big) = nT(\frac{1}{n}v)$. Hence we get $T(\frac{1}{n}v) = \frac{1}{n}Tv$. Thus for every $m \in \mathbb{Z}$ we have

$$T(\tfrac{m}{n}v) = T\big(m(\tfrac{1}{n}v)\big) = mT(\tfrac{1}{n}v) = m(\tfrac{1}{n}Tv) = \tfrac{m}{n}Tv.$$

Thus we have $T(rv) = rTv$ for every $r \in \mathbb{Q}$.

Finally let $a \in \mathbb{R}$. Then we know that there is a sequence of rational numbers $(r_j)_{j\in\mathbb{N}}$ such that $r_j \to a$ as $j \to \infty$. Hence we have

$$\|aTv - r_jTv\| = \|(a - r_j)Tv\| = |a - r_j|\|Tv\| \underset{j\to\infty}{\longrightarrow} 0\|Tv\| = 0,$$

i.e. $r_jTv \to aTv$. Note that in the above reasoning, we have only used the properties of the norm, and the continuity of the multiplication of real numbers. On the other hand, T is an isometry. So we have

$$\|T(av) - T(r_jv)\| = \|av - r_jv\| = \|(a - r_j)v\| = |a - r_j|\|v\| \underset{j\to\infty}{\longrightarrow} 0\|v\| = 0,$$

i.e. $T(r_jv) \to T(av)$. But we have $T(r_jv) = r_jTv$ for every j. Therefore

$$T(av) = \lim_{j\to\infty} T(r_jv) = \lim_{j\to\infty} r_jTv = aTv.$$

Note that we are using the fact that the limit of a sequence in a metric space is unique. ∎

Remark. The above theorem is not true when $F = \mathbb{C}$. Indeed if we define $Tv := f(v) - f(0)$, then T preserves addition and scalar multiplication by real numbers, but in general $T(cv) \neq cTv$ when $c \in \mathbb{C}$. In other words, T is $\mathbb{R}$-linear but it is not $\mathbb{C}$-linear. For example the map $z \mapsto \bar{z}$ from $\mathbb{C}$ to $\mathbb{C}$ is an isometry, but it is not $\mathbb{C}$-linear.

Definition 6.37. Suppose the operator $T \in \mathcal{L}(V)$ has adjoint. Then T is called **unitary** if

$$T^*T = I_V = TT^*.$$

When $F = \mathbb{R}$, a unitary operator is also called an **orthogonal** operator.

A square matrix $A \in F^{n\times n}$ is called **unitary** if $A^*A = I = AA^*$, and it is called **orthogonal** if $A^\intercal A = I = AA^\intercal$.

Remark. It is obvious that for a square matrix $A \in \mathbb{R}^{n\times n}$, being unitary is the same as being orthogonal. It is also trivial that a unitary operator or matrix is normal too.

Remark. Note that unitary operators or matrices are invertible by definition. In fact we can say that an operator T and a matrix A are unitary if

$$T^{-1} = T^*, \qquad A^{-1} = A^*,$$

respectively. Similarly a matrix A is orthogonal if $A^{-1} = A^\intercal$.

Remark. Note that if T is unitary then T^* is also unitary, since $(T^*)^* = T$. The same comment is true about unitary and orthogonal matrices.

Theorem 6.38. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Then $T \in \mathcal{L}(V)$ is unitary if and only if $[T]_\mathcal{B}$ is unitary.*

Remark. This theorem is not true if the basis $\mathcal{B}$ is not orthonormal.

Proof. Let $A := [T]_\mathcal{B}$. Then $A^* = [T^*]_\mathcal{B}$, since $\mathcal{B}$ is orthonormal. Hence we have $AA^* = [T]_\mathcal{B}[T^*]_\mathcal{B} = [TT^*]_\mathcal{B}$. Similarly we have $A^*A = [T^*T]_\mathcal{B}$. Therefore

$$TT^* = I_V = T^*T \iff [TT^*]_\mathcal{B} = [I_V]_\mathcal{B} = [T^*T]_\mathcal{B} \iff AA^* = I = A^*A.$$

Note that we have used the fact that an operator is uniquely determined by its matrix. ∎

Theorem 6.39. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$. Then the following statements are equivalent*

(i) *T is unitary.*
(ii) *T preserves the inner product, i.e. $\langle Tu, Tv\rangle = \langle u, v\rangle$ for every $u, v \in V$.*
(iii) *T preserves the norm, i.e. $\|Tv\| = \|v\|$ for all $v \in V$.*
(iv) *T is an isometry.*
(v) *For every orthonormal basis $\{v_1, \dots, v_n\}$ for V, $\{Tv_1, \dots, Tv_n\}$ is also an orthonormal basis for V.*
(vi) *There exists an orthonormal basis $\{v_1, \dots, v_n\}$ for V, such that $\{Tv_1, \dots, Tv_n\}$ is an orthonormal basis for V.*

Proof. **(i) $\Longrightarrow$ (ii):** We have

$$\langle u, T^*Tv\rangle - \langle u, v\rangle = \langle u, (T^*T - I)v\rangle = \langle u, 0\rangle = 0.$$

Hence $\langle Tu, Tv\rangle = \langle u, T^*Tv\rangle = \langle u, v\rangle$.

(ii) $\Longrightarrow$ (iii): We have $\|Tv\|^2 = \langle Tv, Tv\rangle = \langle v, v\rangle = \|v\|^2$.

(iii) $\Longrightarrow$ (iv): We have $\|Tu - Tv\| = \|T(u - v)\| = \|u - v\|$.

(iv) $\Longrightarrow$ (iii): We have $\|Tv\| = \|Tv - 0\| = \|Tv - T(0)\| = \|v - 0\| = \|v\|$.

(iii) $\Longrightarrow$ (ii): T must preserve the inner product, due to the polarization identities. For example when $F = \mathbb{R}$ we have

$$\begin{aligned}\langle Tu, Tv\rangle &= \frac{1}{4}\|Tu + Tv\|^2 - \frac{1}{4}\|Tu - Tv\|^2 \\ &= \frac{1}{4}\|T(u + v)\|^2 - \frac{1}{4}\|T(u - v)\|^2 = \frac{1}{4}\|u + v\|^2 - \frac{1}{4}\|u - v\|^2 = \langle u, v\rangle.\end{aligned}$$

The case of $F = \mathbb{C}$ can be proved similarly.

(ii) $\Longrightarrow$ (i): We have $\langle u, T^*Tv\rangle = \langle Tu, Tv\rangle = \langle u, v\rangle$. Therefore

$$0 = \langle u, T^*Tv\rangle - \langle u, v\rangle = \langle u, (T^*T - I)v\rangle.$$

Hence $(T^*T - I)v$ is orthogonal to every $u \in V$, so $(T^*T - I)v = 0$. Thus we have $T^*T - I = 0$, since v was an arbitrary vector in V. Therefore $T^*T = I$, and as V is finite dimensional we also have $TT^* = I$.

(iii) $\Longrightarrow$ (v): Note that since T preserves the norm, it also preserves the inner product. Therefore for every $i \neq j$ we have

$$\|Tv_j\| = \|v_j\| = 1, \qquad \langle Tv_i, Tv_j\rangle = \langle v_i, v_j\rangle = 0.$$

Thus $\{Tv_1, \ldots, Tv_n\}$ is an orthonormal set, so in particular it is linearly independent. But $\{Tv_1, \ldots, Tv_n\}$ has the same number of elements as $\dim V$. Hence it is a basis for V.

(v) $\Longrightarrow$ (vi): This is trivial.

(vi) $\Longrightarrow$ (iii): Let $v \in V$. Then there are $a_j \in F$ such that $v = \sum_{j\leq n} a_j v_j$. Hence we have

$$Tv = T\Big(\sum_{j\leq n} a_j v_j\Big) = \sum_{j\leq n} a_j Tv_j.$$

But $\{v_1, \ldots, v_n\}$ and $\{Tv_1, \ldots, Tv_n\}$ are orthonormal bases for V, so by Theorem 5.14 we have

$$\|Tv\|^2 = \sum_{j\leq n} |a_j|^2 = \|v\|^2.$$

■

Theorem 6.40. *Suppose $T \in \mathcal{L}(V)$ is unitary, and λ is an eigenvalue of T. Then we have $|\lambda| = 1$.*

Proof. Suppose $v \neq 0$, and $Tv = \lambda v$. We know that $T^*v = \bar{\lambda}v$, since T is normal. Now we have

$$v = I_V v = T^*Tv = T^*(\lambda v) = \lambda T^*v = \lambda\bar{\lambda}v = |\lambda|^2 v.$$

But $v \neq 0$ so we must have $|\lambda|^2 = 1$. ■

Remark. Suppose $F = \mathbb{C}$, and V is finite dimensional. Let $T \in \mathcal{L}(V)$ be a unitary operator. Then we know that T is diagonalizable, since T is normal too. We also know that the eigenvalues of T have absolute value one. Hence V has an orthonormal basis $\mathcal{B}$ such that

$$[T]_{\mathcal{B}} = \begin{bmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{i\theta_n} \end{bmatrix},$$

where $\theta_j \in [0, 2\pi)$. Remember that multiplication of complex numbers by $e^{i\theta}$ corresponds to the counterclockwise rotation in the complex plane by the angle θ. So, intuitively the above matrix form means that a unitary map on a complex inner product space is the composition of several rotations. Note that unlike the case of orthogonal operators on real inner product spaces discussed below, we do not need reflections here. Because multiplication by -1 in the complex plane is the same as the counterclockwise rotation by π radians.

In contrast when $F = \mathbb{R}$, orthogonal operators are not necessarily diagonalizable. However, using Theorem 6.33 we can obtain the following characterization of them.

Theorem 6.41. *Suppose $F = \mathbb{R}$, and V is finite dimensional. Let $T \in \mathcal{L}(V)$ be an orthogonal operator. Then V has an orthonormal basis*

$$\mathcal{B} = \{u_1, \ldots, u_k, v_1, w_1, \ldots, v_m, w_m\}$$

such that for every $j \leq k$ we either have $Tu_j = u_j$, or $Tu_j = -u_j$. And for every $j \leq m$ there is $\theta_j \in (0, \pi)$ so that

$$Tv_j = (\cos\theta_j)v_j + (\sin\theta_j)w_j, \qquad Tw_j = -(\sin\theta_j)v_j + (\cos\theta_j)w_j.$$

Remark. Note that k, m are nonnegative integers, so they can be zero too.

Remark. If $Tu_j = -u_j$ for some j, and every other vector in $\mathcal{B}$ is fixed under T, then T is a reflection through the subspace $(\text{span}(u_j))^{\perp}$. And if every vector in $\mathcal{B}$ other than v_j, w_j is fixed under T, then T is a rotation in the plane $\text{span}(v_j, w_j)$ by

the angle θ_j. Therefore the meaning of the above theorem is that an orthogonal operator is the composition of several reflections and several rotations. Consequently, this theorem and Theorem 6.36 imply that every isometry of a finite dimensional real inner product space, in particular every isometry of $\mathbb{R}^n$, is the composition of a translation and several reflections and rotations.

Remark. It is obvious that the subspaces $\operatorname{span}(u_j)$ and $\operatorname{span}(v_j, w_j)$ are T-invariant, as shown in Exercise 4.2. In addition, note that u_j is a basis for $\operatorname{span}(u_j)$, and v_j, w_j is a basis for $\operatorname{span}(v_j, w_j)$. Hence by Theorem 2.56 we have

$$V = \operatorname{span}(u_1) \oplus \cdots \oplus \operatorname{span}(u_k) \oplus \operatorname{span}(v_1, w_1) \oplus \cdots \oplus \operatorname{span}(v_m, w_m).$$

Therefore by using the notion of block diagonal matrices defined in Section 8.1, and employing Theorem 8.4, we conclude that $[T]_{\mathcal{B}}$ is a block diagonal matrix of the form

$$[T]_{\mathcal{B}} = \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{m+k} \end{bmatrix},$$

where each R_j is either $[1]$ or $[-1]$ when $j \leq k$, and $R_{j+k} = \begin{bmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{bmatrix}$ when $j \leq m$.

Proof. The proof is by induction on $\dim V$. When $\dim V = 1$ the result holds trivially, because every operator on a one-dimensional space is just multiplication by some scalar. So any vector in the space with norm one is an orthonormal basis for the space, and an eigenvector of the operator T. In addition, note that if λ is an eigenvalue of the orthogonal operator T, then $|\lambda| = 1$. However $\lambda \in \mathbb{R}$, so $\lambda = \pm 1$.

Now suppose the theorem holds for every orthogonal operator on a nonzero real inner product space whose dimension is less than $\dim V$. First suppose that T has at least one eigenvalue. Let λ be an eigenvalue of T. We know that $\lambda = \pm 1$. Set $W := E_\lambda(T)$. Note that $W \neq \{0\}$, since it is an eigenspace. If $W = V$ then every vector in V is an eigenvector of T. Hence if $\mathcal{B}$ is an arbitrary orthonormal basis for V, then its elements are eigenvectors of T, corresponding to the eigenvalues ± 1. Thus $\mathcal{B}$ has the desired properties.

So let us assume that $W \neq V$. Remember that we have $V = W \oplus W^\perp$. Therefore we get

$$0 < \dim W^\perp = \dim V - \dim W < \dim V,$$

since $0 < \dim W < \dim V$. On the other hand note that W is T-invariant. We claim that $W^\perp$ is also T-invariant. Let $v \in W^\perp$. We know that for every $w \in W$ we have $Tw = \lambda w$. Hence

$$\pm\langle Tv, w\rangle = \langle Tv, \pm w\rangle = \langle Tv, \lambda w\rangle = \langle Tv, Tw\rangle = \langle v, w\rangle = 0.$$

Thus $\langle Tv, w\rangle = 0$. Therefore Tv is orthogonal to every $w \in W$. Hence $Tv \in W^\perp$, and consequently $W^\perp$ is T-invariant as desired. Note that since W is T-invariant, and T is normal, Theorem 6.30 implies that $W^\perp$ is T-invariant. But the above reasoning for the T-invariance of $W^\perp$ is much simpler than the one presented in the proof of Theorem 6.30, due to the simple description of W, and the orthogonality of T.

Now let $S := T|_{W^\perp}$. Then S is an orthogonal operator on the nonzero space $W^\perp$. Because for every $v \in W^\perp$ we have $\|Sv\| = \|Tv\| = \|v\|$. Hence by the induction hypothesis, $W^\perp$ has an orthonormal basis $\mathcal{B}$ with the prescribed properties in the theorem. Let $\mathcal{C}$ be an orthonormal basis for W. Note that every vector in W is an eigenvector of T, corresponding to the eigenvalue $\lambda = \pm 1$. Now $\mathcal{C} \cup \mathcal{B}$ is a basis for V, which is also orthonormal due to Theorem 5.22. Furthermore, note that $\mathcal{C} \cup \mathcal{B}$ has the desired properties, since the action of T on an element of $\mathcal{B}$ is the same as the action of S on that element.

Finally suppose that T has no eigenvalue. By Theorem 6.33, we know that there exists a T-invariant subspace $W \subset V$ such that $\dim W$ is either 1 or 2. If $\dim W = 1$ then $W = \text{span}(v)$ for some nonzero $v \in V$. But then we have $Tv \in W = \text{span}(v)$. So $Tv = \lambda v$ for some $\lambda \in \mathbb{R}$, i.e. λ is an eigenvalue of T, which is in contrary to our assumption. Hence we must have $\dim W = 2$. Let v, w be an orthonormal basis for W. Then we have $\|v\| = \|w\| = 1$, and $\langle v, w\rangle = 0$. Thus as T is orthogonal we get

$$\|Tv\| = \|Tw\| = 1, \qquad \text{and} \qquad \langle Tv, Tw\rangle = 0.$$

On the other hand we have $Tv, Tw \in W = \text{span}(v, w)$. Hence there are $a, b, c, d \in \mathbb{R}$ such that

$$Tv = av + bw, \qquad Tw = cv + dw.$$

Therefore the matrices of $R := T|_W$ and R^* in the orthonormal basis $\mathcal{C} := \{v, w\}$ are

$$[R]_\mathcal{C} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \implies [R^*]_\mathcal{C} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}^* = \begin{bmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Hence we have

$$R^*v = av + cw, \qquad R^*w = bv + dw.$$

But v, w is an orthonormal basis, so by Theorem 5.14 we have

$$\|Tv\|^2 = a^2 + b^2, \qquad \|R^*v\|^2 = a^2 + c^2, \qquad \langle Tv, Tw\rangle = ac + bd.$$

On the other hand, as we have seen above, the restrictions of an orthogonal operator are orthogonal. Thus R is orthogonal, and therefore R^* is also orthogonal. Hence we also have $\|R^*v\| = \|R^*w\| = 1$. Therefore we get

$$a^2 + b^2 = 1 = a^2 + c^2, \qquad ac + bd = 0. \tag{$*$}$$

Thus $b^2 = c^2$, so $c = \pm b$.

But if $c = b$ then $[R]_{\mathcal{C}}$ is symmetric, so R is self-adjoint. Hence it has an eigenvalue, and therefore contrary to our assumption T has an eigenvalue too, as shown in Exercise 4.6. Thus $c \neq b$. So we must have $c = -b$. Hence the second equation in $(*)$ implies that $0 = ac + bd = -ab + bd = b(-a + d)$. Now note that $b \neq 0$, since otherwise we would have $c = -b = 0 = b$, which leads to a contradiction as we saw. Therefore we must have $-a + d = 0$, i.e. $a = d$. So we have shown that

$$Tv = av + bw, \qquad Tw = -bv + aw.$$

Thus it only remains to show that

$$a = \cos\theta \quad \text{and} \quad b = \sin\theta \quad \text{for some} \quad \theta \in (0, \pi).$$

We know that the above relations hold for a unique $\theta \in [0, 2\pi)$, since $a^2 + b^2 = 1$. However we cannot have $\theta = 0$ or $\theta = \pi$, since $b \neq 0$. On the other hand, if $\theta \in (\pi, 2\pi)$ then $b = \sin\theta < 0$. But in this case we can simply replace v by $-v$. Then we have

$$\begin{aligned} T(-v) = -Tv = -av - bw = a(-v) + (-b)w, \\ Tw = -bv + aw = b(-v) + aw. \end{aligned}$$

In addition, it is easy to see that $-v, w$ is also an orthonormal basis for W. Hence $-v, w$ has our desired properties, since we have

$$a = \cos\theta = \cos(2\pi - \theta), \qquad -b = -\sin\theta = \sin(2\pi - \theta),$$

and $2\pi - \theta \in (0, \pi)$. Therefore the orthonormal basis $\mathcal{C} := \{\pm v, w\}$ has our desired properties.

The rest of the argument is similar to the case where T has an eigenvalue. If $W = V$ then the orthonormal basis $\mathcal{C}$ for V has the desired properties. If $W \neq V$ then we have $V = W \oplus W^\perp$, and therefore we get

$$0 < \dim W^\perp = \dim V - \dim W < \dim V,$$

since $0 < \dim W < \dim V$. We also know that by Theorem 6.30, $W^\perp$ is also T-invariant; because W is T-invariant, and T is normal. Now let $S := T|_{W^\perp}$. Then S is an orthogonal operator on the nonzero space $W^\perp$, as we have seen before. Hence by the induction hypothesis, $W^\perp$ has an orthonormal basis $\mathcal{B}$ with the prescribed properties in the theorem. Let $\mathcal{C}$ be the orthonormal basis for W that we constructed above. Then $\mathcal{B} \cup \mathcal{C}$ is a basis for V, which is also orthonormal due to Theorem 5.22. Furthermore, note that $\mathcal{B} \cup \mathcal{C}$ has the desired properties, since the action of T on an element of $\mathcal{B}$ is the same as the action of S on that element. ■

Remark. In the last paragraph of the above proof, we can show directly that $W^\perp$ is T-invariant. The proof is similar to the case where T has an eigenvalue, and uses the fact that $T|_W$ is onto.

Theorem 6.42. *Suppose $A \in F^{n\times n}$. Then the following statements are equivalent*

(i) *A is unitary.*
(ii) *The columns of A form an orthonormal basis for F^n.*
(iii) *The rows of A form an orthonormal basis for F^n.*

Proof. **(i) $\Longrightarrow$ (ii):** We know that $A^*A = I$. On the other hand we know that $(A^*A)_{ij} = A^*_{i,\cdot}A_{\cdot,j}$ for every i, j. We also know that $A^*_{i,\cdot} = (A_{\cdot,i})^*$ for every i. Hence we get

$$\langle A_{\cdot,j}, A_{\cdot,i}\rangle = (A_{\cdot,i})^*A_{\cdot,j} = A^*_{i,\cdot}A_{\cdot,j} = (A^*A)_{ij} = I_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Thus $A_{\cdot,1}, \ldots, A_{\cdot,n}$ is an orthonormal set in F^n, so they are linearly independent. Therefore they form a basis for F^n, since their number is the same as the dimension of F^n.

(ii) $\Longrightarrow$ (i): For every $i, j \leq n$ we have

$$(A^*A)_{ij} = A^*_{i,\cdot}A_{\cdot,j} = (A_{\cdot,i})^*A_{\cdot,j} = \langle A_{\cdot,j}, A_{\cdot,i}\rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Therefore $A^*A = I$. So $A^{-1} = A^*$, i.e. A is unitary.

(i) $\Longleftrightarrow$ (iii): The proofs are similar to the previous two parts. We only need to work with AA^* instead of A^*A. ■

Theorem 6.43. *Suppose $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$.*

(i) *If A is normal, then $\mathbb{C}^n$ has an orthonormal basis consisting of the eigenvectors of A. Furthermore, the matrix $C \in \mathbb{C}^{n\times n}$ whose columns are this basis of eigenvectors, is unitary; and C^*AC, which is equal to $C^{-1}AC$, is a diagonal matrix whose diagonal entries are the eigenvalues of A.*
(ii) *If B is symmetric, then $\mathbb{R}^n$ has an orthonormal basis consisting of the eigenvectors of B. Furthermore, the matrix $C \in \mathbb{R}^{n\times n}$ whose columns are this basis of eigenvectors, is orthogonal; and $C^\intercal BC$, which is equal to $C^{-1}BC$, is a diagonal matrix whose diagonal entries are the eigenvalues of B.*

Remark. Note that as explained in Theorem 4.31, the last sentence of the first part of the theorem means that all the eigenvalues of A appear on the diagonal of $C^{-1}AC$, and every diagonal entry of $C^{-1}AC$ is an eigenvalue of A. Similar remark applies to the second part of the theorem.

Proof. **(i)** Let $T \in \mathcal{L}(\mathbb{C}^n)$ be the operator that maps $z \in \mathbb{C}^n$ to $Tz := Az$. Then we have $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of $\mathbb{C}^n$. Therefore T is a normal operator, since A is normal, and $\mathcal{B}$ is an orthonormal basis. Hence we know that $\mathbb{C}^n$ has an orthonormal basis $\mathcal{C}$ consisting of the eigenvectors of T. But if $z \in \mathbb{C}^n$ is an eigenvector of T corresponding to the eigenvalue λ, then we have $Az = Tz = \lambda z$. Thus z is also an eigenvector of A corresponding to the eigenvalue λ. Therefore $\mathcal{C}$ is an orthonormal basis for $\mathbb{C}^n$ consisting of the eigenvectors of A.

Now suppose $\mathcal{C} = \{v_1, \ldots, v_n\}$, and let C be the matrix whose j-th column is $v_j \in \mathbb{C}^n$. By Theorem 4.31 we know that C is invertible, and $C^{-1}AC$ is a diagonal matrix whose diagonal entries are the eigenvalues of A. In addition we know that C is unitary, since its columns form an orthonormal basis for $\mathbb{C}^n$.

(ii) The proof is similar to the previous part. ∎

6.5 Polar Decomposition

Definition 6.44. An operator $T \in \mathcal{L}(V)$ is called **positive** if it is self-adjoint, and for all $v \in V$ we have

$$\langle Tv, v\rangle \geq 0.$$

Also, a square matrix $A \in F^{n\times n}$ is called **positive** if it is self-adjoint, and for all $x \in F^n$ we have

$$x^*Ax \geq 0.$$

Remark. Remember that when T is self-adjoint, $\langle Tv, v\rangle$ is always a real number. Also note that x^*Ax is a 1×1 matrix, i.e. it is a scalar. In addition we have $\bar{A}_{jk} = A_{kj}$. Hence

$$\overline{x^*Ax} = \overline{\sum_{j,k} \bar{x}_j A_{jk} x_k} = \sum_{j,k} x_j \bar{A}_{jk} \bar{x}_k = \sum_{k,j} \bar{x}_k A_{kj} x_j = x^*Ax.$$

Therefore x^*Ax is a real number for every x, when A is self-adjoint.

Remark. A positive operator, as we defined it, is sometimes called *positive semi-definite*. In contrast, a *positive definite* operator is a self-adjoint operator T that satisfies $\langle Tv, v\rangle > 0$ for all nonzero $v \in V$. Similarly we say that the self-adjoint operator T is *negative semi-definite*, or it is *negative definite*, if respectively $\langle Tv, v\rangle \leq 0$, or $\langle Tv, v\rangle < 0$, for all nonzero $v \in V$. These concepts can be also defined for self-adjoint matrices in the obvious way.

Theorem 6.45. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Then $T \in \mathcal{L}(V)$ is positive if and only if $[T]_{\mathcal{B}}$ is positive.*

Proof. Let $A := [T]_{\mathcal{B}}$. We know that T is self-adjoint if and only if A is self-adjoint. If A is positive, then by Theorem 5.14 for every $v \in V$ we have

$$\langle Tv, v\rangle = [v]_{\mathcal{B}}^*[Tv]_{\mathcal{B}} = [v]_{\mathcal{B}}^* A[v]_{\mathcal{B}} \geq 0.$$

So T is positive. Now suppose $\dim V = n$. Then for every $x \in F^n$ there is $v \in V$ such that $x = [v]_{\mathcal{B}}$. Hence if T is positive we have

$$x^* Ax = [v]_{\mathcal{B}}^* A[v]_{\mathcal{B}} = [v]_{\mathcal{B}}^*[Tv]_{\mathcal{B}} = \langle Tv, v\rangle \geq 0.$$

Thus A is positive. ■

***Example* 6.46.** Let $T \in \mathcal{L}(V)$. Then TT^* is a positive operator. Because we have $(TT^*)^* = T^{**}T^* = TT^*$, so TT^* is self-adjoint. In addition, for every $v \in V$ we have $\langle TT^*v, v\rangle = \langle T^*v, T^*v\rangle \geq 0$. Similarly we can show that T^*T is positive; and that for every matrix $A \in F^{n\times n}$, the matrices AA^* and A^*A are positive.

Theorem 6.47. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is self-adjoint. Then T is positive if and only if its eigenvalues are all nonnegative.*

Similarly, suppose $A \in F^{n\times n}$ is self-adjoint. Then A is positive if and only if its eigenvalues are all nonnegative.

Proof. We know that V has an orthonormal basis $\{v_1, \ldots, v_n\}$ consisting of the eigenvectors of T. Suppose $Tv_j = \lambda_j v_j$. We also know that each $\lambda_j \in \mathbb{R}$. If T is positive, then we have

$$0 \leq \langle Tv_j, v_j\rangle = \langle \lambda_j v_j, v_j\rangle = \lambda_j \|v_j\|^2 = \lambda_j,$$

as desired.

Now suppose $\lambda_j \geq 0$ for each j. Let $v \in V$. Then $v = \sum a_j v_j$ for some $a_j \in F$. Hence we have

$$Tv = \sum a_j Tv_j = \sum a_j \lambda_j v_j.$$

Therefore by Theorem 5.14 we have $\langle Tv, v\rangle = \sum a_j \lambda_j \bar{a}_j = \sum \lambda_j |a_j|^2 \geq 0$. So T is positive. The case of matrices is similar. ■

Remark. We have similar characterizations for positive definite, negative definite, or negative semi-definite operators and matrices. For example a self-adjoint operator or matrix is negative definite if and only if its eigenvalues are all negative. The proofs of these results are all similar to the above proof.

Theorem 6.48. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$ is positive. Then T has a unique positive* ***square root****, i.e. there is a unique positive operator $S \in \mathcal{L}(V)$ such that $S^2 = T$.*

Remark. In the following proof we will also show that the eigenvalues of S are the square roots of the eigenvalues of T.

Proof. We know that V has an orthonormal basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ consisting of the eigenvectors of T. Suppose $Tv_j = \lambda_j v_j$. We also know that each λ_j is a nonnegative real number. Let $S \in \mathcal{L}(V)$ be the unique operator that satisfies $Sv_j = \sqrt{\lambda_j} v_j$. Then we have

$$[S]_{\mathcal{B}} = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix}.$$

Hence S is self-adjoint, since its matrix with respect to an orthonormal basis is self-adjoint. In addition note that $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ are all the eigenvalues of S, because $[S]_{\mathcal{B}}$ is diagonal, and these are all the diagonal entries of $[S]_{\mathcal{B}}$. Thus all the eigenvalues of S are nonnegative, so S is positive. Finally note that we have

$$[S^2]_{\mathcal{B}} = \big([S]_{\mathcal{B}}\big)^2 = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix}^2 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = [T]_{\mathcal{B}}.$$

Therefore we must have $S^2 = T$, since the matrix of an operator uniquely determines that operator. Hence we have shown that T has a positive square root S. Note that we have also shown that the eigenvalues of S are the square roots of the eigenvalues of T.

Now we need to prove that the positive square root of T is unique. Suppose $R \in \mathcal{L}(V)$ is a positive square root of T too. Then we know that R is diagonalizable, since it is self-adjoint. Let $\mu_1, \ldots, \mu_k$ be all the distinct eigenvalues of R. We know that $\mu_1, \ldots, \mu_k$ are nonnegative real numbers. We also have

$$V = E_{\mu_1}(R) \oplus \cdots \oplus E_{\mu_k}(R).$$

Let $v \in E_{\mu_j}(R)$ for some j. Then we have

$$Tv = R^2 v = R(Rv) = R(\mu_j v) = \mu_j Rv = \mu_j(\mu_j v) = \mu_j^2 v.$$

Since v can be nonzero, μ_j^2 is an eigenvalue of T. Thus we have shown that

$$E_{\mu_j}(R) \subset E_{\mu_j^2}(T).$$

Furthermore for $i \neq j$ we have $\mu_i^2 \neq \mu_j^2$, because $\mu_i \neq \mu_j$ and $\mu_i, \mu_j \geq 0$. Hence $\mu_1^2, \ldots, \mu_k^2$ are distinct eigenvalues of T. Now the subspace generated by

$\bigcup_{j\leq k} E_{\mu_j}(R)$, i.e. the sum of $E_{\mu_j}(R)$'s, is certainly contained in the subspace generated by $\bigcup_{j\leq k} E_{\mu_j^2}(T)$. Thus we have

$$V = E_{\mu_1}(R) \oplus \cdots \oplus E_{\mu_k}(R) \subset E_{\mu_1^2}(T) \oplus \cdots \oplus E_{\mu_k^2}(T) \subset V.$$

Therefore $E_{\mu_1^2}(T) \oplus \cdots \oplus E_{\mu_k^2}(T) = V$. So $\mu_1^2, \ldots, \mu_k^2$ are all the distinct eigenvalues of T. Because if T had any other eigenvalue λ with corresponding eigenvector v, then v would have been a linear combination of some eigenvectors of T corresponding to $\mu_1^2, \ldots, \mu_k^2$, which contradicts the fact that eigenvectors corresponding to distinct eigenvalues are linearly independent.

In addition we must have $E_{\mu_j}(R) = E_{\mu_j^2}(T)$ for every j. Because if this relation fails for some i, then we would have $\dim E_{\mu_i}(R) < \dim E_{\mu_i^2}(T)$. Furthermore for $j \neq i$ we have $\dim E_{\mu_j}(R) \leq \dim E_{\mu_j^2}(T)$. Hence we get

$$\dim V = \sum_{j\leq k} \dim E_{\mu_j}(R) < \sum_{j\leq k} \dim E_{\mu_j^2}(T) = \dim V,$$

which is a contradiction. Now note that the eigenvalues of S are the square roots of the eigenvalues of T. So $\mu_1, \ldots, \mu_k$ are all the distinct eigenvalues of S. Note that we have $\sqrt{\mu_j^2} = \mu_j$, since $\mu_j \geq 0$. Therefore if we repeat the above argument with S in place of R, we get

$$E_{\mu_j}(S) = E_{\mu_j^2}(T) = E_{\mu_j}(R).$$

Now for any $v \in V$ there are $v_j \in E_{\mu_j}(R) = E_{\mu_j}(S)$ such that $v = v_1 + \cdots + v_k$. Thus we have

$$Sv = Sv_1 + \cdots + Sv_k = \mu_1 v_1 + \cdots + \mu_k v_k = Rv_1 + \cdots + Rv_k = Rv.$$

Hence $R = S$ as desired. ■

Theorem 6.49. *Suppose F is either $\mathbb{R}$ or $\mathbb{C}$, and $A \in F^{n\times n}$ is a positive matrix. Then A has a unique positive* ***square root****, i.e. there is a unique positive matrix $C \in F^{n\times n}$ such that $C^2 = A$.*

Proof. Let $T \in \mathcal{L}(F^n)$ be defined by $T(x) = Ax$ for $x \in F^n$. Then $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of F^n. Now we know that T is positive, since A is positive and the standard basis is orthonormal. Thus by the previous theorem there is a unique positive operator $S \in \mathcal{L}(F^n)$ such that $S^2 = T$. Let $C := [S]_{\mathcal{B}}$. Then C is a positive matrix, and we have

$$C^2 = \big([S]_{\mathcal{B}}\big)^2 = [S^2]_{\mathcal{B}} = [T]_{\mathcal{B}} = A.$$

Hence A has a positive square root.

Now suppose $B \in F^{n\times n}$ is also a positive square root of A. Let $R \in \mathcal{L}(F^n)$ be defined by $R(x) = Bx$ for $x \in F^n$. Then we have $[R]_{\mathcal{B}} = B$, so R is a positive operator too. We also have

$$[R^2]_{\mathcal{B}} = \big([R]_{\mathcal{B}}\big)^2 = B^2 = A = [T]_{\mathcal{B}}.$$

Therefore $R^2 = T$. Thus $R = S$, since the positive square root of T is unique. Hence we have $B = [S]_{\mathcal{B}} = [R]_{\mathcal{B}} = C$ as desired. ■

Remark. Note that if an operator T has a self-adjoint square root S, then T must be positive. Because $T^* = (SS)^* = S^*S^* = SS = T$, so T is self-adjoint. Also, for every vector v we have

$$\langle Tv, v\rangle = \langle SSv, v\rangle = \langle Sv, Sv\rangle = \|Sv\|^2 \geq 0.$$

Similarly if a matrix $A \in F^{n\times n}$ has a self-adjoint square root, then it must be positive.

***Exercise* 6.50.** Suppose $A, B \in F^{n\times n}$ are positive matrices, and A is invertible. Let $C \in F^{n\times n}$ be the positive square root of A. Show that CBC is a positive matrix, and AB is similar to CBC.

Remark. As a result, the eigenvalues of AB are nonnegative real numbers, since they are the same as the eigenvalues of the positive matrix CBC.

Solution. First note that $(CBC)^* = C^*B^*C^* = CBC$. Now for every $x \in F^n$ let $y = Cx$. Then we have

$$x^*(CBC)x = (x^*C)B(Cx) = (x^*C^*)B(Cx) = (Cx)^*B(Cx) = y^*By \geq 0.$$

Thus CBC is positive.

On the other hand, note that C is also invertible. Because otherwise the linear system $Cx = 0$ has at least two solutions in F^n, due to Theorem 3.49. But then for any solution of $Cx = 0$ we would have $Ax = C^2x = C(Cx) = C0 = 0$, i.e. the linear system $Ax = 0$ has more than one solution too. However, by Theorem 3.49, this is in contradiction with the invertibility of A.

Finally note that we have

$$AB = C^2B = C^2BI = C^2BCC^{-1} = C(CBC)C^{-1},$$

i.e. AB and CBC are similar matrices. ■

Polar Decomposition. *Suppose V is finite dimensional, and $T \in \mathcal{L}(V)$. Then there is a positive operator $P \in \mathcal{L}(V)$, and a unitary operator $U \in \mathcal{L}(V)$, such that*

$$T = PU.$$

Remark. The significance of this theorem is that it gives us a complete description of the action of an arbitrary linear map T. Because as we discussed in the remarks before and after Theorem 6.41, we have a clear geometric understanding of the action of a unitary operator. On the other hand, a positive operator can be diagonalized, since it is self-adjoint. In addition, its eigenvalues are nonnegative real numbers, and there is an orthonormal basis of its eigenvectors. Hence the action of a positive operator on a vector is that it scales each coordinate of that vector by a nonnegative scale, when we represent the vector in the orthonormal basis of the eigenvectors of the positive operator. Therefore the meaning of the above theorem is that every linear operator on a finite dimensional inner product space is the composition of several reflections, several rotations, and several scalings.

It should be noted that we can not necessarily diagonalize both P, U simultaneously. Therefore the directions of scalings are not necessarily the same as the directions of rotations.

Proof. First note that if such operators P, U exist, then we must have

$$TT^* = PU(PU)^* = PUU^*P^* = PUU^{-1}P = P^2.$$

Now note that TT^* is a positive operator. Therefore TT^* has a unique positive square root, which we call P.

To find U, let us first suppose that T is invertible. Then T^* is also invertible. Thus P is invertible too, since otherwise there is a nonzero vector v such that $Pv = 0$. But then we have $TT^*v = P^2v = 0$, which contradicts the fact that TT^* is invertible. Hence P is invertible, and therefore the only candidate for U is $P^{-1}T$. Now for $U := P^{-1}T$ we have

$$\begin{aligned} U^*U = T^*(P^{-1})^*P^{-1}T = T^*(P^*)^{-1}P^{-1}T = T^*P^{-1}P^{-1}T \\ = T^*(P^2)^{-1}T = T^*(TT^*)^{-1}T = T^*(T^*)^{-1}T^{-1}T = I_V. \end{aligned}$$

So $U^{-1} = U^*$ as desired, since V is finite dimensional.

Next let us prove the theorem for an arbitrary linear map T. In this case P is not necessarily invertible. So in order to define U, we have to find a replacement for P^{-1}. First note that for every $v \in V$ we have

$$\|T^*v\|^2 = \langle T^*v, T^*v\rangle = \langle v, TT^*v\rangle = \langle v, P^2v\rangle = \langle Pv, Pv\rangle = \|Pv\|^2. \qquad (*)$$

Hence in particular we have $T^*v = 0$ if and only if $Pv = 0$. In other words $\operatorname{null} T^* = \operatorname{null} P$. Thus we have

$$W := T(V) = (\operatorname{null} T^*)^\perp = (\operatorname{null} P)^\perp = P^*(V) = P(V).$$

As a result, T, P have the same rank. Furthermore, W is P-invariant, and $P|_W$ is invertible, since $W \cap \operatorname{null} P = \{0\}$.

Now note that if a unitary operator U exists such that $T = PU$, then for $v \in \operatorname{null} T$ we have $PUv = Tv = 0$, i.e. $Uv \in \operatorname{null} P$. So U maps $\operatorname{null} T$ into $\operatorname{null} P$. Since U preserves the inner product, it must map $(\operatorname{null} T)^\perp = T^*(V)$ into $(\operatorname{null} P)^\perp = W$. But for $v \in T^*(V)$ we have $Tv \in W$, so we have a natural candidate for Uv, namely $(P|_W)^{-1}Tv$. On the other hand when $v \in \operatorname{null} T$, the value of Uv is not important, as long as we have $Uv \in \operatorname{null} P$. With these criteria in mind, we are going to construct the operator U.

Let $\{v_1, \dots, v_k\}$, $\{u_1, \dots, u_k\}$ be orthonormal bases for $\operatorname{null} T$, $\operatorname{null} P$ respectively. Note that these two subspaces have the same dimension, since T, P have the same rank. Let $R \in \mathcal{L}(\operatorname{null} T, \operatorname{null} P)$ be the linear map that sends v_j to u_j. Then R preserves the norm. Because for $v = \sum a_j v_j$ we have $Rv = \sum a_j u_j$. Hence we have $\|Rv\|^2 = \sum |a_j|^2 = \|v\|^2$, since the two bases are orthonormal.

Now any $v \in V$ can be written uniquely as $w + u$, where $w \in T^*(V)$ and $u \in \operatorname{null} T$, because $V = T^*(V) \oplus \operatorname{null} T$. Let $S := (P|_W)^{-1} \in \mathcal{L}(W)$. Then we define

$$Uv := STw + Ru.$$

Note that U is well defined, since the decomposition of v into $w+u$ is unique. Also note that U is a linear map, as can be easily checked from the definition. Now we have

$$PUv = P(STw + Ru) = PSTw + PRu = I_W Tw = Tw = Tw + Tu = Tv.$$

Note that we have used the facts that $Ru \in \operatorname{null} P$, and $u \in \operatorname{null} T$.

To finish the proof, we only need to show that U is unitary. It suffices to show that U preserves the norm. For $w \in T^*(V)$ we have $w = T^*\tilde{w}$ for some $\tilde{w} \in V$. Hence by $(*)$ we have

$$\|STw\| = \|STT^*\tilde{w}\| = \|SP^2\tilde{w}\| = \|P\tilde{w}\| = \|T^*\tilde{w}\| = \|w\|.$$

We also know that $\|Ru\| = \|u\|$ for $u \in \operatorname{null} T$. Therefore as $Ru \in \operatorname{null} P$ and $STw \in W$ are orthogonal, and also $u \in \operatorname{null} T$ and $w \in T^*(V)$ are orthogonal, we have

$$\|Uv\|^2 = \|STw\|^2 + \|Ru\|^2 = \|w\|^2 + \|u\|^2 = \|v\|^2,$$

as desired. ∎

Remark. As we have seen in the above proof, P is uniquely determined by T, since it is the unique positive square root of TT^*. Also when T is invertible, $U = P^{-1}T$ is uniquely determined by T too. But when T is not invertible, then U is not uniquely determined. The reason is the freedom that we have for the construction of the operator R in the proof.

Remark. We can also show that for every operator T on a finite dimensional inner product space, there is a positive operator P_1 and a unitary operator U_1 such that

$$T = U_1 P_1.$$

But P_1, U_1 are not necessarily the same as P, U. To prove this version we can apply the above theorem to T^* to obtain $T^* = P_2 U_2$. Then we have

$$T = T^{**} = U_2^* P_2^* = U_2^{-1} P_2.$$

Hence we can take $P_1 = P_2$ and $U_1 = U_2^{-1}$. Note that the inverse of a unitary map is also unitary. This construction also shows that P_1 is the positive square root of $T^*T^{**} = T^*T$.

Definition 6.51. Suppose $T \in \mathcal{L}(V)$. The **singular values** of T are the square roots of the eigenvalues of the positive operator TT^*.

Similarly, the **singular values** of a square matrix $A \in F^{n \times n}$ are the square roots of the eigenvalues of the positive matrix AA^*.

Remark. Note that the singular values of an operator or a matrix are nonnegative real numbers.

Remark. It can be shown that the eigenvalues of TT^* are the same as the eigenvalues of T^*T, although T, T^* do not commute necessarily. So we could have defined the singular values of T using the eigenvalues of T^*T. A similar remark is true about the square matrices.

Remark. Note that if $T = PU$ is the polar decomposition of T, then the singular values of T are the eigenvalues of P. Because we know that P is the unique positive square root of TT^*. So the eigenvalues of P are the square roots of the eigenvalues of TT^*, due to the remark after Theorem 6.48.

Theorem 6.52. *Suppose V is finite dimensional, and $\mathcal{B}$ is an orthonormal basis for V. Then the singular values of $T \in \mathcal{L}(V)$ are the same as the singular values of $[T]_{\mathcal{B}}$.*

Proof. Let $A := [T]_{\mathcal{B}}$. Then $A^* = [T^*]_{\mathcal{B}}$, since $\mathcal{B}$ is orthonormal. Hence we have

$$AA^* = [T]_{\mathcal{B}}[T^*]_{\mathcal{B}} = [TT^*]_{\mathcal{B}}.$$

Therefore the eigenvalues of TT^* are the same as the eigenvalues of AA^*. Hence their square roots are also the same, i.e. the singular values of T are the same as the singular values of A. ■

Singular Value Decomposition. *Suppose F is either $\mathbb{R}$ or $\mathbb{C}$, and $A \in F^{n\times n}$. Then there is a diagonal matrix $\Sigma \in \mathbb{R}^{n\times n}$ whose diagonal entries are the singular values of A, and there are unitary matrices $U_1, U_2 \in F^{n\times n}$, such that*

$$A = U_1 \Sigma U_2^*.$$

Remark. In general U_1, U_2 are not equal.

Proof. Let $T \in \mathcal{L}(F^n)$ be defined by $T(x) = Ax$ where $x \in F^n$. Then $[T]_{\mathcal{B}} = A$, where $\mathcal{B}$ is the standard basis of F^n. Now we know that $T = PU$, where P is a positive operator, and U is a unitary operator. Let $B := [P]_{\mathcal{B}}$ and $M := [U]_{\mathcal{B}}$. Then B is a positive matrix, and M is a unitary matrix, since the standard basis is orthonormal. By Theorem 6.43 we know that there is a unitary matrix $C \in F^{n\times n}$ such that $\Sigma := CBC^*$ is diagonal. But the diagonal entries of Σ are the eigenvalues of B, which are the eigenvalues of P. On the other hand, the eigenvalues of P are the singular values of T, which are the same as the singular values of A. Finally we have

$$A = [T]_{\mathcal{B}} = [PU]_{\mathcal{B}} = [P]_{\mathcal{B}}[U]_{\mathcal{B}} = BM = C^*\Sigma CM = U_1 \Sigma U_2^*,$$

where $U_1 := C^*$, and $U_2 := (CM)^*$. Note that U_1, U_2 are unitary matrices because $U_1^* = C = (C^*)^{-1} = U_1^{-1}$, and

$$U_2 = (CM)^* = M^*C^* = M^{-1}C^{-1} = (CM)^{-1} = (U_2^*)^{-1}.$$ ■

Definition 6.53. Let $A \in F^{n\times n}$. The **operator norm** of A is

$$\|A\| := \sup_{x \in F^n - \{0\}} \frac{\|Ax\|}{\|x\|}.$$

Remark. A nice property of the operator norm is that for every $x \in F^n$ we have

$$\|Ax\| \leq \|A\|\|x\|.$$

This property makes the operator norm a very useful tool in analysis, when we want to estimate the norm of the image of a vector.

Remark. It is not hard to show that the operator norm is a norm on the space of matrices. Although it is not induced by any inner product. In the next theorem we compute the operator norm of a matrix using its singular value decomposition.

Theorem 6.54. *Let $A \in F^{n\times n}$, and suppose $s_1, \ldots, s_n \in [0, \infty)$ are the singular values of A. Then we have*

$$\|A\| = \max\{s_1, \ldots, s_n\}.$$

Proof. Suppose $A = U_1 \Sigma U_2^*$ is the singular value decomposition of A. Let $x \in F^n$. Then for $y := U_2^* x$ we have $x = U_2 y$, and so $\|y\| = \|x\|$, since U_2 is unitary. Suppose $s_1, \ldots, s_n$ are arranged so that

$$\Sigma = \begin{bmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_n \end{bmatrix}.$$

Then $\Sigma y = [s_1 y_1, \ldots, s_n y_n]^\intercal$. Therefore $\|\Sigma y\| \leq s\|y\|$, where $s := \max\{s_1, \ldots, s_n\}$. Hence as U_1 is unitary we have

$$\|Ax\| = \|U_1 \Sigma y\| = \|\Sigma y\| \leq s\|y\| = s\|x\|.$$

Thus $\|A\| \leq s$. On the other hand suppose $s = s_j$. Then for $x = U_2 e_j$ we have

$$\begin{aligned} \|Ax\| = \|U_1 \Sigma U_2^* U_2 e_j\| = \|U_1 \Sigma e_j\| &= \|\Sigma e_j\| \\ &= \|s_j e_j\| = s\|e_j\| = s\|U_2 e_j\| = s\|x\|. \end{aligned}$$

So $s \leq \|A\|$ too. ∎

Chapter 7

Determinants

7.1 Multilinear Maps

Definition 7.1. Let $V_1, \dots, V_k, W$ be vector spaces over a field F. A map $L : V_1 \times \cdots \times V_k \to W$ is called **multilinear** or ***k*-linear** if L is linear with respect to each variable when the other variables are fixed, i.e. for any $j \le k$ we have

$$\begin{aligned} &L(v_1, \dots, v_{j-1}, au + bv, v_{j+1}, \dots, v_k) \\ &\quad = a\,L(v_1, \dots, v_{j-1}, u, v_{j+1}, \dots, v_k) + b\,L(v_1, \dots, v_{j-1}, v, v_{j+1}, \dots, v_k), \end{aligned}$$

for every vectors $v_i \in V_i$, $u, v \in V_j$, and scalars $a, b \in F$.

Remark. We are mostly interested in the case where $V_1 = \cdots = V_k = F^n$, and $W = F$. So in this section, the theorems are only stated for this case. Although, the general case can be treated similarly.

Theorem 7.2. *Suppose $L : (F^n)^k \to F$ is multilinear, and $e_1, \dots, e_n$ is the standard basis of F^n. Also suppose $v_i = [a_{i1}, \dots, a_{in}]^\intercal = \sum_{j \le n} a_{ij} e_j$ for $i \le k$. Then we have*

$$L(v_1, \dots, v_k) = \sum_{j_1 \le n} \cdots \sum_{j_k \le n} a_{1j_1} \cdots a_{kj_k} L(e_{j_1}, \dots, e_{j_k}).$$

Proof. We have

$$\begin{aligned}
L(v_1,\dots,v_k) &= L\Big(\sum_{j_1\le n} a_{1j_1}e_{j_1}, v_2,\dots,v_k\Big) = \sum_{j_1\le n} a_{1j_1}L(e_{j_1},v_2,\dots,v_k)\\
&= \sum_{j_1\le n} a_{1j_1}L\Big(e_{j_1},\sum_{j_2\le n} a_{2j_2}e_{j_2},v_3,\dots,v_k\Big)\\
&= \sum_{j_1\le n} a_{1j_1}\Big(\sum_{j_2\le n} a_{2j_2}L(e_{j_1},e_{j_2},v_3,\dots,v_k)\Big)\\
&= \sum_{j_1\le n}\sum_{j_2\le n} a_{1j_1}a_{2j_2}L(e_{j_1},e_{j_2},v_3,\dots,v_k)\\
&\ \ \vdots\\
&= \sum_{j_1\le n}\cdots\sum_{j_k\le n} a_{1j_1}\cdots a_{kj_k}L(e_{j_1},\dots,e_{j_k}).
\end{aligned}$$

■

Theorem 7.3. *Suppose $c_{j_1\dots j_k}\in F$ are given. Then there exists a unique multilinear map $L:(F^n)^k\to F$ such that*

$$L(e_{j_1},\dots,e_{j_k}) = c_{j_1\dots j_k},$$

for every $j_1,\dots,j_k\in\{1,\dots,n\}$. Moreover, for $v_i=[a_{i1},\dots,a_{in}]^\intercal$ we have

$$L(v_1,\dots,v_k) = \sum_{j_1\le n}\cdots\sum_{j_k\le n} a_{1j_1}\cdots a_{kj_k}c_{j_1\dots j_k}.$$

7.2 Determinants

7.3 The Characteristic Polynomial

Chapter 8

The Jordan Form

8.1 Generalized Eigenvectors

Notation. In this chapter we assume that F is a field, V is a nonzero vector space over F, and $T \in \mathcal{L}(V)$ is a linear operator.

Remember that a diagonalizable linear operator can be represented by a diagonal matrix. So we can easily understand the behavior of a diagonalizable operator. In addition, calculations with diagonal matrices are much simpler than calculations with arbitrary matrices. But we know that in general a linear operator is not necessarily diagonalizable (see for instance Example 4.28). Hence in order to understand the behavior of an arbitrary operator T, and to simplify calculations involving T, we will try to find a basis $\mathcal{B}$ so that $[T]_{\mathcal{B}}$ has a simple "canonical" form. We will see that these matrices in canonical form are not in general diagonal, but they have a simple structure that is very close to a diagonal matrix.

To obtain a canonical form, first we will show that if we decompose the space as the direct sum of several T-invariant subspaces, then the matrix of T becomes a block diagonal matrix, as we define below. The next step is to find an appropriate decomposition for the space. Finally we will study the restriction of T to those invariant subspaces, and we will find a suitable description for it.

Definition 8.1. Suppose $n_1, \dots, n_k, m_1, \dots, m_l \in \mathbb{N}$. Let $n = n_1 + \cdots + n_k$ and $m = m_1 + \cdots + m_l$. A matrix $A \in F^{n \times m}$ is called a **block matrix** if for $\alpha \leq k$ and $\beta \leq l$ there exist matrices $B_{\alpha\beta} \in F^{n_\alpha \times m_\beta}$, called the **blocks** of A, such that for $r \leq n_\alpha, s \leq m_\beta$ we have

$$A_{ij} = (B_{\alpha\beta})_{rs}, \text{ where } i = r + \sum_{\tilde{\alpha}<\alpha} n_{\tilde{\alpha}},\ j = s + \sum_{\tilde{\beta}<\beta} m_{\tilde{\beta}}.$$

Suppose in addition that $m = n$, $l = k$, and $m_\alpha = n_\alpha$ for each α. Then we say A

is a **block diagonal matrix** if $B_{\alpha\beta} = 0$ when $\beta \neq \alpha$. In this case, the blocks $B_{\alpha\alpha}$ are called the *diagonal blocks* of A.

Notation. When A is a block diagonal matrix with diagonal blocks $A_1, \dots, A_k$, we write

$$A = A_1 \oplus \cdots \oplus A_k.$$

Note that in this notation the order of $A_1, \dots, A_k$ is important. Also, note that each diagonal block is itself a square matrix.

Remark. It is easy to check that in a block matrix, for every $i \leq n, j \leq m$ there are unique $\alpha \leq k, \beta \leq l$ and $r \leq n_\alpha, s \leq m_\beta$ such that $i = r + \sum_{\tilde{\alpha}<\alpha} n_{\tilde{\alpha}}$ and $j = s + \sum_{\tilde{\beta}<\beta} m_{\tilde{\beta}}$. Note that $\sum_{\tilde{\alpha}<\alpha} n_{\tilde{\alpha}}$ is the number of rows of A above the block $B_{\alpha\beta}$, and $\sum_{\tilde{\beta}<\beta} m_{\tilde{\beta}}$ is the number of columns of A to the left of the block $B_{\alpha\beta}$.

Remark. Note that a block matrix is itself a matrix. So when we talk about a block matrix, we are considering a matrix, but we want to think that its entries are partitioned into several smaller matrices, which we call them the blocks of our block matrix. Also note that every matrix can be considered as a block matrix; and if the matrix is not 1×1, then there are several different ways to consider it as a block matrix.

Remark. A block matrix and a block diagonal matrix look respectively as follows

$$\begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1l} \\ B_{21} & B_{22} & \dots & B_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \cdots & B_{kl} \end{bmatrix}, \qquad \begin{bmatrix} B_{11} & 0 & \cdots & 0 \\ 0 & B_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{kk} \end{bmatrix},$$

where $B_{\alpha\beta} \in F^{n_\alpha \times m_\beta}$.

***Example* 8.2.** The following matrices are block diagonal

$$\begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & -1 & 0 & 0 \\ 0 & 0 & 5 & 7 \\ 0 & 0 & 0 & -6 \end{bmatrix}, \qquad \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 9 & 0 & 8 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

The first one has two 2×2 nonzero blocks, and the second one has a 1×1 nonzero block and a 3×3 nonzero block. To emphasize their block diagonal structure, we can also write the above matrices as follows

$$\begin{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix} & 0 & 0 \\ 0 & 0 & \begin{bmatrix} 5 & 7 \\ 0 & -6 \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \begin{bmatrix} 3 \end{bmatrix} & 0 & 0 & 0 \\ 0 & \begin{bmatrix} 9 & 0 & 8 \\ 5 & 6 & 7 \\ 0 & 1 & 2 \end{bmatrix} \end{bmatrix}.$$

***Exercise* 8.3.** Suppose $n_1, \dots, n_k, m_1, \dots, m_l \in \mathbb{N}$. Let $n = n_1 + \cdots + n_k$ and $m = m_1 + \cdots + m_l$. Let $A \in F^{n\times n}$, $B \in F^{m\times m}$, $A_\alpha \in F^{n_\alpha \times n_\alpha}$ for $\alpha \le k$, and $B_\beta \in F^{m_\beta \times m_\beta}$ for $\beta \le l$. Suppose that $A = A_1 \oplus \cdots \oplus A_k$ and $B = B_1 \oplus \cdots \oplus B_l$. Then we have

$$A \oplus B = A_1 \oplus \cdots \oplus A_k \oplus B_1 \oplus \cdots \oplus B_l.$$

Solution. Let $C := A \oplus B$. Then C has 4 blocks, where two of them are A, B, and the other two are zero. Hence for $i, j \le m + n$ we have

$$C_{ij} = \begin{cases} A_{ij} & \text{if } i, j \le n, \\ B_{i-n,j-n} & \text{if } i, j > n, \\ 0 & \text{otherwise.} \end{cases}$$

Now let $D := A_1 \oplus \cdots \oplus A_k \oplus B_1 \oplus \cdots \oplus B_l$. Then D has $(k + l)^2$ blocks. Let $A_{\alpha\hat{\alpha}} \in F^{n_\alpha \times n_{\hat{\alpha}}}$ and $B_{\beta\hat{\beta}} \in F^{m_\beta \times m_{\hat{\beta}}}$ be the blocks of A and B respectively. Note that $\alpha, \hat{\alpha} \le k$ and $\beta, \hat{\beta} \le l$. Also note that

$$A_{\alpha\hat{\alpha}} = \begin{cases} A_\alpha & \text{if } \hat{\alpha} = \alpha, \\ 0 & \text{if } \hat{\alpha} \ne \alpha, \end{cases} \qquad B_{\beta\hat{\beta}} = \begin{cases} B_\beta & \text{if } \hat{\beta} = \beta, \\ 0 & \text{if } \hat{\beta} \ne \beta. \end{cases}$$

Let $D_{\gamma\hat{\gamma}}$ be the blocks of D, where $\gamma, \hat{\gamma} \le k + l$. Then we have

$$D_{\gamma\hat{\gamma}} = \begin{cases} A_\gamma = A_{\gamma\hat{\gamma}} & \text{if } \gamma = \hat{\gamma} \le k, \\ B_{\gamma-k} = B_{\gamma-k,\hat{\gamma}-k} & \text{if } \gamma = \hat{\gamma} > k, \\ 0 = A_{\gamma\hat{\gamma}} & \text{if } \gamma, \hat{\gamma} \le k, \gamma \ne \hat{\gamma}, \\ 0 = B_{\gamma-k,\hat{\gamma}-k} & \text{if } \gamma, \hat{\gamma} > k, \gamma \ne \hat{\gamma}, \\ 0 \in F^{n_\gamma \times m_{\hat{\gamma}-k}} & \text{if } \gamma \le k, \hat{\gamma} > k, \\ 0 \in F^{m_{\gamma-k} \times n_{\hat{\gamma}}} & \text{if } \gamma > k, \hat{\gamma} \le k. \end{cases}$$

Also, note that C, D have the same size.

To simplify the notation, for every $\alpha \le k$ let $N_\alpha := \sum_{\tilde{\alpha}<\alpha} n_{\tilde{\alpha}}$, and for every $\beta \le l$ let $M_\beta := \sum_{\tilde{\beta}<\beta} m_{\tilde{\beta}}$. Suppose $i, j \le n$. Then for some $\alpha, \hat{\alpha} \le k$ we have $N_\alpha < i \le N_{\alpha+1}$ and $N_{\hat{\alpha}} < j \le N_{\hat{\alpha}+1}$. Hence we have

$$C_{ij} = A_{ij} = (A_{\alpha\hat{\alpha}})_{rs} = (D_{\alpha\hat{\alpha}})_{rs},$$

where $r := i - N_\alpha$ and $s := j - N_{\hat{\alpha}}$. Next suppose $n < i, j \le n + m$. Then for some $\beta, \hat{\beta} \le l$ we have $M_\beta < i - n \le M_{\beta+1}$ and $M_{\hat{\beta}} < j - n \le M_{\hat{\beta}+1}$. Hence we have

$$C_{ij} = B_{i-n,j-n} = (B_{\beta\hat{\beta}})_{rs} = (D_{\beta+k,\hat{\beta}+k})_{rs},$$

where $r := i - n - M_\beta$ and $s := j - n - M_{\hat{\beta}}$. Note that $n + M_\beta$ is the number of rows of D above the block $D_{\beta+k,\hat{\beta}+k}$, and $n + M_{\hat{\beta}}$ is the number of columns of D to the left of the block $D_{\beta+k,\hat{\beta}+k}$.

Now suppose that $i \leq n$ and $n < j \leq n + m$. Then for some $\alpha \leq k$ and $\beta \leq l$ we have $N_\alpha < i \leq N_{\alpha+1}$ and $M_\beta < j - n \leq M_{\beta+1}$. Hence we have

$$C_{ij} = 0 = (D_{\alpha,\beta+k})_{rs},$$

where $r := i - N_\alpha$ and $s := j - n - M_\beta$. Finally suppose that $n < i \leq n + m$ and $j \leq n$. Then for some $\alpha \leq k$ and $\beta \leq l$ we have $M_\beta < i - n \leq M_{\beta+1}$ and $N_\alpha < j \leq N_{\alpha+1}$. Hence we have

$$C_{ij} = 0 = (D_{\beta+k,\alpha})_{rs},$$

where $r := i - n - M_\beta$ and $s := j - N_\alpha$. Therefore by the definition of a block matrix we have $C = D$, as desired. ■

Theorem 8.4. *Suppose V is finite dimensional, and $W_1, \ldots, W_k$ are T-invariant nonzero subspaces of V such that*

$$V = W_1 \oplus \cdots \oplus W_k.$$

Let $\mathcal{B}_\alpha$ be a basis for W_α for each $\alpha \leq k$. Then the matrix of T with respect to the basis $\mathcal{B} = \bigcup_{\alpha=1}^{k} \mathcal{B}_\alpha$ for V, is a block diagonal matrix of the form

$$[T]_{\mathcal{B}} = \begin{bmatrix} [T|_{W_1}]_{\mathcal{B}_1} & 0 & \cdots & 0 \\ 0 & [T|_{W_2}]_{\mathcal{B}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [T|_{W_k}]_{\mathcal{B}_k} \end{bmatrix}.$$

In other words we have $[T]_{\mathcal{B}} = [T|_{W_1}]_{\mathcal{B}_1} \oplus \cdots \oplus [T|_{W_k}]_{\mathcal{B}_k}$.

Remark. Note that in the list of vectors of the basis $\mathcal{B}$, we first put the elements of $\mathcal{B}_1$, then the elements of $\mathcal{B}_2$, then the elements of $\mathcal{B}_3$, and so forth.

Proof. Let $n = \dim V$ and $n_\alpha = \dim W_\alpha$ for each $\alpha \leq k$. Note that by Theorem 2.55 we have $n = n_1 + \cdots + n_k$, and $\mathcal{B} = \bigcup_{\alpha=1}^{k} \mathcal{B}_\alpha$ is a basis for V. Now suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$. Let us denote the block diagonal matrix described in the theorem by A. We will show that the j-th column of $[T]_{\mathcal{B}}$ is equal to the j-th column of A, for every $j \leq n$. We know that the j-th column of $[T]_{\mathcal{B}}$ is $[Tv_j]_{\mathcal{B}}$. Suppose $Tv_j = \sum_{l \leq n} a_l v_l$, where $a_l \in F$. Also suppose that $v_j \in \mathcal{B}_\beta$ for some β, and we have $\mathcal{B}_\beta = \{v_{j_1}, \ldots, v_j, \ldots, v_{j_2}\}$. Then we know that

$$\big([T]_{\mathcal{B}}\big)_{\cdot,j} = [Tv_j]_{\mathcal{B}} = [a_1, \ldots, a_{j_1}, \ldots, a_{j_2}, \ldots, a_n]^{\intercal} \in F^n.$$

On the other hand we have $v_j \in W_\beta$, so we must have $Tv_j \in W_\beta$, since W_β is T-invariant. Therefore Tv_j is a linear combination of $v_{j_1}, \dots, v_{j_2}$, i.e. $Tv_j = \sum_{j_1 \le l \le j_2} b_l v_l$ for some $b_l \in F$. Now the equality

$$\sum_{l \le n} a_l v_l = Tv_j = \sum_{j_1 \le l \le j_2} b_l v_l$$

implies that $a_l = 0$ for $l < j_1$ and $l > j_2$, and $b_l = a_l$ for $j_1 \le l \le j_2$. Because when we write a vector as a linear combination of elements of a basis, the coefficients are uniquely determined. Hence we have

$$\big([T]_{\mathcal{B}}\big)_{\cdot,j} = [Tv_j]_{\mathcal{B}} = [0, \dots, 0, a_{j_1}, \dots, a_{j_2}, 0, \dots, 0]^{\intercal}.$$

Now consider A. Note that A has k^2 blocks. Let $B_{\alpha\beta} \in F^{n_\alpha \times n_\beta}$ be the blocks of A, where $\alpha, \beta \le k$. Then we have $B_{\alpha\beta} = 0$ for $\alpha \ne \beta$, since A is block diagonal. We also have $B_{\beta\beta} = [T|_{W_\beta}]_{\mathcal{B}_\beta}$. Furthermore, by definition, for $r \le n_\alpha, s \le n_\beta$ we have

$$A_{ij} = (B_{\alpha\beta})_{rs}, \text{ where } i = r + \sum_{\tilde{\alpha} < \alpha} n_{\tilde{\alpha}},\ j = s + \sum_{\tilde{\beta} < \beta} n_{\tilde{\beta}}.$$

To simplify the notation let us set $N_\beta := \sum_{\tilde{\beta} < \beta} n_{\tilde{\beta}}$. Consider a fixed j, and suppose $N_\beta < j \le N_\beta + n_\beta$ for some β. If $i \le N_\beta$ then A_{ij} is an entry of $B_{\alpha\beta}$ for some $\alpha < \beta$, hence we have $A_{ij} = 0$. Similarly if $i > N_\beta + n_\beta$ then $A_{ij} = 0$, since it is an entry of $B_{\alpha\beta}$ for some $\alpha > \beta$. Finally suppose that $N_\beta < i \le N_\beta + n_\beta$. Let

$$r := i - \sum_{\tilde{\beta} < \beta} n_{\tilde{\beta}}, \qquad s := j - \sum_{\tilde{\beta} < \beta} n_{\tilde{\beta}}.$$

Then we have $A_{ij} = (B_{\beta\beta})_{rs} = \big([T|_{W_\beta}]_{\mathcal{B}_\beta}\big)_{rs}$.

On the other hand note that if $N_\beta < j \le N_\beta + n_\beta$ then $v_j \in \mathcal{B}_\beta$, because N_β is the number of vectors in $\bigcup_{\alpha=1}^{\beta-1} \mathcal{B}_\alpha$. In addition remember that $Tv_j = \sum_{j_1 \le l \le j_2} a_l v_l$. Also note that $j_1 = N_\beta + 1$, so $s = j - j_1 + 1$. Thus the s-th column of $[T|_{W_\beta}]_{\mathcal{B}_\beta}$ is

$$\begin{aligned}\big([T|_{W_\beta}]_{\mathcal{B}_\beta}\big)_{\cdot,s} &= [T|_{W_\beta}(v_{j_1+s-1})]_{\mathcal{B}_\beta} = [T|_{W_\beta} v_j]_{\mathcal{B}_\beta} \\ &= [Tv_j]_{\mathcal{B}_\beta} = [a_{j_1}, \dots, a_{j_2}]^{\intercal} \in F^{n_\beta}.\end{aligned}$$

Therefore $A_{\cdot,j} = [0, \dots, 0, a_{j_1}, \dots, a_{j_2}, 0, \dots, 0]^{\intercal} = \big([T]_{\mathcal{B}}\big)_{\cdot,j}$ as desired. ■

Our next step is to find appropriate T-invariant subspaces such that their direct sum is V. Remember that the eigenspaces of T are T-invariant subspaces, and their sum is a direct sum. But the direct sum of the eigenspaces of T is not equal to V, if T is not diagonalizable. So in a sense, we can say that if T is not diagonalizable,

then it does not have enough eigenvectors to generate the vector space V. Therefore we need to add new vectors to the set of eigenvectors of T, so that we can generate V. These new vectors are called the generalized eigenvectors of T, and they are defined below.

Definition 8.5. A vector $v \in V$ is called a **generalized eigenvector** of T corresponding to the scalar $\lambda \in F$, if $v \neq 0$, and there exists $m \in \mathbb{N}$ such that

$$(T - \lambda I)^m v = 0.$$

Theorem 8.6. *If T has a generalized eigenvector corresponding to a scalar λ, then λ is an eigenvalue of T.*

Proof. Suppose v is a generalized eigenvector corresponding to λ, and m is the smallest positive integer such that $(T-\lambda I)^m v = 0$. If $m = 1$ then v is an eigenvector of T, and λ is an eigenvalue of T. Otherwise we have $w := (T - \lambda I)^{m-1} v \neq 0$, since $m - 1$ is smaller than m. But then we have

$$(T - \lambda I)w = (T - \lambda I)(T - \lambda I)^{m-1} v = (T - \lambda I)^m v = 0.$$

Therefore w is an eigenvector of T, and λ is an eigenvalue of T, as desired. ■

Remark. The above theorem means that there is no notion of generalized eigenvalue, and there is no need for it.

Definition 8.7. Suppose that λ is an eigenvalue of T. The set that consists of $0 \in V$ and all the generalized eigenvectors of T corresponding to λ, is called the **generalized eigenspace** of T corresponding to λ. We denote this set by $G_\lambda(T)$, or simply by G_λ when T is clear from the context.

Remark. Note that every eigenvector is also a generalized eigenvector. Thus every eigenspace of T is a subset of the corresponding generalized eigenspace, i.e.

$$E_\lambda(T) \subset G_\lambda(T).$$

Theorem 8.8. *The generalized eigenspaces of T are subspaces, and they are T-invariant.*

Proof. Suppose λ is an eigenvalue of T, and $u, v \in G_\lambda(T)$. Then there are $m, k \in \mathbb{N}$ such that

$$(T - \lambda I)^m u = 0, \qquad (T - \lambda I)^k v = 0.$$

Let us suppose that $m \geq k$; the other case is similar. Then we have

$$(T - \lambda I)^m v = (T - \lambda I)^{m-k} (T - \lambda I)^k v = (T - \lambda I)^{m-k}\, 0 = 0.$$

Hence for every $a \in F$ we have

$$(T - \lambda I)^m(u + av) = (T - \lambda I)^m u + a(T - \lambda I)^m v = 0.$$

Thus $u + av \in G_\lambda(T)$. Now note that $G_\lambda(T)$ is nonempty, since $0 \in G_\lambda(T)$ by definition. Therefore $G_\lambda(T)$ is a subspace.

In addition we have

$$(T - \lambda I)^m Tu = T(T - \lambda I)^m u = T(0) = 0.$$

Note that $T, (T - \lambda I)^m$ commute, since $(T - \lambda I)^m$ is a polynomial in T. So we have shown that $Tu \in G_\lambda(T)$. Hence $G_\lambda(T)$ is T-invariant. ■

Remark. Note that if λ is an eigenvalue of T, then $G_\lambda(T)$ is a nonzero subspace, since it contains at least one nonzero eigenvector of T corresponding to λ. Thus in particular when $G_\lambda(T)$ is finite dimensional we have $\dim G_\lambda(T) \geq 1$.

Theorem 8.9. *Suppose V is finite dimensional. Let λ be an eigenvalue of T. Then we have*

$$G_\lambda(T) = \operatorname{null}(T - \lambda I)^n,$$

where $n = \dim V$.

Remark. In other words, for every generalized eigenvector v corresponding to λ we have

$$(T - \lambda I)^n v = 0.$$

Proof. It is trivial that $\operatorname{null}(T - \lambda I)^n \subset G_\lambda(T)$. To show the reverse inclusion, suppose $v \in G_\lambda(T)$ is nonzero. Let m be the smallest positive integer such that $(T - \lambda I)^m v = 0$. It is enough to show that $m \leq n$. Consider the following m nonzero vectors

$$v, (T - \lambda I)v, \ldots, (T - \lambda I)^{m-1} v.$$

We claim that these vectors are linearly independent. To prove this, let

$$w_j := (T - \lambda I)^{j-1} v,$$

for $j = 1, \ldots, m$. Suppose $a_1 w_1 + \cdots + a_m w_m = 0$, where $a_1, \ldots, a_m \in F$. We have to show that every a_j is zero. Suppose we have shown that $a_1 = \cdots = a_{k-1} = 0$. Then we have $a_k w_k + \cdots + a_m w_m = 0$. Therefore

$$\begin{aligned} 0 &= (T - \lambda I)^{m-k} 0 = (T - \lambda I)^{m-k}(a_k w_k + \cdots + a_m w_m) \\ &= a_k(T - \lambda I)^{m-k} w_k + \cdots + a_m(T - \lambda I)^{m-k} w_m \\ &= a_k(T - \lambda I)^{m-1} v + a_{k+1}(T - \lambda I)^m v + \cdots + a_m(T - \lambda I)^{2m-k-1} v \\ &= a_k w_m + 0 + \cdots + 0 = a_k w_m. \end{aligned}$$

But w_m is nonzero, hence we must have $a_k = 0$. So if we continue this argument inductively, we will show that $a_1 = \cdots = a_m = 0$. Note that this argument also works when $k = 1$, thus we do not need to check the base of induction separately. Hence $w_1, \ldots, w_m$ are linearly independent. Therefore we must have $m \leq n$, as desired. ∎

Remark. Suppose V is finite dimensional, and we know that λ is an eigenvalue of T. Then the above description of $G_\lambda(T)$ as a null space, enables us to easily find a basis for it by using Theorem 3.48.

Proposition 8.10. *Suppose v is an eigenvector of T corresponding to the eigenvalue λ. Then for every polynomial $p \in F[x]$ we have*

$$p(T)v = p(\lambda)v,$$

i.e. v is an eigenvector of $p(T)$ corresponding to the eigenvalue $p(\lambda)$.

Proof. We know that $Tv = \lambda v$. Let us show by induction that $T^n v = \lambda^n v$ for every $n \in \mathbb{N}$. The case of $n = 1$ is trivial. And if the claim holds for some n, then for $n + 1$ we have

$$T^{n+1}v = TT^n v = T(T^n v) = T(\lambda^n v) = \lambda^n Tv = \lambda^n(\lambda v) = \lambda^{n+1}v.$$

Now suppose $p(x) = a_0 + \cdots + a_m x^m$. Then we have

$$p(T)v = \Big(\sum_{j\leq m} a_j T^j\Big)v = \sum_{j\leq m} a_j T^j v = \sum_{j\leq m} a_j \lambda^j v = \Big(\sum_{j\leq m} a_j \lambda^j\Big)v = p(\lambda)v,$$

as desired. ∎

Theorem 8.11. *Suppose $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of T. Then the generalized eigenspaces $G_{\lambda_1}(T), \ldots, G_{\lambda_k}(T)$ are independent subspaces.*

Proof. Suppose $v_j \in G_{\lambda_j}$, and $v_1 + \cdots + v_k = 0$. We have to show that $v_j = 0$ for every j. For each j, suppose m_j is the smallest nonnegative integer such that

$$(T - \lambda_j I)^{m_j} v_j = 0.$$

If $v_j \neq 0$ then we must have $m_j > 0$, since otherwise we would have

$$v_j = Iv_j = (T - \lambda_j I)^0 v_j = (T - \lambda_j I)^{m_j} v_j = 0.$$

Now suppose to the contrary that $v_l \neq 0$ for some $l \leq k$. Then $w := (T - \lambda_l I)^{m_l - 1} v_l$ is nonzero, and we have $(T - \lambda_l I)w = 0$. Hence w is an eigenvector of T corresponding to the eigenvalue λ_l. Let

$$q(x) := \prod_{j\neq l} (x - \lambda_j)^{m_j}, \qquad p(x) := q(x)(x - \lambda_l)^{m_l - 1}.$$

Note that $q(\lambda_l) \neq 0$. Now we have

$$0 = p(T)(0) = p(T)(v_1 + \cdots + v_k) = p(T)v_1 + \cdots + p(T)v_k.$$

For $j \neq l$ we have $p(x) = g(x)(x - \lambda_j)^{m_j}$, where g is some polynomial. Thus

$$p(T)v_j = g(T)(T - \lambda_j I)^{m_j} v_j = g(T)(0) = 0.$$

On the other hand

$$p(T)v_l = q(T)(T - \lambda_l)^{m_l - 1} v_l = q(T)w = q(\lambda_l)w.$$

Hence we have $q(\lambda_l)w = 0$. But $q(\lambda_l) \neq 0$, so we must have $w = 0$, which is a contradiction. Thus $v_l = 0$, and therefore every v_j must be zero, as desired. ■

Theorem 8.12. *Suppose V is finite dimensional, and F is algebraically closed. Then T has distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, and we have*

$$V = G_{\lambda_1}(T) \oplus \cdots \oplus G_{\lambda_k}(T).$$

Remark. Note that the theorem expresses that $\lambda_1, \ldots, \lambda_k$ are all the eigenvalues of T, and that they are also distinct.

Remark. Also note that the sum of generalized eigenspaces is a direct sum, since they are independent subspaces. Hence the nontrivial statement in this theorem is that the sum of generalized eigenspaces of any operator is the whole space V, when the field of scalars is algebraically closed.

Remark. A trivial consequence of this theorem is that

$$\dim V = \sum_{j=1}^{k} \dim G_{\lambda_j}(T).$$

Proof. Let $n = \dim V$. We prove the theorem by induction on n. The case of $n = 1$ is trivial, because every operator on a one-dimensional space is just multiplication by some scalar. So any nonzero vector in the space is a basis for the space, and an eigenvector of the operator. Now suppose the theorem is true for operators on vector spaces with dimension less than n. We know that T has at least one eigenvalue λ_1, since F is algebraically closed. Let $W := (T - \lambda_1 I)^n(V)$. We claim that

$$V = G_{\lambda_1}(T) \oplus W.$$

To see this, first note that if

$$w \in G_{\lambda_1}(T) \cap W,$$

then we have $w = (T - \lambda_1 I)^n v$ for some $v \in V$. We also have $(T - \lambda_1 I)^n w = 0$, since $G_{\lambda_1}(T) = \text{null}(T - \lambda_1 I)^n$. Therefore we get

$$(T - \lambda_1 I)^{2n} v = 0 \implies v \in G_{\lambda_1}(T) \implies w = (T - \lambda_1 I)^n v = 0.$$

Hence $G_{\lambda_1}(T), W$ are independent subspaces. But we have

$$\begin{aligned} \dim\big(G_{\lambda_1}(T) \oplus W\big) &= \dim G_{\lambda_1}(T) + \dim W \\ &= \dim \text{null}(T - \lambda_1 I)^n + \dim (T - \lambda_1 I)^n(V) = \dim V, \end{aligned}$$

where we used the rank-nullity theorem in the last line. Thus $G_{\lambda_1}(T) \oplus W$ is a subspace of V that has the same dimension as V. Hence $G_{\lambda_1}(T) \oplus W = V$ as desired.

Now note that $\dim W < n$, since T has at least one eigenvector corresponding to λ_1, and therefore $G_{\lambda_1}(T)$ is a nonzero subspace. Also note that W is T-invariant, because it is the image of a polynomial in T. Let

$$S := T|_W.$$

Then by the induction hypothesis, S has distinct eigenvalues, which we call them $\lambda_2, \ldots, \lambda_k$, and we have

$$W = G_{\lambda_2}(S) \oplus \cdots \oplus G_{\lambda_k}(S).$$

Next note that none of $\lambda_2, \ldots, \lambda_k$ equals λ_1. The reason is that if for $w \in W$ we have $Sw = \lambda_1 w$, then $Tw = T|_W w = Sw = \lambda_1 w$ too. Hence $w \in E_{\lambda_1}(T) \subset G_{\lambda_1}(T)$. Thus $w = 0$, since W and $G_{\lambda_1}(T)$ are independent subspaces.

In addition, note that $\lambda_2, \ldots, \lambda_k$ are also eigenvalues of T. Because if for a nonzero $w \in W$ we have $Sw = \lambda_j w$, then we also have $Tw = T|_W w = Sw = \lambda_j w$. Furthermore, by Exercise 2.57 we have

$$V = G_{\lambda_1}(T) \oplus W = G_{\lambda_1}(T) \oplus G_{\lambda_2}(S) \oplus \cdots \oplus G_{\lambda_k}(S). \qquad (*)$$

Therefore we only need to show that

$$G_{\lambda_j}(S) = G_{\lambda_j}(T),$$

for every $j \geq 2$. It is obvious that $G_{\lambda_j}(S) \subset G_{\lambda_j}(T)$, since if for some $w \in W$ and $m \in \mathbb{N}$ we have $(S - \lambda_j I_W)^m w = 0$, then we also have $(T - \lambda_j I_V)^m w = 0$. To show the reverse inclusion suppose $v \in G_{\lambda_j}(T)$. By equality $(*)$ we have $v = v_1 + v_2 + \cdots + v_k$, where $v_1 \in G_{\lambda_1}(T)$, and $v_i \in G_{\lambda_i}(S)$ for $i \geq 2$. We can rewrite this equality as

$$v_1 + \cdots + v_{j-1} + (v_j - v) + v_{j+1} + \cdots + v_k = 0.$$

But $v_i \in G_{\lambda_i}(T)$ for $i \neq j$, and $v_j - v \in G_{\lambda_j}(T)$. Furthermore, we know that the generalized eigenspaces of T are independent subspaces. Therefore we must have $v_i = 0$ for $i \neq j$, and $v_j - v = 0$. Hence

$$v = v_j \in G_{\lambda_j}(S).$$

Thus $G_{\lambda_j}(T) \subset G_{\lambda_j}(S)$ too, as desired.

Finally let us show that $\lambda_1, \ldots, \lambda_k$ are all the eigenvalues of T. Suppose to the contrary that T has a different eigenvalue λ. Let v be an eigenvector of T corresponding to λ. Then we have shown that

$$v = v_1 + \cdots + v_k,$$

where $v_i \in G_{\lambda_i}(T)$. Now we have $v_1 + \cdots + v_k + (-v) = 0$, where $v_i \in G_{\lambda_i}(T)$ and $-v \in E_\lambda(T) \subset G_\lambda(T)$. But the generalized eigenspaces corresponding to distinct eigenvalues are independent, so we must have $v_i = 0$ for every i, and $-v = 0$. However this contradicts the fact that v is an eigenvector, and thus it must be nonzero. Hence T cannot have any other eigenvalue besides $\lambda_1, \ldots, \lambda_k$. ■

8.2 The Jordan Form

Suppose V is finite dimensional, and F is algebraically closed. In the last section we have found appropriate T-invariant subspaces such that their direct sum is V. These subspaces are the generalized eigenspaces of T. Our final step is to understand the behavior of the restriction of T to its generalized eigenspaces. Suppose λ is an eigenvalue of T. Let $W := G_\lambda(T)$, and $N := T|_W - \lambda I$. Then for every $w \in W$ we have $N^n w = 0$, where $n = \dim V$. In other words $N^n = 0$ on W. So if we understand the behavior of operators with this property, then we can understand $T|_W$.

Definition 8.13. An operator $N \in \mathcal{L}(V)$ is called **nilpotent** if there exists $m \in \mathbb{N}$ such that

$$N^m = 0.$$

Theorem 8.14. *Suppose V is finite dimensional, and $N \in \mathcal{L}(V)$ is nilpotent. Then we have*

$$N^n = 0,$$

where $n = \dim V$.

Proof. Let m be the smallest positive integer such that $N^m = 0$. It is enough to show that $m \leq n$, because then we would have $N^n = N^{n-m}N^m = N^{n-m}0 = 0$.

Now since $N^{m-1} \neq 0$, there is a vector $v \in V$ such that $N^{m-1}v \neq 0$. Hence we have the following m nonzero vectors

$$v, Nv, \ldots, N^{m-1}v.$$

We claim that these vectors are linearly independent. To prove this, suppose

$$a_1 v + \cdots + a_m N^{m-1} v = 0,$$

for some $a_1, \ldots, a_m \in F$. We have to show that every a_j is zero. Suppose we have shown that $a_1 = \cdots = a_{k-1} = 0$. Then we have $a_k N^{k-1}v + \cdots + a_m N^{m-1}v = 0$. Hence we get

$$\begin{aligned} 0 = N^{m-k}(0) &= N^{m-k}(a_k N^{k-1}v + \cdots + a_m N^{m-1}v) \\ &= a_k N^{m-1}v + a_{k+1}N^m v + \cdots + a_m N^{2m-k-1}v \\ &= a_k N^{m-1}v + 0 + \cdots + 0 = a_k N^{m-1}v. \end{aligned}$$

But $N^{m-1}v$ is nonzero, thus we must have $a_k = 0$. So if we continue this argument inductively, we will show that $a_1 = \cdots = a_m = 0$. Note that the above argument also works when $k = 1$, thus we do not need to check the base of induction separately. Therefore $v, Nv, \ldots, N^{m-1}v$ are linearly independent. Hence we must have $m \leq n$, as desired. ∎

Theorem 8.15. *Suppose V is finite dimensional, and $N \in \mathcal{L}(V)$ is nilpotent. Then there are $v_1, \ldots, v_k \in V$ and $m_1, \ldots, m_k \in \mathbb{N}$ such that*

$$N^{m_1-1}v_1, \ldots, Nv_1, v_1, \quad N^{m_2-1}v_2, \ldots, v_2, \quad \ldots \quad N^{m_k-1}v_k, \ldots, v_k$$

is a basis for V. Furthermore we have $N^{m_j}v_j = 0$ for every j.

Proof. The proof is by induction on $\dim V$. When $\dim V = 1$ the result holds trivially, because the previous theorem implies that $N = N^1 = N^{\dim V} = 0$. So any nonzero vector in the space is a basis for the space, and has our desired property. Now suppose the theorem holds for every nilpotent operator on a nonzero vector space whose dimension is less than $\dim V$. If $N = 0$ then every basis for V has our desired property, since if $v_1, \ldots, v_n$ is a basis for V, then we have $Nv_j = 0$ for all j.

So suppose that $N \neq 0$. Let m be the smallest positive integer such that $N^m = 0$. Note that $m > 1$, since $N \neq 0$. Thus $N^{m-1} \neq 0$. Hence there is $v \in V$ so that $N^{m-1}v \neq 0$. But we know that $N(N^{m-1}v) = N^m v = 0(v) = 0$. Therefore $N^{m-1}v \in \operatorname{null} N$. Thus $\operatorname{null} N \neq \{0\}$. Hence we have

$$\dim N(V) = \dim V - \dim \operatorname{null} N < \dim V.$$

On the other hand we know that $N(V)$ is N-invariant, as shown in Proposition 4.17. Let $W := N(V)$, and let $M := N|_W$. Then by Exercise 4.16 we have $M^m = N^m|_W = 0|_W = 0$. Hence M is also nilpotent. Thus we can apply the induction hypothesis to M, and find the following basis for W

$$M^{m_1-1}w_1, \ldots, Mw_1, w_1, \quad \ldots \quad M^{m_k-1}w_k, \ldots, w_k,$$

where $M^{m_j}w_j = 0$ for each j. But we know that $M^l = N^l|_W$ for every l. So for every $w \in W$ we have $M^l w = N^l|_W w = N^l w$. Therefore the above basis is the same as

$$N^{m_1-1}w_1, \ldots, Nw_1, w_1, \quad \ldots \quad N^{m_k-1}w_k, \ldots, w_k, \qquad (*)$$

where $N^{m_j}w_j = 0$ for each j. Next note that each w_j is in $W = N(V)$, so there are $v_j \in V$ such that $w_j = Nv_j$ for each j. Consider the list of vectors

$$N^{m_1}v_1, \ldots, N^2v_1, Nv_1, v_1, \quad \ldots \quad N^{m_k}v_k, \ldots, Nv_k, v_k.$$

Note that if we remove the vectors $v_1, \ldots, v_k$ from this list, we get the list $(*)$.

Now note that $N^{m_j+1}v_j = N^{m_j}Nv_j = N^{m_j}w_j = 0$. Thus $N(N^{m_j}v_j) = N^{m_j+1}v_j = 0$. Hence $N^{m_j}v_j \in \operatorname{null} N$ for every j. Obviously, $N^{m_1}v_1, \ldots, N^{m_k}v_k$ are linearly independent, since they are the same as $N^{m_1-1}w_1, \ldots, N^{m_k-1}w_k$, and these vectors belong to a basis for W. Now we extend the linearly independent list $N^{m_1}v_1, \ldots, N^{m_k}v_k$ to a basis for $\operatorname{null} N$, by adding the vectors $u_1, \ldots, u_p$.

Finally we claim that

$$N^{m_1}v_1, \ldots, v_1, \quad \ldots \quad N^{m_k}v_k, \ldots, v_k, \, u_1, \ldots, u_p$$

is a basis for V, which has our desired property. Note that for every j, l we have $N^{m_j+1}v_j = 0$, and $Nu_l = 0$. So the above list has our desired property. Now let $v \in V$ be an arbitrary vector. Then $Nv \in N(V) = W$. Hence there are $b_{ij} \in F$ such that

$$\begin{aligned} Nv &= \sum_{i \le k} \sum_{j < m_i} b_{ij} N^j w_i = \sum_{i \le k} \sum_{j < m_i} b_{ij} N^j N v_i \\ &= \sum_{i \le k} \sum_{j < m_i} b_{ij} N N^j v_i = N\Big(\sum_{i \le k} \sum_{j < m_i} b_{ij} N^j v_i \Big). \end{aligned}$$

Note that we have used the fact that N, N^j commute for every j. Therefore we can conclude that

$$N\Big(v - \sum_{i \le k} \sum_{j < m_i} b_{ij} N^j v_i \Big) = 0.$$

Thus we have $v - \sum_{i\le k}\sum_{j<m_i} b_{ij}N^j v_i \in \operatorname{null} N$. Hence there are $b_{im_i}, b_l \in F$ so that

$$v - \sum_{i\le k}\sum_{j<m_i} b_{ij}N^j v_i = \sum_{i\le k} b_{im_i}N^{m_i}v_i + \sum_{l\le p} b_l u_l$$
$$\implies v = \sum_{i\le k}\sum_{j\le m_i} b_{ij}N^j v_i + \sum_{l\le p} b_l u_l.$$

Thus v is in the span of the proposed basis.

It only remains to show that the proposed basis is linearly independent. Suppose

$$\sum_{i\le k}\sum_{j\le m_i} a_{ij}N^j v_i + \sum_{l\le p} a_l u_l = 0, \qquad (**)$$

for some $a_{ij}, a_l \in F$. Then we get

$$\begin{aligned}\sum_{i\le k}\sum_{j<m_i} a_{ij}N^j w_i &= \sum_{i\le k}\sum_{j<m_i} a_{ij}N^j N v_i \\ &= \sum_{i\le k}\sum_{j<m_i} a_{ij}NN^j v_i = N\Big(\sum_{i\le k}\sum_{j<m_i} a_{ij}N^j v_i\Big) \\ &= N\Big(\sum_{i\le k}\sum_{j<m_i} a_{ij}N^j v_i\Big) + N\Big(\sum_{i\le k} a_{im_i}N^{m_i}v_i + \sum_{l\le p} a_l u_l\Big) \\ &= N\Big(\sum_{i\le k}\sum_{j\le m_i} a_{ij}N^j v_i + \sum_{l\le p} a_l u_l\Big) = N(0) = 0.\end{aligned}$$

Note that we are using the fact that $N\big(\sum_{i\le k} a_{im_i}N^{m_i}v_i + \sum_{l\le p} a_l u_l\big) = 0$, since $N^{m_i}v_i, u_l \in \operatorname{null} N$. Hence $a_{ij} = 0$ for $i \le k$ and $j < m_i$, because $N^j w_i$'s form a basis for W. Thus equation $(**)$ implies that

$$\sum_{i\le k} a_{im_i}N^{m_i}v_i + \sum_{l\le p} a_l u_l = 0.$$

Therefore we must have $a_{im_i} = 0$ for $i \le k$ and $a_l = 0$ for $l \le p$, since $N^{m_i}v_i$'s and u_l's form a basis for $\operatorname{null} N$. ∎

Remark. Let $\mathcal{B}$ be the basis

$$N^{m_1-1}v_1, \ldots, v_1, \quad N^{m_2-1}v_2, \ldots, v_2, \quad \ldots \quad N^{m_k-1}v_k, \ldots, v_k.$$

Then the entires of the matrix $[N]_{\mathcal{B}}$ are all 0, except for some 1's on the diagonal immediately above its main diagonal. Because the image of any vector in the basis is either the previous vector in the basis, or zero. In addition, the matrix $[N]_{\mathcal{B}}$ is

block diagonal with k diagonal blocks, and its j-th diagonal block has size $m_j \times m_j$. To see this let

$$W_j := \operatorname{span}(N^{m_j-1}v_j, N^{m_j-2}v_j, \ldots, v_j).$$

Then for $i < m_j - 1$ we have $N(N^i v_j) = N^{i+1}v_j \in W_j$, and for $i = m_j - 1$ we have $N(N^i v_j) = N^{m_j}v_j = 0 \in W_j$. Hence as shown in Exercise 4.2, each W_j is N-invariant. In addition by Theorem 2.56 we have

$$V = W_1 \oplus \cdots \oplus W_k,$$

since $\mathcal{B}_j := \{N^{m_j-1}v_j, \ldots, v_j\}$ is a basis for W_j, and $\bigcup_{j\le k} \mathcal{B}_j$ is a basis for V. Thus by Theorem 8.4 the matrix $[N]_{\mathcal{B}}$ is block diagonal, and its diagonal blocks are $[N|_{W_1}]_{\mathcal{B}_1}, \ldots, [N|_{W_k}]_{\mathcal{B}_k}$. Finally note that $\mathcal{B}_j$ has m_j elements.

Now consider a fixed j. Then for every $0 \le i < m_j - 1$ we have $N(N^i v_j) = N^{i+1}v_j$, i.e. the image of the $(m_j - i)$-th vector in the basis $\mathcal{B}_j$ is the $(m_j - i - 1)$-th vector in the basis. We also have $N(N^{m_j-1}v_j) = 0$, i.e. the image of the first vector in the basis is zero. Remember that the coordinate vector with respect to a basis, of the i-th element of that basis, is e_i, as shown in Example 3.30. Hence we have

$$\begin{aligned}
[N|_{W_j}]_{\mathcal{B}_j} &= \left[\; [N(N^{m_j-1}v_j)]_{\mathcal{B}_j} \;\middle|\; [N(N^{m_j-2}v_j)]_{\mathcal{B}_j} \;\middle|\; \ldots \;\middle|\; [N(Nv_j)]_{\mathcal{B}_j} \;\middle|\; [Nv_j]_{\mathcal{B}_j} \;\right] \\
&= \left[\; [N^{m_j}v_j]_{\mathcal{B}_j} \;\middle|\; [N^{m_j-1}v_j]_{\mathcal{B}_j} \;\middle|\; \ldots \;\middle|\; [N^2v_j]_{\mathcal{B}_j} \;\middle|\; [Nv_j]_{\mathcal{B}_j} \;\right] \\
&= \left[\; 0 \;\middle|\; e_1 \;\middle|\; \ldots \;\middle|\; e_{m_j-2} \;\middle|\; e_{m_j-1} \;\right] \\
&= \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & & 0 \\
 & & \ddots & \ddots & \\
\vdots & & & \ddots & 1 \\
0 & 0 & & \cdots & 0
\end{bmatrix} \in F^{m_j \times m_j}.
\end{aligned}$$

Therefore we have a complete characterization of $[N]_{\mathcal{B}}$. ■

***Example* 8.16.** Let us consider a more concrete example related to the above observations. Suppose $k = 4$, $m_1 = 3$, $m_2 = 2$, and $m_3 = m_4 = 1$. Then our basis $\mathcal{B}$ becomes

$$N^2v_1, Nv_1, v_1, Nv_2, v_2, v_3, v_4.$$

In addition we know that $N^3v_1 = N^2v_2 = Nv_3 = Nv_4 = 0$. Therefore the matrix

of N is

$$\begin{bmatrix} 0 & 1 & 0 & & & & \\ 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & & & & \\ & & & 0 & 1 & & \\ & & & 0 & 0 & & \\ & & & & & 0 & \\ & & & & & & 0 \end{bmatrix}.$$

Note that every entry of the above matrix that is not displayed is zero. Also notice the block diagonal structure of the matrix. It has 4 diagonal blocks, of which two blocks are 1×1 zero matrix.

Jordan Form. *Suppose V is a finite dimensional vector space over an algebraically closed field, and T is an operator on V. Then V has a basis in which the entires of the matrix of T are all 0 except on the main diagonal and the diagonal immediately above the main diagonal. The entries on the main diagonal are the eigenvalues of T, and the entries on the diagonal immediately above the main diagonal are either 0 or 1.*

Proof. We can decompose V as

$$V = G_{\lambda_1}(T) \oplus \cdots \oplus G_{\lambda_k}(T).$$

Each $G_{\lambda_j}(T)$ is T-invariant, and the restriction of T to it is of the form

$$\lambda_j I + N_j,$$

where N_j is nilpotent. We can choose a basis for each generalized eigenspace in which the matrix of N_j has the required form. The union of these bases is the desired basis for V. ■

Remark. In particular, this theorem shows that every operator on a vector space over an algebraically closed field can be represented by an upper triangular matrix.

8.3 The Minimal Polynomial

Definition 8.17. Let T be an operator on a space V. A nonzero monic polynomial p is called the **minimal polynomial** of T if $p(T) = 0$, and p has the smallest degree among the polynomials with this property.

Remark. If T has a minimal polynomial, it is easy to see that its minimal polynomial is unique.

Theorem 8.18. *If an operator T has the minimal polynomial p, and f is a polynomial such that $f(T) = 0$, then f is a multiple of p.*

Proof. We can divide f by p to obtain $f = pq + r$, where $\deg r < \deg p$. Suppose to the contrary that $r \neq 0$. Then as $r(T) = f(T) - p(T)q(T) = 0$ and $\deg r < \deg p$, we get a contradiction. ∎

Theorem 8.19. *Every operator on a finite dimensional space has a minimal polynomial.*

Proof. Suppose T is an operator on an n-dimensional space V. Then $T \in \mathcal{L}(V)$, and $\dim \mathcal{L}(V) = n^2$. Hence $I, T, T^2, \ldots, T^{n^2}$ are linearly dependent. Let m be the smallest integer for which

$$I, T, T^2, \ldots, T^m$$

are linearly dependent. Then there are scalars a_i, not all of them zero, such that

$$a_0 I + a_1 T + \cdots + a_m T^m = 0.$$

First note that $a_m \neq 0$, since otherwise $I, T, T^2, \ldots, T^{m-1}$ would be linearly dependent, contradicting our choice of m.

We claim that $p(x) := x^m + \frac{a_{m-1}}{a_m} x^{m-1} + \cdots + \frac{a_0}{a_m}$ is the minimal polynomial of T. It is obvious that $p(T) = 0$. If there exist a polynomial q with $\deg q < m$, then we would have a linear dependence relation between $I, T, T^2, \ldots, T^{m-1}$, which is again a contradiction. ∎

Theorem 8.20. *Suppose an operator T has the minimal polynomial p. Then a scalar λ is an eigenvalue of T if and only if $p(\lambda) = 0$.*

Proof. Suppose $Tv = \lambda v$ for a nonzero vector v. Then $0 = p(T)v = p(\lambda)v$ and as $v \neq 0$ we have $p(\lambda) = 0$.

Now suppose $p(\lambda) = 0$. Then $p(x) = (x - \lambda)q(x)$, where $\deg q < \deg p$. Hence $q(T) \neq 0$, so there is a vector w such that $q(T)w \neq 0$. Let $v := q(T)w$. Then

$$(T - \lambda I)v = (T - \lambda I)q(T)w = p(T)w = 0.$$ ∎

Appendix A

Rings

A.1 Rings

Definition A.1. A **ring** is a nonempty set R equipped with two binary operations

$$\begin{array}{c} R \times R \longrightarrow R \\ (a,b) \mapsto a+b \end{array}, \qquad \begin{array}{c} R \times R \longrightarrow R \\ (a,b) \mapsto ab \end{array},$$

called respectively **addition** and **multiplication**, such that

(i) The operations are **associative**, i.e. for all $a, b, c \in R$

$$a + (b + c) = (a + b) + c, \qquad a(bc) = (ab)c.$$

(ii) Addition is **commutative**, i.e. for all $a, b \in R$

$$a + b = b + a.$$

(iii) There exist elements $0, 1 \in R$, called respectively additive identity and multiplicative identity, such that for all $a \in R$

$$a + 0 = a, \qquad a1 = a = 1a.$$

(iv) For every $a \in R$ there exists $b \in R$, called its additive inverse, such that

$$a + b = 0.$$

(v) Multiplication is **distributive** over addition, i.e. for all $a, b, c \in R$

$$a(b + c) = ab + ac, \qquad (b + c)a = ba + ca.$$

***Example* A.2.** $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are all rings with the usual addition and multiplication. $\mathbb{N}$ is not a ring as it does not have an additive identity, and its elements do not have additive inverse.

Definition A.3. A multiplicative inverse of an element $a \in R$ is an element $b \in R$ such that

$$ab = 1 = ba.$$

In this case a is called **invertible**.

Proposition A.4. *Let R be a ring. Then for all $a, b, c \in R$ we have*

(i) **(Cancellation Laws)**

$$\begin{aligned} &a + c = b + c \implies a = b, \\ &ac = bc, \ c \text{ is invertible} \implies a = b, \\ &ca = cb, \ c \text{ is invertible} \implies a = b. \end{aligned}$$

(ii) *Additive and multiplicative identities of R are unique.*

(iii) *Additive inverse of any element of R is unique, and multiplicative inverse of any invertible element of R is unique*

(iv) $0a = 0 = a0$.

Proof. **(i)** Suppose d is an additive inverse of c. Then we can add d to both sides of $a + c = b + c$ to obtain $(a + c) + d = (b + c) + d$. Now by associativity of addition we have $a + (c + d) = b + (c + d)$. Since $c + d = 0$, we get $a + 0 = b + 0$; and hence $a = b$. The multiplicative cases can be proved similarly using a multiplicative inverse of c.

(ii) Suppose $0, \tilde{0}$ are both additive identities, then

$$\tilde{0} = \tilde{0} + 0 = 0 + \tilde{0} = 0.$$

Similarly for two multiplicative identities $1, \tilde{1}$ we have $\tilde{1} = \tilde{1}1 = 1$.

(iii) Suppose $b, \tilde{b}$ are both additive inverses of a. Then $a + b = 0 = a + \tilde{b}$, and by (i) we get $b = \tilde{b}$. The multiplicative case is similar.

(iv) We have

$$0 + 0a = 0a = (0 + 0)a = 0a + 0a.$$

Hence by cancellation law we get $0 = 0a$. The other equality can be proved similarly. ∎

Definition A.5. Additive and multiplicative identities of a ring are respectively called **zero** and **identity** of the ring.

The unique additive inverse of an element a is denoted by $-a$, and is called its **opposite**. Also for two elements a, b we set $a - b := a + (-b)$.

If an element a has multiplicative inverse, we denote it by a^{-1}, and we call it the **inverse** of a.

Proposition A.6. *Let R be a ring. Then for all $a, b \in R$ we have*

(i) $-(-a) = a$ *for all* $a \in R$.
(ii) $-(a+b) = (-a) + (-b) = -a - b$ *for all* $a, b \in R$.
(iii) *If* a *is invertible, then* a^{-1} *is also invertible and*

$$(a^{-1})^{-1} = a.$$

(iv) *If* a *and* b *are invertible, then* ab *is also invertible and we have*

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

(v) $(-a)b = -ab = a(-b)$ *for all* $a, b \in R$. *As a result*

$$(-a)(-b) = ab,$$
$$-a = (-1)a.$$

Proof. **(i)** This is similar to (iii).

(ii) This is similar to (iv). Note that the last equality in (ii) holds by definition.

(iii) Since $a^{-1}a = 1 = aa^{-1}$, a^{-1} is invertible, and we must have $(a^{-1})^{-1} = a$ due to the uniqueness of inverse.

(iv) First note that

$$\begin{aligned}(b^{-1}a^{-1})(ab) &= b^{-1}(a^{-1}(ab)) = b^{-1}((a^{-1}a)b)\\ &= b^{-1}(1b) = b^{-1}b = 1.\end{aligned}$$

Similarly $(ab)(b^{-1}a^{-1}) = 1$. Therefore ab is invertible. Now the result follows from the uniqueness of inverse.

(v) We have

$$ab + (-a)b = (a + (-a))b = 0b = 0.$$

Thus uniqueness of additive inverse implies $(-a)b = -ab$. The equality $a(-b) = -ab$ can be proved similarly. Now we have

$$\begin{aligned}(-a)(-b) &= -(-a)b = -(-ab) = ab,\\ (-1)a &= -(1a) = -a.\end{aligned}$$

■

***Exercise* A.7.** Show that if a is invertible, then $-a$ is also invertible and we have

$$(-a)^{-1} = -a^{-1}.$$

Solution. By the previous theorem we have

$$(-a^{-1})(-a) = a^{-1}a = 1 = aa^{-1} = (-a)(-a^{-1}).$$

Thus we get the desired result due to the uniqueness of inverse. ■

Definition A.8. A **commutative ring** is a ring in which multiplication is commutative, i.e. for all elements a, b we have

$$ab = ba.$$

Also, we say two elements a and b in a ring **commute** if $ab = ba$.

Definition A.9. In a ring R, for a positive integer n we inductively define

$$[0] := 0,\ [1] := 1, \quad \dots \quad [n] := [n-1] + 1.$$

We also define $[-n] := -[n]$.

Remark. Note that in the above definition, each one of the $0, 1$, and $\pm$, has two different meanings.

Notation. We abuse the notation and write n instead of $[n]$. We also set

$$na := [n]a.$$

Remark. The next proposition shows that the operations of $\mathbb{Z}$ on n's and the operations of R on n's are compatible. Therefore this abbreviation does not lead to any confusion.

Remark. na is actually the n-th *additive power* of a, i.e. the n-th power of a with respect to the binary operation $+$ as defined in Section A.6. To see this note that $0a = [0]a = 0$, since $[0] = 0$ (note that 0 has two meanings here). Also for $n > 0$ we have

$$na = [n]a = ([n-1] + 1)a = [n-1]a + 1a = (n-1)a + a.$$

In addition we have

$$(-n)a = [-n]a = (-[n])a = [n](-a) = n(-a).$$

Remark. A consequence of the above remark is that for $n > 0$ we have

$$na = \overbrace{a + a + \cdots + a}^{n \text{ times}},$$

since the right hand side is just the n-th power of a with respect to $+$.

Proposition A.10. *In any ring R we have*

(i) *For all $n \in \mathbb{Z}$, $[n]$ commutes with all elements of R.*

(ii) *For all $n, m \in \mathbb{Z}$*

$$[n + m] = [n] + [m], \qquad [nm] = [n][m].$$

Proof. **(i)** Let a be an arbitrary element of R. Then $[0]a = 0a = 0 = a0 = a[0]$, so $[0]$ commutes with every a. Now suppose for some $n > 0$, $[n]$ commutes with every a. Then we have

$$[n+1]a = ([n]+1)a = [n]a + 1a = a[n] + a1 = a([n]+1) = a[n+1].$$

Hence by induction, $[n]$ commutes with every a for all $n > 0$. Next suppose $n = -m < 0$. Then

$$[n]a = (-[m])a = -[m]a = -a[m] = a(-[m]) = a[n].$$

(ii) Since $[n] = n1$ is the n-th additive power of $1 \in R$, these relations are special cases of the properties of powers as proved in Section A.6. For example, for the second equality we can say that $[nm]$ is the nm-th power of 1, so it is equal to the m-th power of the n-th power of 1. But the n-th power of 1 is $[n]$. Hence the nm-th power of 1 equals the m-th power of $[n]$, which is $[m][n]$. But by (i) we have $[m][n] = [n][m]$. Therefore $[nm] = [n][m]$ as desired. ■

Proposition A.11. *Let R be a ring. Then for all $m, n \in \mathbb{Z}$ and all $a, b \in R$ we have*

(i) $(-n)a = n(-a) = -(na)$.
(ii) $(n+m)a = na + ma$.
(iii) $m(na) = (mn)a$.
(iv) $n(a+b) = na + nb$.
(v) $(ma)(nb) = mn(ab) = (na)(mb)$.
(vi) *If a commutes with b, then na commutes with mb.*

Proof. Parts (i) to (iv) are true for any notion of power as proved in Section A.6. They can also be proved directly, as we do below. We have

$$\begin{aligned}
(-n)a &= [-n]a = (-[n])a = [n](-a) = n(-a),\\
(-n)a &= [-n]a = (-[n])a = -[n]a = -na,\\
(n+m)a &= [n+m]a = ([n]+[m])a = [n]a + [m]a = na + ma,\\
m(na) &= [m]([n]a) = ([m][n])a = [mn]a = (mn)a,\\
n(a+b) &= [n](a+b) = [n]a + [n]b = na + nb.
\end{aligned}$$

For part (v) we have

$$(ma)(nb) = ([m]a)([n]b) = [m]a[n]b = [m][n]ab = [mn]ab = mn(ab).$$

Note that we used the generalized associativity of the product of R, and the fact that $[n]$ commutes with all elements of R. Now for the second equality we use the first one to obtain

$$(ma)(nb) = mn(ab) = nm(ab) = (na)(mb).$$

Finally, for part (vi) we have

$$(na)(mb) = nm(ab) = mn(ab) = mn(ba) = (mb)(na).$$ ■

Definition A.12. We define the **powers** of $a \in R$ as follows. For a positive integer n we inductively define

$$a^0 := 1,\ a^1 := a,\ \ldots\ a^n := a^{n-1}a.$$

If a is invertible, we define

$$a^{-n} := (a^{-1})^n.$$

Theorem A.13. *Let R be a ring. Then for all $a, b \in R$ we have*

(i) *If a commutes with b, then a^n commutes with b^m, for all $m, n \geq 0$. If one or both of a, b are invertible, we can allow n and/or m to be negative too.*

(ii) *If a is invertible, then a^n is also invertible for all $n \in \mathbb{Z}$, and*

$$(a^n)^{-1} = a^{-n} = (a^{-1})^n.$$

(iii) *$a^n a^m = a^{n+m}$ for all $m, n \geq 0$. If a is invertible, we can allow m, n to be negative too.*

(iv) *$(a^n)^m = a^{nm}$ for all $m, n \geq 0$. If a is invertible, we can allow m, n to be negative too.*

(v) *If a, b commute, we have $a^n b^n = (ab)^n$ for all $n \geq 0$. If a, b are invertible, we can allow n to be negative too.*

Proof. All the proofs are by induction. We will only write the induction steps below, since the base of inductions can be checked easily.

(i) For $m \geq 0$ we have

$$ab^{m+1} = ab^m b = b^m ab = b^m ba = b^{m+1}a.$$

If b is invertible we have

$$ab^{-1} = 1ab^{-1} = b^{-1}bab^{-1} = b^{-1}abb^{-1} = b^{-1}a.$$

Thus by the first part $b^{-m} = (b^{-1})^m$ commutes with a. By repeating this argument with fixed m, we see that a^n commutes with b^m too.

(ii) When $n \geq 0$ we have

$$(a^{n+1})a^{-n-1} = a^n a(a^{-1})^{n+1} = aa^n(a^{-1})^n a^{-1} = aa^n a^{-n}a^{-1} = aa^{-1} = 1.$$

When $n = -m < 0$ we have $a^{-m} = (a^{-1})^m$. Hence by the previous part we get

$$(a^{-m})^{-1} = ((a^{-1})^m)^{-1} = (a^{-1})^{-m} = ((a^{-1})^{-1})^m = a^m.$$

The second equality holds by definition when $n > 0$. When $n = 0$ we have

$$(a^{-1})^0 = 1 = a^0 = a^{-0}.$$

And when $n = -m < 0$ we have

$$(a^{-1})^{-m} = ((a^{-1})^{-1})^m = a^m = a^{-n}.$$

(iii) When $n, m \geq 0$ we have

$$a^n a^{m+1} = a^n a^m a = a^{n+m} a = a^{n+m+1}.$$

Now suppose a is invertible. Then we have

$$a^{-n}a^{m+1} = a^{-n}a^m a$$
$$= a^{-n+m}a = \begin{cases} (a^{-1})^{n-m}a = (a^{-1})^{n-m-1}a^{-1}a & \text{if } -n+m < 0, \\ \qquad = (a^{-1})^{n-m-1} = a^{-n+m+1} & \\ a^{-n+m+1} & \text{if } -n+m \geq 0. \end{cases}$$

We also have

$$a^n a^{-m} = (a^{-1})^{-n}(a^{-1})^m = (a^{-1})^{-n+m} = a^{n-m},$$
$$a^{-n}a^{-m} = (a^{-1})^n(a^{-1})^m = (a^{-1})^{n+m} = a^{-n-m}.$$

(iv) For $n, m \geq 0$ we have

$$(a^n)^{m+1} = (a^n)^m a^n = a^{nm}a^n = a^{nm+n} = a^{n(m+1)}.$$

If a is invertible we have

$$(a^{-n})^m = ((a^{-1})^n)^m = (a^{-1})^{nm} = a^{-nm},$$
$$(a^{\pm n})^{-m} = ((a^{\pm n})^m)^{-1} = (a^{\pm nm})^{-1} = a^{\mp nm}.$$

(v) For $n \geq 0$ we have

$$a^{n+1}b^{n+1} = a^n ab^n b = a^n b^n ab = (ab)^n ab = (ab)^{n+1},$$

and if a, b are invertible we have

$$a^{-n}b^{-n} = (a^{-1})^n(b^{-1})^n = (a^{-1}b^{-1})^n = ((ba)^{-1})^n = (ba)^{-n} = (ab)^{-n}. \qquad \blacksquare$$

Definition A.14. Let $n \in \mathbb{N}$. The ***n* factorial** is

$$n! := n \times (n-1) \times \cdots \times 2 \times 1.$$

We also set $0! := 1$. Suppose $n, k \in \mathbb{Z}$, and $0 \leq k \leq n$. The number

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

is called a **binomial coefficient**.

Remark. Note that for all $n \geq 1$ we have $n! = n(n-1)!$. It is also trivial to see that $\binom{n}{0} = 1 = \binom{n}{n}$ for all $n \geq 0$.

Proposition A.15. *For all integers $1 \leq k \leq n$ we have*

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$

As a result $\binom{n}{k}$ is always a positive integer for all $0 \leq k \leq n$.

Proof. We have

$$\begin{aligned}
\binom{n}{k} + \binom{n}{k-1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} \\
&= \frac{n!}{(k-1)!(n-k)!}\left(\frac{1}{k} + \frac{1}{n-k+1}\right) \\
&= \frac{n!}{(k-1)!(n-k)!}\,\frac{n+1}{k(n-k+1)} \\
&= \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}.
\end{aligned}$$

Next, we show by induction on n that $\binom{n}{k}$ is a positive integer for all $0 \leq k \leq n$. For $n = 1$ we have $\binom{1}{0} = \binom{1}{1} = 1 \in \mathbb{N}$. Suppose the claim holds for n. Then for $0 < k < n+1$ we have

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \in \mathbb{N}.$$

Also note that $\binom{n+1}{0} = \binom{n+1}{n+1} = 1 \in \mathbb{N}$. ∎

Theorem A.16. *For two commuting elements a, b in a ring, and a positive integer n, we have*

(i) **(Binomial Theorem)** $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$.

(ii) $a^n - b^n = (a-b)\big(\sum_{k=0}^{n-1} a^{n-1-k}b^k\big)$.

Proof. **(i)** The proof is by induction on n. The case of $n = 1$ is obvious. For the induction step we have

$$
\begin{aligned}
(a+b)^{n+1} = (a+b)^n(a+b) &= \Big(\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k\Big)(a+b)\\
&= \sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k(a+b) = \sum_{k=0}^{n}\binom{n}{k}(a^{n-k}b^k a + a^{n-k}b^k b)\\
&= \sum_{k=0}^{n}\binom{n}{k}(a^{n-k}ab^k + a^{n-k}b^{k+1})\\
&= \sum_{k=0}^{n}\left[\binom{n}{k}a^{n-k+1}b^k + \binom{n}{k}a^{n-k}b^{k+1}\right]\\
&= \sum_{k=0}^{n}\binom{n}{k}a^{n-k+1}b^k + \sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^{k+1}\\
&= \sum_{k=0}^{n}\binom{n}{k}a^{n+1-k}b^k + \sum_{j=1}^{n+1}\binom{n}{j-1}a^{n+1-j}b^j\\
&\qquad\qquad \text{(We replaced } k \text{ with } j-1 \text{ in the 2nd sum.)}\\
&= a^{n+1} + \left(\sum_{k=1}^{n}\binom{n}{k}a^{n+1-k}b^k + \sum_{k=1}^{n}\binom{n}{k-1}a^{n+1-k}b^k\right) + b^{n+1}\\
&\qquad\qquad \text{(We replaced } j \text{ with } k \text{ in the 2nd sum.)}\\
&= a^{n+1} + \left(\sum_{k=1}^{n}\left[\binom{n}{k} + \binom{n}{k-1}\right]a^{n+1-k}b^k\right) + b^{n+1}\\
&= \sum_{k=0}^{n+1}\binom{n+1}{k}a^{n+1-k}b^k.
\end{aligned}
$$

Note that since the binomial coefficients are positive integers, we can multiply the ring elements with them.

(ii) We have

$$
\begin{aligned}
(a-b)\Big(\sum_{k=0}^{n-1}a^{n-1-k}b^k\Big) &= \sum_{k=0}^{n-1}(a+(-b))a^{n-1-k}b^k\\
&= \sum_{k=0}^{n-1}(aa^{n-1-k}b^k + (-1)ba^{n-1-k}b^k)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=0}^{n-1} (a^{n-k}b^k + (-1)a^{n-1-k}bb^k) \\
&= \sum_{k=0}^{n-1} a^{n-k}b^k + \sum_{k=0}^{n-1} (-1)a^{n-1-k}b^{k+1} \\
&= \sum_{k=0}^{n-1} a^{n-k}b^k + \sum_{j=1}^{n} (-1)a^{n-j}b^j \\
&\qquad\qquad \text{(We replaced } k \text{ with } j-1 \text{ in the 2nd sum.)} \\
&= a^n + \left(\sum_{k=1}^{n-1} a^{n-k}b^k + \sum_{k=1}^{n-1} (-1)a^{n-k}b^k \right) + (-1)b^n \\
&\qquad\qquad \text{(We replaced } j \text{ with } k \text{ in the 2nd sum.)} \\
&= a^n + \left(\sum_{k=1}^{n-1} a^{n-k}b^k + (-1) \sum_{k=1}^{n-1} a^{n-k}b^k \right) - b^n \\
&= a^n - b^n. \qquad \blacksquare
\end{aligned}
$$

Definition A.17. A **field** is a commutative ring in which $1 \neq 0$, and all nonzero elements are invertible.

Notation. For two elements a, b in a field when $b \neq 0$ we set $a/b = \frac{a}{b} := ab^{-1}$.

Definition A.18. An **integral domain** is a commutative ring in which for all a, b

$$ab = 0 \implies a = 0 \text{ or } b = 0.$$

Theorem A.19. *Suppose R is an integral domain, and $a, b, c \in R$. Then the cancellation law holds for the multiplication of R, i.e.*

$$ac = bc,\ c \neq 0 \implies a = b.$$

Proof. We have $(a - b)c = 0$. Hence $a - b = 0$, since $c \neq 0$. $\blacksquare$

Theorem A.20. *Any field is an integral domain.*

Proof. If $ab = 0$ and $a \neq 0$ then

$$b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0. \qquad \blacksquare$$

Definition A.21. Suppose R is a ring, and $S \subset R$. We say S is a **subring** of R if S contains the identity of R, and for all $a, b \in S$ we have $-a, a + b, ab \in S$.

Proposition A.22. *Suppose R is a ring, and S is a subring of R. Then $0 \in S$, and S is itself a ring with the addition and multiplication inherited from R.*

Proof. We have $-1 \in S$, since $1 \in S$. Then $0 = 1 + (-1) \in S$. The associativity, commutativity, and distributivity laws are trivially satisfied in S, since they are satisfied in R. Also by definition S contains the opposite of each of its elements. Hence S is a ring. ■

Definition A.23. Suppose in a ring R we have $n = \overbrace{1 + 1 + \cdots + 1}^{n \text{ times}} = 0$ for some positive integer n. Then the smallest such n is called the **characteristic** of R. If this never happens we say that R has characteristic zero.

Theorem A.24. *The characteristic of an integral domain is either zero or a prime positive integer.*

Proof. If the conclusion does not hold, the characteristic is $n = pq$ for some positive integers $p, q < n$. But $pq = n = 0$ hence $p = 0$ or $q = 0$, which is in contradiction with the fact that n is the smallest positive integer with this property. ■

Remark. Suppose F is a field in which $n \neq 0$. Then we have

$$\overbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}^{n \text{ times}} = n\big(\frac{1}{n}\big) = n(1n^{-1}) = nn^{-1} = 1.$$

In other words, the n-th additive power of $\frac{1}{n}$ is 1.

***Exercise* A.25.** Suppose R is a ring, and S is a nonempty set. On the space of all functions from S into R we define the binary operations of pointwise addition and multiplication of functions, i.e. for two functions $f, g : S \to R$ and all $s \in S$ we define

$$(f + g)(s) := f(s) + g(s),$$
$$(fg)(s) := f(s)g(s).$$

Show that this space is a ring with these operations.

A.2 Matrices

Definition A.26. Let R be a ring, and $m, n \in \mathbb{N}$. An $m \times n$ **matrix** with entries in R is a function

$$A : \{(i, j) : i, j \in \mathbb{N},\ i \leq m,\ j \leq n\} \to R.$$

We denote by A_{ij} (or $A_{i,j}$) the value of A at (i, j), and call it the ij-th **entry** of A. The matrix A is usually denoted as a rectangular array of elements of R with m rows and n columns

$$A = [A_{ij}] = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}.$$

The $1 \times n$ matrix $[A_{i1}, \ldots, A_{in}]$ is called the i-th **row** of A, and is denoted by $A_{i,.}$. Also, the $m \times 1$ matrix

$$\begin{bmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{bmatrix}$$

is called the j-th **column** of A, and is denoted by $A_{.,j}$. A $1 \times n$ matrix is also called a **row vector**, and an $m \times 1$ matrix is also called a **column vector**. The set of $m \times n$ matrices with entries in R is denoted by $R^{m\times n}$. The **size** of a matrix $A \in R^{m\times n}$ is $m \times n$.

Remark. We know that R^n is the set of *ordered n-tuples* of elements of R. In order to make this precise, we can define R^n to be the set of functions

$$r : \{1, 2, \ldots, n\} \to R.$$

Then we denote by r_i the value of r at i, and we call it the i-th **component** of r. We will also denote r by the following familiar notation

$$r = (r_1, \ldots, r_n).$$

We can identify R^n with both $R^{1\times n}$ and $R^{n\times 1}$ via the maps

$$(r_1, \ldots, r_n) \mapsto [r_1, \ldots, r_n],$$
$$(r_1, \ldots, r_n) \mapsto \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}.$$

In particular, we always identify R with $R^{1\times 1}$. We also refer to the i,1-th entry of a column vector, or the 1,i-th entry of a row vector, as the i-th **component** of them. Furthermore, the operations that we are going to define on matrices can also be applied to the elements of R^n via the above identifications, and they have the same properties.

Remark. Note that as matrices are functions into R, it suffices to define them by specifying their ij-th entry for every i, j. Also, when we want to show that two matrices are equal, it is enough to check the equality of their ij-th entry for each i, j. The same things apply to the elements of R^n.

Definition A.27. Let R be a ring, and $m, n \in \mathbb{N}$. The $m \times n$ **zero matrix** is a matrix whose entries are all zero. We often denote the zero matrix simply by 0. A **square matrix** is a matrix for which $m = n$, i.e. a matrix that has the same number of rows and columns. The **(main) diagonal** of a square matrix A is the n-tuple $(A_{11}, A_{22}, \ldots, A_{nn}) \in R^n$. The entries A_{ii} are referred to as the diagonal entries of A. The square matrix A is called **upper triangular** if $A_{ij} = 0$ for $j < i$. In other words, the entries of A below its main diagonal are zero, so A has the form

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \ldots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}.$$

Similarly, a square matrix A is called **lower triangular** if $A_{ij} = 0$ for $j > i$. A **diagonal matrix** is a square matrix A for which $A_{ij} = 0$ when $i \neq j$, so it has the form

$$\begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}.$$

A special diagonal matrix is the $n \times n$ **identity matrix**, which is defined by

$$I_{ij} = (I_n)_{ij} := \begin{cases} 0 & i \neq j, \\ 1 & i = j. \end{cases}$$

Definition A.28. Let R be a ring, and $m, n \in \mathbb{N}$. The **addition** of two $m \times n$ matrices A, B with entries in R, is defined by

$$(A + B)_{ij} := A_{ij} + B_{ij}.$$

The **multiplication** of an $m \times n$ matrix A with an $n \times l$ matrix B is an $m \times l$ matrix AB, which is defined by

$$(AB)_{ij} := \sum_{k=1}^{n} A_{ik} B_{kj}.$$

The **scalar multiplication** of $r \in R$ and $A \in R^{m \times n}$ is defined by

$$(rA)_{ij} := rA_{ij}.$$

The **transpose** of an $m \times n$ matrix A is the $n \times m$ matrix $A^\intercal$ that satisfies

$$(A^\intercal)_{ij} := A_{ji}.$$

Notation. For a matrix A we set $-A := (-1)A$, so $(-A)_{ij} = -A_{ij}$. Also, for two $m \times n$ matrices A, B we set $A - B := A + (-B)$.

Remark. Let $A, B \in R^{m\times n}$ and $r \in R$. It is easy to show that for every i, j we have

$$\begin{aligned} (A+B)_{i,.} &= A_{i,.} + B_{i,.}, \qquad (rA)_{i,.} = rA_{i,.}, \qquad (A_{i,.})^\intercal = A^\intercal_{.,i}, \\ (A+B)_{.,j} &= A_{.,j} + B_{.,j}, \qquad (rA)_{.,j} = rA_{.,j}, \qquad (A_{.,j})^\intercal = A^\intercal_{j,.}. \end{aligned}$$

Remark. When the ring R is not commutative we can also define a scalar multiplication from the right by $(Ar)_{ij} := A_{ij}r$. This scalar multiplication has the properties described in the next theorem too. But we are mainly interested in commutative rings and do not pursue this direction here.

Theorem A.29. *Let R be a ring. Then for all $L \in R^{p\times m}$, $A, B, E \in R^{m\times n}$, $C \in R^{n\times l}$, and $r, s \in R$ we have*

(i) *The addition of matrices is associative and commutative, i.e.*

$$A + (B + E) = (A + B) + E, \qquad A + B = B + A.$$

(ii) *Let $0 \in R^{m\times n}$ be the zero matrix, then*

$$A + 0 = A, \qquad A + (-A) = 0.$$

(iii) *$1A = A$, and $I_m A = A = AI_n$.*

(iv) *We have*

$$L(A + B) = LA + LB, \qquad (A + B)C = AC + BC.$$

(v) *We have*

$$\begin{aligned} r(A + B) &= rA + rB, \qquad (r + s)A = rA + sA, \\ (rA)C &= r(AC), \qquad r(sA) = (rs)A. \end{aligned}$$

(vi) *If A or C is the zero matrix, then AC is the zero matrix. Also, if r is zero, or A is the zero matrix, then rA is the zero matrix.*

(vii) *We have*

$$(A + B)^\intercal = A^\intercal + B^\intercal, \qquad (rA)^\intercal = rA^\intercal, \qquad (A^\intercal)^\intercal = A.$$

(viii) *When R is a commutative ring, we also have*

$$(AC)^\intercal = C^\intercal A^\intercal,$$

and

$$A(rC) = r(AC), \qquad (rA)(sC) = (rs)(AC).$$

Proof. **(i)** For each i, j we have

$$\begin{aligned}\big(A + (B + E)\big)_{ij} &= A_{ij} + (B + E)_{ij} = A_{ij} + (B_{ij} + E_{ij}) \\ &= (A_{ij} + B_{ij}) + E_{ij} = (A + B)_{ij} + E_{ij} = \big((A + B) + E\big)_{ij}.\end{aligned}$$

The other one is similar.

(ii) This is similar to (i).

(iii) It is obvious that $1A = A$. For the second part we have

$$(I_m A)_{ij} = \sum_{k\le m} (I_m)_{ik} A_{kj} = 0A_{1j} + \cdots + 1A_{ij} + \cdots + 0A_{mj} = A_{ij}.$$

The other half is similar.

(iv) We have

$$\begin{aligned}\big((A + B)C\big)_{ij} &= \sum_{k\le n} (A + B)_{ik} C_{kj} = \sum_{k\le n} (A_{ik} + B_{ik}) C_{kj} \\ &= \sum_{k\le n} A_{ik} C_{kj} + \sum_{k\le n} B_{ik} C_{kj} = (AC)_{ij} + (BC)_{ij}.\end{aligned}$$

The other one is similar.

(v) We only prove $(rA)C = r(AC)$, the others can be proved similarly. We have

$$\begin{aligned}\big((rA)C\big)_{ij} &= \sum_{k\le n} (rA)_{ik} C_{kj} = \sum_{k\le n} r A_{ik} C_{kj} \\ &= r \sum_{k\le n} A_{ik} C_{kj} = r(AC)_{ij} = \big(r(AC)\big)_{ij}.\end{aligned}$$

(vi) These are all easy to show.

(vii) We have $\big((A^\intercal)^\intercal\big)_{ij} = (A^\intercal)_{ji} = A_{ij}$. Also

$$\big((rA)^\intercal\big)_{ij} = (rA)_{ji} = rA_{ji} = r(A^\intercal)_{ij} = (rA^\intercal)_{ij}.$$

The other one is similar.

(viii) We have

$$\begin{aligned}\big((AC)^{\intercal}\big)_{ij} &= (AC)_{ji} = \sum_{k\le n} A_{jk}C_{ki} \\ &= \sum_{k\le n} C_{ki}A_{jk} = \sum_{k\le n}(C^{\intercal})_{ik}(A^{\intercal})_{kj} = (C^{\intercal}A^{\intercal})_{ij}.\end{aligned}$$

Also

$$\begin{aligned}\big(A(rC)\big)_{ij} &= \sum_{k\le n} A_{ik}(rC)_{kj} = \sum_{k\le n} A_{ik}rC_{kj} \\ &= \sum_{k\le n} rA_{ik}C_{kj} = r\sum_{k\le n} A_{ik}C_{kj} = r(AC)_{ij} = \big(r(AC)\big)_{ij}.\end{aligned}$$

Hence

$$(rA)(sC) = s\big((rA)C\big) = s\big(r(AC)\big) = (sr)(AC) = (rs)(AC). \qquad \blacksquare$$

Theorem A.30. *The multiplication of matrices is associative, i.e. for any ring R and all matrices $A \in R^{p\times m}$, $B \in R^{m\times n}$, and $C \in R^{n\times l}$, we have*

$$(AB)C = A(BC).$$

Proof. We have

$$\begin{aligned}\big((AB)C\big)_{ij} &= \sum_{k=1}^{n}(AB)_{ik}C_{kj} = \sum_{k=1}^{n}\Big(\sum_{l=1}^{m} A_{il}B_{lk}\Big)C_{kj} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m} A_{il}B_{lk}C_{kj} = \sum_{l=1}^{m}\sum_{k=1}^{n} A_{il}B_{lk}C_{kj} \\ &= \sum_{l=1}^{m} A_{il}\Big(\sum_{k=1}^{n} B_{lk}C_{kj}\Big) = \sum_{l=1}^{m} A_{il}(BC)_{lj} = \big(A(BC)\big)_{ij}.\end{aligned}$$

$\blacksquare$

***Example* A.31.** Let $A = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$ and $B = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}$ be matrices in $R^{2\times 2}$, for some ring R. Then we have

$$AB = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix} \neq \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix} = BA.$$

Hence the multiplication of matrices is not in general commutative. This example also shows that the product of two nonzero matrices can be zero.

Theorem A.32. *Suppose R is a ring, and $A \in R^{m\times n}$, $C \in R^{n\times l}$. Then we have*

$$(AC)_{ij} = A_{i,\cdot}C_{\cdot,j}, \qquad (AC)_{\cdot,j} = AC_{\cdot,j}, \qquad (AC)_{i,\cdot} = A_{i,\cdot}C.$$

Remark. In other words, the j-th column of AC is the product of A and the j-th column of C. And the i-th row of AC is the product of the i-th row of A, and C.

Proof. Since $A_{i,.}$ and $C_{.,j}$ are respectively $1 \times n$ and $n \times 1$ matrices, their product is a 1×1 matrix, i.e. an element of R, and we have

$$(A_{i,.}C_{.,j})_{1,1} = \sum_{k \leq n} (A_{i,.})_{1,k}(C_{.,j})_{k,1} = \sum_{k \leq n} A_{i,k}C_{k,j} = (AC)_{ij}.$$

Similarly, $(AC)_{.,j}$ and $(AC)_{i,.}$ are respectively $m \times 1$ and $1 \times l$ matrices. Hence we have

$$\big((AC)_{.,j}\big)_{i,1} = (AC)_{i,j} = \sum_{k \leq n} A_{ik}C_{kj} = \sum_{k \leq n} A_{ik}(C_{.,j})_{k,1} = (AC_{.,j})_{i,1},$$
$$\big((AC)_{i,.}\big)_{1,j} = (AC)_{i,j} = \sum_{k \leq n} A_{ik}C_{kj} = \sum_{k \leq n} (A_{i,.})_{1,k}C_{kj} = (A_{i,.}C)_{1,j}. \qquad \blacksquare$$

Theorem A.33. *Suppose R is a ring, and $A \in R^{m \times n}$. Let $x = [x_1, \ldots, x_n]^\intercal \in R^{n \times 1}$ be a column vector, and let $y = [y_1, \ldots, y_m] \in R^{1 \times m}$ be a row vector. Then we have*

$$Ax = \sum_{j \leq n} A_{.,j}x_j, \qquad yA = \sum_{i \leq m} y_i A_{i,.}.$$

In particular, for $k \in \mathbb{N}$ and $j \leq k$ let $e_j \in R^{k \times 1}$ be the column vector whose components are all zero except for its j-th component which is one. Then

$$Ae_j = A_{.,j}, \qquad e_i^\intercal A = A_{i,.}.$$

Remark. We say that Ax is a *linear combination* of the columns of A, and yA is a linear combination of the rows of A.

Proof. We know that Ax and yA are respectively $m \times 1$ and $1 \times n$ matrices. Then we have

$$(Ax)_{i,1} = \sum_{j \leq n} A_{ij}x_j = \sum_{j \leq n} (A_{.,j})_{i,1}x_j = \Big(\sum_{j \leq n} A_{.,j}x_j\Big)_{i,1},$$
$$(yA)_{1,j} = \sum_{i \leq m} y_i A_{ij} = \sum_{i \leq m} y_i (A_{i,.})_{1,j} = \Big(\sum_{i \leq m} y_i A_{i,.}\Big)_{1,j}. \qquad \blacksquare$$

Theorem A.34. *Let R be a ring, then $R^{n \times n}$ is a ring with the addition and multiplication of matrices.*

Proof. This is a trivial consequence of the previous theorems. Just note that the zero and identity of this ring are respectively the $n \times n$ zero matrix and I_n. Also, the opposite of each matrix A is $-A$. $\blacksquare$

Remark. Let R be a ring. We know that $R^{n\times n}$ is a ring. Therefore a square matrix $A \in R^{n\times n}$ is called invertible if there is $B \in R^{n\times n}$ such that

$$AB = I_n = BA.$$

We also know that the inverse of an invertible matrix A is unique, and as usual we denote it by A^{-1}.

Proposition A.35. *Suppose R is a commutative ring, and $A \in R^{n\times n}$ is invertible. Then $A^\intercal$ is also invertible, and we have*

$$(A^\intercal)^{-1} = (A^{-1})^\intercal.$$

Proof. We have

$$(A^{-1})^\intercal A^\intercal = (AA^{-1})^\intercal = I^\intercal = I.$$

Similarly we have $A^\intercal(A^{-1})^\intercal = I$. Hence we get the desired due to the uniqueness of the inverse of a matrix. ■

A.3 Polynomials

Definition A.36. Let R be a commutative ring. The ring of **polynomials** with coefficients in R is the set of all sequences in R that terminate eventually, i.e.

$$R[x] := \{f : \mathbb{N} \cup \{0\} \to R \ : \ \text{there is } N \geq 0 \text{ such that } f_n = 0 \text{ for } n \geq N\}.$$

The elements f_n are called the **coefficients** of f. The **zero polynomial** is the polynomial whose coefficients are all zero. For a nonzero polynomial f, the largest nonnegative integer n for which $f_n \neq 0$ is called the **degree** of f and is denoted by

$$\deg f.$$

We also define the degree of the zero polynomial to be $-\infty$, with the understanding that for all $n \in \mathbb{Z}$ we have

$$-\infty < n, \qquad -\infty + n = -\infty.$$

The addition and multiplication of two polynomials f, g are defined as follows

$$(f+g)_n := f_n + g_n, \qquad (fg)_n := \sum_{k\leq n} f_k g_{n-k}.$$

Remark. Note that as polynomials are sequences of elements of R, it suffices to define them by specifying their n-th terms for every n. Also, when we want to show that two polynomials are equal, it is enough to check the equality of their n-th terms for each n.

Remark. It is easy to see that $R[x]$ is closed under the addition and multiplication defined above. Because $(f+g)_n$ and $(fg)_n$ are zero for all large values of n. In fact for nonnegative $n > \max\{\deg f, \deg g\}$ we have

$$(f+g)_n = f_n + g_n = 0 + 0 = 0,$$

and for nonnegative $n > \deg f + \deg g$ we have

$$\begin{aligned}(fg)_n &= \sum_{k=0}^{n} f_k g_{n-k} = \sum_{k\le \deg f} f_k g_{n-k} + \sum_{k>\deg f} f_k g_{n-k} \\ &= \sum_{k\le \deg f} f_k 0 + \sum_{k>\deg f} 0 g_{n-k} = 0.\end{aligned}$$

Note that for $k \le \deg f$ we have $n - k > \deg g$, hence $g_{n-k} = 0$. Also, note that as a result we have

$$\begin{aligned}&\deg(f+g) \le \max\{\deg f, \deg g\},\\ &\deg(fg) \le \deg f + \deg g.\end{aligned}$$

Theorem A.37. *Let R be a commutative ring. Then the ring of polynomials $R[x]$ is also a commutative ring.*

Proof. Let $f, g, h \in R[x]$. It is easy to check that addition of polynomials is commutative and associative, and the zero polynomial

$$0 := (0, 0, 0, \ldots, 0, \ldots)$$

is an additive identity. We have

$$\begin{aligned}(f+g)_n &= f_n + g_n = g_n + f_n = (g+f)_n,\\ (f+(g+h))_n &= f_n + (g+h)_n = f_n + (g_n + h_n)\\ &= (f_n + g_n) + h_n = (f+g)_n + h_n = ((f+g)+h)_n,\\ (f+0)_n &= f_n + 0_n = f_n + 0 = f_n.\end{aligned}$$

Also, for any f, the polynomial defined by $(-f)_n := -f_n$ is its opposite, since

$$(f + (-f))_n = f_n + (-f)_n = f_n + (-f_n) = 0 = 0_n.$$

Note that $-f$ is a polynomial as $(-f)_n = 0$ for $n > \deg f$.

The sequence

$$1 := (1, 0, 0, \ldots, 0, \ldots)$$

is the multiplicative identity of $R[x]$, since

$$(1f)_n = 1f_n + 0f_{n-1} + \cdots + 0f_1 + 0f_0 = f_n.$$

The commutativity of the multiplication is also easy to show

$$\begin{aligned}(fg)_n &= \sum_{k=0}^{n} f_k g_{n-k} = \sum_{k=0}^{n} g_{n-k} f_k \\ &= \sum_{l=n}^{0} g_l f_{n-l} = \sum_{l=0}^{n} g_l f_{n-l} = (gf)_n. \qquad (l := n-k)\end{aligned}$$

(Note that we have used both the commutativity of the multiplication of R, and the generalized commutativity of the addition of R.) To see that multiplication is distributive over addition, we have

$$\begin{aligned}(f(g+h))_n &= \sum_{k=0}^{n} f_k (g+h)_{n-k} = \sum_{k=0}^{n} f_k (g_{n-k} + h_{n-k}) \\ &= \sum_{k=0}^{n} (f_k g_{n-k} + f_k h_{n-k}) = \sum_{k=0}^{n} f_k g_{n-k} + \sum_{k=0}^{n} f_k h_{n-k} = (fg)_n + (fh)_n.\end{aligned}$$

It remains to show that multiplication is associative. Let

$$\Delta_{l,k} := \begin{cases} 1 & l \le k, \\ 0 & l > k, \end{cases}$$

where l, k are nonnegative integers and $0, 1 \in R$. Then we have

$$\begin{aligned}\big((fg)h\big)_n &= \sum_{k\le n} (fg)_k h_{n-k} = \sum_{k\le n} \Big[\sum_{l\le k} f_l g_{k-l}\Big] h_{n-k} \\ &= \sum_{k\le n}\sum_{l\le n} \Delta_{l,k} f_l g_{k-l} h_{n-k} = \sum_{l\le n}\sum_{k\le n} \Delta_{l,k} f_l g_{k-l} h_{n-k} \\ &= \sum_{l\le n} f_l \sum_{l\le k\le n} g_{k-l} h_{n-k} = \sum_{l\le n} f_l \sum_{0\le j\le n-l} g_j h_{n-l-j} \qquad (j := k-l) \\ &= \sum_{l\le n} f_l (gh)_{n-l} = \big(f(gh)\big)_n,\end{aligned}$$

as desired. ∎

Notation. We use the abbreviations

$$\begin{aligned} r &:= (r, 0, 0, \ldots, 0, \ldots), \\ x &:= (0, 1, 0, \ldots, 0, \ldots), \end{aligned}$$

where $r \in R$. Then we have

$$rx^n = x^n r = (0, 0, \ldots, \overset{n\text{-th}}{\underset{\downarrow}{r}}, \ldots, 0, \ldots).$$

Thus any polynomial can be written as

$$f = f_0 + f_1 x + \cdots + f_m x^m,$$

where $m = \deg f$ for nonzero f. Note that the coefficients of f in this representation are exactly the coefficients of f, so they are uniquely determined by f. We sometimes write $f(x)$ instead of f.

The zero polynomial, and polynomials of degree zero, are called **constant polynomials**. So constant polynomials are polynomials of degree less than one. Also, polynomials of degree one, two, and three are respectively called *linear*, *quadratic*, and *cubic* polynomials.

Remark. By identifying $r \in R$ with the constant polynomial $r \in R[x]$, we can consider R as a subring of $R[x]$.

Remark. Every polynomial defines a function on R by evaluation. That is for

$$f(x) = f_0 + f_1 x + \cdots + f_m x^m,$$

we have $f(r) := f_0 + f_1 r + \cdots + f_m r^m$, where $r \in R$. Note that $f(r)$ is uniquely determined by f and r, since the coefficients of f are uniquely determined by f. Also note that in the above, m need not be the $\deg f$. Because for $i > \deg f$ we have $f_i = 0$, hence the terms with $i > \deg f$ do not change the value of $f(r)$.

Theorem A.38. *For any two polynomials $f, g \in R[x]$ and all $r \in R$ we have*

$$(f + g)(r) = f(r) + g(r), \qquad (fg)(r) = f(r)g(r).$$

Proof. The proof is the same as of a more general version in Section A.5. ■

Remark. Note that the multiplication of polynomials is defined in a way that makes the above theorem valid, and this is one of the reasons behind its definition.

Definition A.39. Let f be a polynomial with coefficients in a ring R, and suppose $r \in R$. Then when $f(r) = 0$ we say r is a **root** of f.

Theorem A.40. *Suppose R is an integral domain. Then $R[x]$ is an integral domain too. In addition, for any two polynomials f, g we have*

$$\deg(fg) = \deg f + \deg g.$$

Proof. We have already shown that $R[x]$ is a commutative ring. If f or g, for example f, is zero, then $fg = 0$. Hence

$$\deg fg = -\infty = -\infty + \deg g = \deg f + \deg g.$$

Now suppose f, g are nonzero polynomials. Let $\deg f = n$ and $\deg g = m$. Then f_m, g_n are nonzero elements of R. Hence $(fg)_{m+n} = f_m g_n \neq 0$. Thus in particular fg is nonzero. Therefore $R[x]$ is an integral domain. Furthermore we have $\deg(fg) \geq m + n$. On the other hand we know that in general $\deg(fg) \leq m + n$, so $\deg(fg)$ is exactly $m + n$. ∎

Theorem A.41. *Suppose F is a field, and $f, g \in F[x]$ with $g \neq 0$. Then there are unique $q, r \in F[x]$ such that*

$$f = gq + r, \qquad \textit{where} \quad \deg r < \deg g.$$

Proof. If there is h such that $f = gh$, then we put $q = h$ and $r = 0$. Now suppose no such h exists. Then all the polynomials in

$$\{f - gp : p \in F[x]\}$$

are nonzero. Let q be an element of this set for which $f - gq$ has the least degree. This is possible due to the well ordering of nonnegative integers. Then set

$$r := f - gq.$$

We must show that $\deg r < \deg g$. Suppose to the contrary that $\deg r \geq \deg g$. Let

$$r(x) = r_n x^n + \cdots + r_0, \qquad g(x) = g_m x^m + \cdots + g_0.$$

Note that $r_n, g_m \neq 0$. Set $s(x) := r(x) - \frac{r_n}{g_m} x^{n-m} g(x)$. If $s = 0$ then we have

$$f(x) = g(x)\big(q(x) + \frac{r_n}{g_m} x^{n-m}\big),$$

which is in contradiction with our assumption. Thus $s \neq 0$ and we have $\deg s < \deg r$, since we have eliminated the x^n term. But this implies

$$f(x) = g(x)\big(q(x) + \frac{r_n}{g_m} x^{n-m}\big) + s(x),$$

which is in contradiction with the choice of q. Hence we have $\deg r < \deg g$ as desired.

For the uniqueness, suppose we have

$$gq_1 + r_1 = f = gq_2 + r_2.$$

Then $g(q_1 - q_2) = r_2 - r_1$. Since $g \neq 0$, we have $r_2 - r_1 = 0$ if and only if $q_1 - q_2 = 0$. Now if $r_1 \neq r_2$ and $q_1 \neq q_2$ then we get

$$\deg g + \deg(q_1 - q_2) = \deg(r_2 - r_1),$$

which is in contradiction with the fact that

$$\deg(r_2 - r_1) \leq \max\{\deg r_2, \deg r_1\} < \deg g.$$ ∎

Theorem A.42. *Suppose F is a field, $a \in F$, and $f \in F[x]$. Then*

$$f(x) = (x - a)g(x) + f(a).$$

Thus $f(a) = 0$ if and only if there is $g \in F[x]$ such that

$$f(x) = (x - a)g(x).$$

As a result, the number of distinct roots of a nonzero polynomial f is at most $\deg f$.

Proof. We divide f by $x - a$ to get $f(x) = (x - a)g(x) + r(x)$. But $\deg r < \deg(x - a) = 1$, so r is a constant. By evaluating the above equality at a we get $r = f(a)$. Thus

$$f(x) = (x - a)g(x) + f(a).$$

Now the second statement follows easily.

The last statement can be proved by induction on $\deg f$. Nonzero polynomials of degree zero are constant polynomials which have no root. Suppose the claim holds for all polynomials with degree less than $\deg f$. If f has no root, then there is nothing to prove. So let a be a root of f. Then $f = (x - a)g$. If $g = 0$ then $f = 0$, which is contrary to our assumption. So $g \neq 0$, and we have $\deg g = \deg f - 1$. Now if b is another root of f we must have $g(b) = 0$. But g has at most $\deg g$ distinct roots, hence f has at most $\deg g + 1 = \deg f$ distinct roots. ■

Remark. When the field F has infinitely many elements, the function defined by a polynomial uniquely determines the polynomial. Since if there are two distinct polynomials $f, g \in F[x]$ such that $f(a) = g(a)$ for all $a \in F$, then the nonzero polynomial $f - g$ has infinitely many roots, which is in contradiction with the above theorem.

Definition A.43. A field F is called **algebraically closed**, if every nonconstant polynomial with coefficients in F has at least one root in F.

Theorem A.44. *Let f be a polynomial with coefficients in an algebraically closed field F. Suppose* $\deg f = n \geq 1$. *Then there are (not necessarily distinct) elements $a_1, \ldots, a_n \in F$, and $c \in F - \{0\}$, such that*

$$f(x) = c(x - a_1) \cdots (x - a_n).$$

Proof. The proof is by induction on n. For $n = 1$ the claim holds trivially. Now suppose it also holds for polynomials of degree $n - 1$. Then we know that f has at least one root a_1. Hence there is a polynomial g of degree $n - 1$ such that

$$f(x) = (x - a_1)g(x).$$

Now by the induction hypothesis g has a factorization

$$g(x) = c(x - a_2) \cdots (x - a_n).$$

Thus we get the desired factorization for f. Finally note that if $c = 0$ then $f = 0$, which contradicts the fact that $\deg f \geq 1$. ■

Definition A.45. Suppose R is a commutative ring. We inductively define the ring of **polynomials in n variables** with coefficients in R to be the commutative ring

$$R[x_1, \ldots, x_n] := R[x_1, \ldots, x_{n-1}][x_n].$$

Remark. Intuitively we consider a polynomial p in n variables to be a formal sum of finitely many expressions of the form $rx_1^{m_1} \cdots x_n^{m_n}$, where $r \in R$. Then we can collect all terms in which x_n has power m_n, and factor $x_n^{m_n}$ out, to get $q(x_1, \ldots, x_{n-1})x_n^{m_n}$, where q is (what we intuitively consider) a polynomial in $n-1$ variables. Thus we can write p as a sum of terms of this form, with different powers of x_n. In other words, p is a polynomial in x_n whose coefficients are polynomials in $n - 1$ variables. Continuing inductively we can see that our intuitive notion of polynomials in n variables is the same notion described rigorously in the above definition.

Remark. When R is an integral domain, $R[x_1]$ is also an integral domain. Hence $R[x_1, x_2] = R[x_1][x_2]$ is an integral domain too. By an easy induction it follows that for any n, $R[x_1, \ldots, x_n]$ is also an integral domain.

Remark. Every polynomial in n variables defines a function on R^n by evaluation. Let $f(x_1, \ldots, x_n)$ be a polynomial in n variables. Then by definition we have

$$f(x_1, \ldots, x_n) = \\ f_0(x_1, \ldots, x_{n-1}) + f_1(x_1, \ldots, x_{n-1})x_n + \cdots + f_m(x_1, \ldots, x_{n-1})x_n^m,$$

where f_j's are polynomials in $n - 1$ variables which are uniquely determined by f. Then we can inductively define the value of f at $(a_1, \ldots, a_n) \in R^n$ to be

$$f(a_1, \ldots, a_n) := \\ f_0(a_1, \ldots, a_{n-1}) + f_1(a_1, \ldots, a_{n-1})a_n + \cdots + f_m(a_1, \ldots, a_{n-1})a_n^m.$$

It can also be proved inductively that the value of f at a point is uniquely determined by f and that point, since the same is true for each f_j, and f_j's are uniquely determined by f.

Theorem A.46. *Suppose R is a commutative ring. Then every polynomial $f \in R[x_1, \ldots, x_n]$ can be written as a sum of finitely many **monomials**, i.e.*

$$f(x_1, \ldots, x_n) = \sum_{m_1 \leq k_1} \cdots \sum_{m_n \leq k_n} r_{m_1 \ldots m_n} x_1^{m_1} \cdots x_n^{m_n},$$

where $r_{m_1 \dots m_n} \in R$ and $k_1, \dots, k_n$ are nonnegative integers. Furthermore, this representation of f is unique, and for every $(a_1, \dots, a_n) \in R^n$ we have

$$f(a_1, \dots, a_n) = \sum_{m_1 \le k_1} \cdots \sum_{m_n \le k_n} r_{m_1 \dots m_n} a_1^{m_1} \cdots a_n^{m_n}.$$

Proof. The proof is by induction on n. The case of $n = 1$ is obvious. So suppose the theorem is true for $n-1$. First note that the monomials are actually polynomials in n variables, since if $r x_1^{m_1} \cdots x_{n-1}^{m_{n-1}} \in R[x_1, \dots, x_{n-1}]$ then by definition we have $r x_1^{m_1} \cdots x_{n-1}^{m_{n-1}} x_n^{m_n} \in R[x_1, \dots, x_n]$. Now we know that

$$f(x_1, \dots, x_n) = \\ f_0(x_1, \dots, x_{n-1}) + f_1(x_1, \dots, x_{n-1}) x_n + \cdots + f_m(x_1, \dots, x_{n-1}) x_n^m,$$

where f_j's are polynomials in $n-1$ variables. By the induction hypothesis each f_j can be written as a sum of finitely many monomials in $n-1$ variables. If we substitute those expansions into the above formula for f, and multiply them by x_n^j, then we get an expansion of f into a sum of finitely many monomials in n variables.

Now let us prove the second statement. We have

$$\begin{aligned} f(x_1, \dots, x_n) &= \sum_{m_1 \le k_1} \cdots \sum_{m_n \le k_n} r_{m_1 \dots m_n} x_1^{m_1} \cdots x_n^{m_n} \\ &= \sum_{j=0}^{k_n} \Big(\sum_{m_1 \le k_1} \cdots \sum_{m_{n-1} \le k_{n-1}} r_{m_1 \dots m_{n-1} j} x_1^{m_1} \cdots x_{n-1}^{m_{n-1}} \Big) x_n^j \\ &\qquad\qquad \text{(We replaced } m_n \text{ with } j.\text{)} \\ &= \sum_{j \le k_n} f_j(x_1, \dots, x_{n-1}) x_n^j, \end{aligned}$$

where $f_j := \sum_{m_1 \le k_1} \cdots \sum_{m_{n-1} \le k_{n-1}} r_{m_1 \dots m_{n-1} j} x_1^{m_1} \cdots x_{n-1}^{m_{n-1}}$ is a polynomial in $n-1$ variables. Hence by the induction hypothesis, $r_{m_1 \dots m_{n-1} j}$'s are uniquely determined by f_j, and we have

$$f_j(a_1, \dots, a_{n-1}) = \sum_{m_1 \le k_1} \cdots \sum_{m_{n-1} \le k_{n-1}} r_{m_1 \dots m_{n-1} j} a_1^{m_1} \cdots a_{n-1}^{m_{n-1}}.$$

On the other hand, f_j's must be the coefficients of $f \in R[x_1, \dots, x_{n-1}][x_n]$, since the coefficients of f are uniquely determined by f. Thus $r_{m_1 \dots m_n}$'s are uniquely

determined by f, and we have

$$\begin{aligned} f(a_1,\dots,a_n) &= \sum_{j\le k_n} f_j(a_1,\dots,a_{n-1})a_n^j \\ &= \sum_{j\le k_n}\Big(\sum_{m_1\le k_1}\cdots\sum_{m_{n-1}\le k_{n-1}} r_{m_1\dots m_{n-1}j}a_1^{m_1}\cdots a_{n-1}^{m_{n-1}}\Big)a_n^j \\ &= \sum_{m_1\le k_1}\cdots\sum_{m_n\le k_n} r_{m_1\dots m_n}a_1^{m_1}\cdots a_n^{m_n}. \quad \text{(We replaced } j \text{ with } m_n.\text{)} \end{aligned}$$

Note that we have used Theorem A.68 several times, to change the order of summations. ■

Remark. When we expand the polynomial $f \in R[x_1,\dots,x_n]$ into a sum of monomials as described in the above theorem, the elements $r_{m_1\dots m_n} \in R$ are referred to as the **coefficients** of f. Note that we sometimes consider the coefficients of f to be polynomials of $n-1$ variables, but usually we consider the coefficients of f to be $r_{m_1\dots m_n}$'s.

Definition A.47. The **elementary symmetric polynomials** in n variables $x_1,\dots,x_n$ are

$$\begin{aligned} s_1 &:= x_1 + x_2 + \cdots + x_n, \\ s_2 &:= x_1x_2 + x_1x_3 + \cdots + x_2x_3 + \cdots + x_{n-1}x_n, \\ &\vdots \\ s_k &:= \sum_{i_1=1}^{n-k+1}\sum_{i_2=i_1+1}^{n-k+2}\cdots\sum_{i_k=i_{k-1}+1}^{n} x_{i_1}x_{i_2}\cdots x_{i_k}, \\ &\vdots \\ s_n &:= x_1x_2\cdots x_n. \end{aligned}$$

Theorem A.48. *Suppose R is a commutative ring and $a_1,\dots,a_n$ are (not necessarily distinct) elements of R. Then*

$$(x-a_1)\cdots(x-a_n) = x^n - b_1x^{n-1} + b_2x^{n-2} - \cdots + (-1)^n b_n,$$

where $b_k = s_k(a_1,\dots,a_n)$.

Proof. The proof is by induction on n. The case of $n=1$ is trivial. Suppose the claim holds for $n-1$. Let $\tilde{s}_1,\dots,\tilde{s}_n$ be the elementary symmetric polynomials in $n-1$ variables. Then we have

$$(x-a_1)\cdots(x-a_{n-1}) = x^{n-1} - c_1x^{n-2} + \cdots + (-1)^{n-1}c_{n-1},$$

where $c_k = \tilde{s}_k(a_1, \ldots, a_{n-1})$. Now we have

$$\begin{aligned}(x-a_1)\cdots(x-a_n) &= (x^{n-1} - c_1x^{n-2} + \cdots + (-1)^{n-1}c_{n-1})(x-a_n)\\ &= x^n - (c_1+a_n)x^{n-1} + (c_2 + c_1a_n)x^{n-2}\\ &\quad - (c_3 + c_2a_n)x^{n-3} + \cdots + (-1)^n c_{n-1}a_n.\end{aligned}$$

But for $1 < k < n$ we have

$$\begin{aligned}b_k &= \sum_{i_1=1}^{n-k+1}\sum_{i_2=i_1+1}^{n-k+2}\cdots\sum_{i_k=i_{k-1}+1}^{n} a_{i_1}a_{i_2}\cdots a_{i_k}\\ &= \sum_{i_1=1}^{n-k}\sum_{i_2=i_1+1}^{n-k+1}\cdots\sum_{i_{k-1}=i_{k-2}+1}^{n-2}\sum_{i_k=i_{k-1}+1}^{n-1} a_{i_1}a_{i_2}\cdots a_{i_k}\\ &\quad + \Big(\sum_{i_1=1}^{n-k+1}\sum_{i_2=i_1+1}^{n-k+2}\cdots\sum_{i_{k-1}=i_{k-2}+1}^{n-1} a_{i_1}a_{i_2}\cdots a_{i_{k-1}}\Big)a_n.\end{aligned}$$

Hence $b_k = c_k + c_{k-1}a_n$. It is also obvious that $b_1 = c_1 + a_n$, and $b_n = c_{n-1}a_n$. Therefore we get the desired formula. ∎

A.4 Field of Fractions

Let R be an integral domain. Let $X = \{(a,b) : a, b \in R,\ b \neq 0\}$. We define the relation $\sim$ on X as follows

$$(a,b) \sim (c,d) \text{ if } ad = bc.$$

Proposition A.49. *$\sim$ is an equivalence relation.*

Proof. We only check the transitivity of $\sim$ and leave the rest as an exercise. Suppose $(a,b) \sim (c,d)$ and $(c,d) \sim (e,f)$. Then we have $ad = bc$ and $cf = de$. If we multiply the first equation by f we get $adf = bcf = bde$. Hence $d(af - be) = 0$. But $d \neq 0$ and R is an integral domain, thus $af = be$. Therefore $(a,b) \sim (e,f)$. ∎

Let F be the set of equivalence classes of $\sim$. We denote the equivalence class of (a,b) by $[a,b]$. We want to formally consider $[a,b]$ to be the fraction $\frac{a}{b}$. Note that $[a,b] = [c,d]$ if and only if $ad = bc$, which is formally equivalent to the equality of the fractions $\frac{a}{b}, \frac{c}{d}$. Now we define the addition and multiplication on F as follows

$$\begin{aligned}[a,b] + [c,d] &:= [ad + bc, bd],\\ [a,b][c,d] &:= [ac, bd].\end{aligned}$$

Notice that $bd \neq 0$, since $b, d \neq 0$ and R is an integral domain. It must be checked that these operations are well-defined, i.e. they do not depend on the particular representatives of the equivalence classes $[a, b]$ and $[c, d]$. Note that these operations are the same as formally adding and multiplying the fractions $\frac{a}{b}, \frac{c}{d}$.

Theorem A.50. *F equipped with the above operations is a field.*

Proof. We leave the details of the proof as an exercise. Only note that the zero and identity of F are respectively

$$[0, 1], \qquad [1, 1].$$

Also the opposite and the inverse (if $a \neq 0$) of an element $[a, b]$ are respectively

$$-[a, b] = [-a, b], \qquad [a, b]^{-1} = [b, a]. \qquad \blacksquare$$

Note that the map $a \mapsto [a, 1]$ from R into F preserves addition and multiplication, i.e.

$$[a + b, 1] = [a, 1] + [b, 1], \qquad [ab, 1] = [a, 1][b, 1].$$

We abuse the notation and denote $[a, 1]$ simply by a, and we consider R to be a subring of F. Now for every element $[a, b] \in F$ we have

$$[a, b] = [a, 1][1, b] = [a, 1][b, 1]^{-1} = ab^{-1},$$

i.e. $[a, b]$ is the fraction $\frac{a}{b}$.

Definition A.51. F is called the **field of fractions** of R.

***Example* A.52.** $\mathbb{Q}$ is the field of fractions of $\mathbb{Z}$.

***Example* A.53.** Let F be a field. The field of fractions of the ring of polynomials $F[x_1, \ldots, x_n]$ is denoted by $F(x_1, \ldots, x_n)$, and is called the field of **rational functions in n-variables** over F. The elements of $F(x_1, \ldots, x_n)$ are of the form

$$\frac{f(x_1, \ldots, x_n)}{g(x_1, \ldots, x_n)},$$

where $f, g \in F[x_1, \ldots, x_n]$ are polynomials in n-variables. These elements are called rational functions. For $(a_1, \ldots, a_n) \in F$ if $g(a_1, \ldots, a_n) \neq 0$, we can compute the value of the rational function $\frac{f}{g}$ at $(a_1, \ldots, a_n)$ to be

$$\frac{f(a_1, \ldots, a_n)}{g(a_1, \ldots, a_n)} \in F.$$

A.5 Algebras

Definition A.54. Let R be a commutative ring. Let A be a ring. We say A is an **algebra** over R if there is an operation

$$\begin{array}{c} R \times A \longrightarrow A \\ (r, \alpha) \mapsto r\alpha \end{array},$$

called **scalar multiplication**, that satisfies

(i) For all $r \in R$ and $\alpha, \beta \in A$ we have

$$r(\alpha + \beta) = r\alpha + r\beta,$$

and

$$(r\alpha)\beta = r(\alpha\beta) = \alpha(r\beta).$$

(ii) For all $r, s \in R$ and $\alpha \in A$ we have

$$(r + s)\alpha = r\alpha + s\alpha,$$

and

$$r(s\alpha) = (rs)\alpha.$$

(iii) For all $\alpha \in A$ we have

$$1\alpha = \alpha,$$

where 1 is the identity of R.

Remark. Note that the second equation of (i) informally means that the elements of R commute with the elements of A. If the ring A is itself a commutative ring, we say A is a commutative algebra over R.

***Example* A.55.** Suppose R is a commutative ring. Then $R^{n\times n}$ equipped with the standard addition, multiplication, and scalar multiplication of matrices, is an algebra over R. Also, $R[x]$ with its usual addition and multiplication is a commutative algebra over R. The scalar multiplication of $r \in R$ and $f(x) = f_0 + \cdots + f_m x^m \in R[x]$ is

$$rf(x) := rf_0 + \cdots + rf_m x^m.$$

Since this is the same as the product of the constant polynomial r, and f, the scalar multiplication satisfies all the required properties trivially.

***Example* A.56.** Any commutative ring R is an algebra over itself. The scalar multiplication of $r \in R$ and $a \in R$ is just their product ra.

***Exercise* A.57.** Suppose R is a commutative ring, and S is a nonempty set. Show that the space of all functions from S into R with pointwise addition, multiplication, and scalar multiplication of functions, is a commutative algebra over R. Remember that for two functions $f, g : S \to R$, and all $s \in S$ and $r \in R$, these operations are defined as

$$\begin{aligned}(f+g)(s) &:= f(s) + g(s),\\ (fg)(s) &:= f(s)g(s),\\ (rf)(s) &:= rf(s).\end{aligned}$$

Proposition A.58. *Let A be an algebra over the commutative ring R. Then for all $r, s \in R$ and $\alpha, \beta \in A$ we have*

(i) $0\alpha = 0$.
(ii) $r0 = 0$.
(iii) $(-1)\alpha = -\alpha$, *where* $-1 \in R$.
(iv) $(r\alpha)(s\beta) = (rs)(\alpha\beta)$.

Proof. **(i)** We have

$$0\alpha + 0 = 0\alpha = (0+0)\alpha = 0\alpha + 0\alpha.$$

Hence by cancellation law we get $0 = 0\alpha$.

(ii) We have

$$r0 + 0 = r0 = r(0+0) = r0 + r0.$$

Thus again by cancellation law we get $0 = r0$.

(iii) We have

$$\alpha + (-1)\alpha = 1\alpha + (-1)\alpha = (1 + (-1))\alpha = 0\alpha = 0.$$

Therefore the result follows from the uniqueness of the inverse of α.

(iv) We have

$$(r\alpha)(s\beta) = r(\alpha(s\beta)) = r(s(\alpha\beta)) = (rs)(\alpha\beta).$$

■

Remark. We can also easily show by induction that

$$\begin{aligned}r(\alpha_1 + \cdots + \alpha_k) &= r\alpha_1 + \cdots + r\alpha_k,\\ (r_1 + \cdots + r_k)\alpha &= r_1\alpha + \cdots + r_k\alpha,\end{aligned}$$

for $r, r_i \in R$, and $\alpha, \alpha_i \in A$.

Definition A.59. Suppose A is an algebra over the commutative ring R. Then every polynomial $f \in R[x]$ defines a function from A into A by evaluation. That is for

$$f(x) = f_0 + f_1 x + \cdots + f_m x^m$$

with $f_i \in R$, and for $\alpha \in A$, we define

$$f(\alpha) := f_0 I + f_1 \alpha + \cdots + f_m \alpha^m,$$

where $I \in A$ is the identity of A. We say that the element $f(\alpha)$ is a *polynomial in* α.

Remark. Note that $f(\alpha)$ is uniquely determined by f and α, since the coefficients of f are uniquely determined by f. Also note that in the above definition, m need not be the $\deg f$. Because for $i > \deg f$ we have $f_i = 0$, hence the terms with $i > \deg f$ do not change the value of $f(\alpha)$.

Theorem A.60. *Suppose A is an algebra over the commutative ring R. Then for any two polynomials $f, g \in R[x]$ and all $\alpha \in A$ we have*

$$(f+g)(\alpha) = f(\alpha) + g(\alpha), \qquad (fg)(\alpha) = f(\alpha)g(\alpha).$$

As a result, $f(\alpha)$ and $g(\alpha)$ always commute.

Remark. The significance of this theorem is that the addition and multiplication of polynomials convert to the addition and multiplication of A via the map $f \mapsto f(\alpha)$.

Proof. Let $m = \deg f$, and $n = \deg g$. Then $f_i = 0$ for $i > m$, and $g_j = 0$ for $j > n$. Let $l = \max\{m, n\}$, then $\deg(f+g) \le l$. Now we have

$$\begin{aligned} f(\alpha) + g(\alpha) &= \sum_{i \le m} f_i \alpha^i + \sum_{i \le n} g_i \alpha^i = \sum_{i \le l} f_i \alpha^i + \sum_{i \le l} g_i \alpha^i \\ &= \sum_{i \le l} (f_i \alpha^i + g_i \alpha^i) = \sum_{i \le l} (f_i + g_i)\alpha^i = (f+g)(\alpha). \end{aligned}$$

Next, remember that $\deg(fg) \le m+n$. For $0 \le k \le m+n$ let $a = \max\{0, k-m\}$, and $b = \min\{n, k\}$. Then by the generalized distributivity and Theorem A.68 we have

$$\begin{aligned} f(\alpha)g(\alpha) &= \Big(\sum_{i \le m} f_i \alpha^i\Big)\Big(\sum_{j \le n} g_j \alpha^j\Big) = \sum_{i \le m} \sum_{j \le n} (f_i \alpha^i)(g_j \alpha^j) \\ &= \sum_{i \le m} \sum_{j \le n} (f_i g_j)(\alpha^i \alpha^j) = \sum_{i \le m} \sum_{j \le n} (f_i g_j)\alpha^{i+j} \\ &= \sum_{k \le m+n} \sum_{i+j=k} (f_i g_j)\alpha^k = \sum_{k \le m+n} \Big(\sum_{a \le i \le b} f_i g_{k-i} \Big)\alpha^k \\ &= \sum_{k \le m+n} \Big(\sum_{i \le k} f_i g_{k-i}\Big)\alpha^k = \sum_{k \le m+n} (fg)_k \alpha^k = (fg)(\alpha). \end{aligned}$$

In the last line of the above formula we used the fact that $\sum_{a\leq i\leq b} f_i g_{k-i} = \sum_{i\leq k} f_i g_{k-i}$. The reason is that for $i > b \geq n$ we have $f_i = 0$, and for $i < a \leq k-m$ we have $g_{k-i} = 0$ since $k - i > m$.

Finally, to prove the last statement of the theorem, note that we have

$$f(\alpha)g(\alpha) = (fg)(\alpha) = (gf)(\alpha) = g(\alpha)f(\alpha),$$

since $R[x]$ is a commutative ring. ■

Remark. The above theorem is in particular true for polynomials of square matrices with entries in a commutative ring.

Remark. As a consequence of the above theorem, we can easily show by induction that if $p_1, \ldots, p_k \in R[x]$ then we have

$$\begin{aligned}(p_1 + \cdots + p_k)(\alpha) &= p_1(\alpha) + \cdots + p_k(\alpha),\\ (p_1 p_2 \cdots p_k)(\alpha) &= p_1(\alpha)p_2(\alpha)\cdots p_k(\alpha).\end{aligned}$$

A.6 Binary Operations

Definition A.61. A **binary operation** on a set S is a function

$$\star : S \times S \to S.$$

For two elements $a, b \in S$, we usually write $a \star b$ instead of $\star(a, b)$.

Notation. In the rest of this section we assume that S is a set, and $\star$ is a binary operation on S.

Definition A.62. A binary operation $\star$ on S is called **associative** if for all $a, b, c \in S$ we have

$$a \star (b \star c) = (a \star b) \star c,$$

and it is called **commutative** if for all $a, b \in S$ we have

$$a \star b = b \star a.$$

We also say two elements a, b commute if $a \star b = b \star a$. An element $e \in S$ is an **identity** if for all $a \in S$ we have

$$a \star e = a = e \star a.$$

Finally, a subset $A \subset S$ is said to be **closed** under $\star$ if for all $a, b \in A$ we have $a \star b \in A$.

Theorem A.63. *A binary operation has at most one identity.*

Proof. If there exist two identities e, e' we have $e' = e' \star e = e$. ■

Definition A.64. Suppose $\star$ is a binary operation on S with identity e, and $a, b \in S$. We say a is **invertible,** and b is an **inverse** of a, if

$$a \star b = e = b \star a.$$

Theorem A.65. *Suppose $\star$ is an associative operation with identity e, and $a, b \in S$.*

(i) *If a has an inverse, its inverse is unique, and we denote it by a^{-1}.*

(ii) *If a is invertible, then a^{-1} is also invertible and*

$$(a^{-1})^{-1} = a.$$

(iii) *If a and b are invertible, then $a \star b$ is also invertible and we have*

$$(a \star b)^{-1} = b^{-1} \star a^{-1}.$$

Proof. **(i)** Suppose that a has two inverses denoted by b and c, then

$$b = b \star e = b \star (a \star c) = (b \star a) \star c = e \star c = c.$$

(ii) This follows from the definition of a^{-1} and the uniqueness of inverse.

(iii) First note that

$$\begin{aligned}(b^{-1} \star a^{-1}) \star (a \star b) &= b^{-1} \star (a^{-1} \star (a \star b)) \\ &= b^{-1} \star ((a^{-1} \star a) \star b) = b^{-1} \star (e \star b) = b^{-1} \star b = e.\end{aligned}$$

Similarly $(a \star b) \star (b^{-1} \star a^{-1}) = e$. Therefore $(a \star b)$ is invertible. Now the result follows from uniqueness of inverse and the above computations. ■

Cancellation Law. *Suppose $\star$ is an associative operation with identity e. Let $a, b, c \in S$, and suppose a is invertible. Then we have*

$$\begin{aligned}a \star b = a \star c &\implies b = c, \\ b \star a = c \star a &\implies b = c.\end{aligned}$$

Proof. We have

$$\begin{aligned}a \star b = a \star c &\implies a^{-1} \star (a \star b) = a^{-1} \star (a \star c) \\ &\implies (a^{-1} \star a) \star b = (a^{-1} \star a) \star c \\ &\implies e \star b = e \star c \implies b = c.\end{aligned}$$

The other one is similar. ■

Notation. Suppose $\star$ is a binary operation on S, and $a_1, \ldots, a_n \in S$. We inductively define the *standard product* of n elements of S to be

$$\prod_{i=1}^{1} a_i := a_1, \ \ldots \ \prod_{i=1}^{n} a_i := \Big(\prod_{i=1}^{n-1} a_i\Big) \star a_n.$$

We also inductively define the set of *all possible products* of n elements of S as follows

$$\begin{aligned} &P(a_1) := \{a_1\}, \\ &\vdots \\ &P(a_1, \ldots, a_n) := \\ &\qquad \{b \star c : b \in P(a_1, \ldots, a_k) \ , \ c \in P(a_{k+1}, \ldots, a_n) \text{ for some } 1 \le k < n\}. \end{aligned}$$

Note that the n elements $a_1, \ldots, a_n$ have an order, and can have repetitions; so the above notions are actually assigned to the ordered n-tuple $(a_1, \ldots, a_n)$.

Remark. When the binary operation is denoted by $a \cdot b$, or simply by ab, then we keep using the notation $\prod$ for the standard product of several elements. But when the binary operation is denoted by $a + b$, we use the notation $\sum$ instead of $\prod$, and we use the term "sum" instead of "product".

Generalized Associativity. *Suppose $\star$ is an associative binary operation, and $a_1, \ldots, a_n \in S$. Then $P(a_1, \ldots, a_n)$ has exactly one element $\prod_{i=1}^{n} a_i$.*

Proof. It is easy to show by induction that $\prod_{i=1}^{n} a_i \in P(a_1, \ldots, a_n)$. For uniqueness we can argue inductively as follows. When $n = 1$, $P(a_1)$ has one element. Suppose the theorem is true for $1 \le k < n$, i.e. $P(b_1, \ldots, b_k)$ has exactly one element for any $b_1, \ldots, b_k$. Then the elements of $P(a_1, \ldots, a_n)$ are of the form $b \star c$ where $b \in P(a_1, \ldots, a_k)$ and $c \in P(a_{k+1}, \ldots, a_n)$. Thus by induction hypothesis we have

$$b = \prod_{i=1}^{k} a_i, \qquad c = \prod_{i=k+1}^{n} a_i.$$

When $k = n-1$ we have $c = a_n$, hence $b \star c = \big(\prod_{i=1}^{n-1} a_i\big) \star a_n = \prod_{i=1}^{n} a_i$. Otherwise we have

$$\begin{aligned} \Big(\prod_{i=1}^{k} a_i\Big) \star \Big(\prod_{i=k+1}^{n} a_i\Big) &= \Big(\prod_{i=1}^{k} a_i\Big) \star \Big[\Big(\prod_{i=k+1}^{n-1} a_i\Big) \star a_n\Big] \\ &= \Big[\Big(\prod_{i=1}^{k} a_i\Big) \star \Big(\prod_{i=k+1}^{n-1} a_i\Big)\Big] \star a_n \end{aligned}$$

$$= \Big(\prod_{i=1}^{n-1} a_i\Big) \star a_n = \prod_{i=1}^{n} a_i. \qquad \blacksquare$$

Remark. The above theorem means that when $\star$ is associative, the value of the product of $a_1, \ldots, a_n$ is independent of the arrangement of the parentheses. In this case we sometimes denote $\prod_{i=1}^{n} a_i$ by $a_1 \star \cdots \star a_n$.

***Exercise* A.66.** Suppose $\star$ is an associative binary operation on S, and $a_1, \ldots, a_n \in S$ are invertible. Show that $\prod_{i=1}^{n} a_i$ is also invertible and we have

$$(a_1 \star \cdots \star a_n)^{-1} = a_n^{-1} \star \cdots \star a_1^{-1}.$$

Definition A.67. A **permutation** is a one-to-one and onto map from $\{1, \ldots, n\}$ to itself, for some positive integer n. We denote the set of all permutations on $\{1, \ldots, n\}$ by S_n.

Generalized Commutativity. *Suppose $\star$ is an associative and commutative binary operation, and $a_1, \ldots, a_n \in S$. Then for every permutation $\sigma \in S_n$ we have*

$$a_{\sigma(1)} \star \cdots \star a_{\sigma(n)} = a_1 \star \cdots \star a_n.$$

Proof. We use induction on n. The case $n = 1$ is obvious, so suppose the conclusion holds for all permutations on $\{1, \ldots, n-1\}$. Now for the induction step we have

$$\prod_{i=1}^{n} a_{\sigma(i)} = \prod_{i=1}^{n-1} a_{\sigma(i)} \star a_{\sigma(n)}.$$

Suppose $\sigma(j) = n$. Then

$$\begin{aligned} \prod_{i=1}^{n-1} a_{\sigma(i)} &= \prod_{i=1}^{j-1} a_{\sigma(i)} \star a_n \star \prod_{i=j+1}^{n-1} a_{\sigma(i)} \\ &= \prod_{i=1}^{j-1} a_{\sigma(i)} \star \prod_{i=j+1}^{n-1} a_{\sigma(i)} \star a_n = \prod_{i=1, i\neq j}^{n-1} a_{\sigma(i)} \star a_n. \end{aligned}$$

Let $\hat{\sigma}$ be the permutation on $\{1, \ldots, n-1\}$ defined by

$$\hat{\sigma}(i) = \begin{cases} \sigma(i) & i < j \\ \sigma(i+1) & i \geq j. \end{cases}$$

Then we have

$$\begin{aligned}\prod_{i=1}^{n} a_{\sigma(i)} &= \prod_{i=1,i\neq j}^{n-1} a_{\sigma(i)} \star a_n \star a_{\sigma(n)} \\ &= \prod_{i=1,i\neq j}^{n-1} a_{\sigma(i)} \star a_{\sigma(n)} \star a_n \\ &= \prod_{i=1}^{j-1} a_{\sigma(i)} \star \prod_{i=j+1}^{n} a_{\sigma(i)} \star a_n \\ &= \prod_{i=1}^{n-1} a_{\hat{\sigma}(i)} \star a_n = \prod_{i=1}^{n-1} a_i \star a_n = \prod_{i=1}^{n} a_i. \end{aligned}$$

■

Remark. The above theorem means that the order of $a_1, \ldots, a_n$ does not affect the value of $a_1 \star \cdots \star a_n$, when the operation $\star$ is both commutative and associative.

Remark. Suppose $\star$ is an associative and commutative binary operation. Sometimes we want to compute the product of several elements of S that do not have an order, or are not ordered linearly. Suppose I is a finite set, and $\mathfrak{a} : I \to S$ is a function. We want to compute the product of all the elements $\mathfrak{a}(\alpha)$ for $\alpha \in I$. Note that $\mathfrak{a}$ need not be one-to-one, so some of the $\mathfrak{a}(\alpha)$'s might be the same. In other words, we may have repetition of factors in our product. Suppose I has n elements, and $f : \{1, \ldots, n\} \to I$ is a one-to-one and onto function. Let us denote $\mathfrak{a}(f(k))$ by a_k. Now we define

$$\prod_{\alpha \in I} \mathfrak{a}(\alpha) := \prod_{k \leq n} a_k.$$

We have to check that this definition is independent of f. Let $g : \{1, \ldots, n\} \to I$ be another one-to-one and onto function. Then

$$\sigma := f^{-1} \circ g : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

is also one-to-one and onto, i.e. it is a permutation. Let us denote $\mathfrak{a}(g(k))$ by $\tilde{a}_k$. Hence by the above theorem we have

$$\prod_{k \leq n} \tilde{a}_k = \prod_{k \leq n} \mathfrak{a}(g(k)) = \prod_{k \leq n} \mathfrak{a}(f(\sigma(k))) = \prod_{k \leq n} a_{\sigma(k)} = \prod_{k \leq n} a_k = \prod_{\alpha \in I} \mathfrak{a}(\alpha),$$

as desired. The element $\prod_{\alpha \in I} \mathfrak{a}(\alpha)$ is sometimes called the *unordered product* of the elements $\mathfrak{a}(\alpha)$ for $\alpha \in I$. A particular case is when $A \subset S$ is a finite set, and $\mathfrak{a} : A \to S$ is the inclusion map. Then the unordered product of the elements of A is $\prod_{a \in A} a := \prod_{a \in A} \mathfrak{a}(a)$.

Theorem A.68. *Suppose $\star$ is an associative and commutative binary operation, and $a_i, b_i, a_{ij} \in S$ for $i \le n$, $j \le m$. Then*

$$\prod_{i=1}^{n} a_i \star \prod_{i=1}^{n} b_i = \prod_{i=1}^{n} (a_i \star b_i),$$

and

$$\prod_{j=1}^{m}\prod_{i=1}^{n} a_{ij} = \prod_{i=1}^{n}\prod_{j=1}^{m} a_{ij} = \prod_{k=2}^{m+n}\prod_{i+j=k} a_{ij}.$$

Here $\prod_{i+j=k} a_{ij}$ is a shorthand notation for $\prod_{i=r}^{l} a_{i,k-i}$, where $r = \max\{1, k-m\}$, and $l = \min\{n, k-1\}$.

Proof. The proofs are by induction on n. We only write the induction step. For the first equality we have

$$\begin{aligned}
\prod_{i=1}^{n+1} a_i \star \prod_{i=1}^{n+1} b_i &= \Big(\prod_{i=1}^{n} a_i\Big) \star a_{n+1} \star \Big(\prod_{i=1}^{n} b_i\Big) \star b_{n+1} \\
&= \Big(\prod_{i=1}^{n} a_i\Big) \star \Big(\prod_{i=1}^{n} b_i\Big) \star a_{n+1} \star b_{n+1} \\
&= \Big(\prod_{i=1}^{n} (a_i \star b_i)\Big) \star (a_{n+1} \star b_{n+1}) = \prod_{i=1}^{n+1} (a_i \star b_i).
\end{aligned}$$

For the second equality we have

$$\begin{aligned}
\prod_{i=1}^{n+1}\prod_{j=1}^{m} a_{ij} &= \Big(\prod_{i=1}^{n}\prod_{j=1}^{m} a_{ij}\Big) \star \prod_{j=1}^{m} a_{n+1,j} \\
&= \Big(\prod_{j=1}^{m}\prod_{i=1}^{n} a_{ij}\Big) \star \prod_{j=1}^{m} a_{n+1,j} \\
&= \prod_{j=1}^{m}\Big[\Big(\prod_{i=1}^{n} a_{ij}\Big) \star a_{n+1,j}\Big] = \prod_{j=1}^{m}\prod_{i=1}^{n+1} a_{ij}.
\end{aligned}$$

For the third equality let $r := \max\{1, k-m\}$, $l := \min\{n, k-1\}$, and $L := \min\{n+1, k-1\}$. Note that for $k \le n+1$ we have $l = L$, and for $k \ge n+2$ we

have $l = n$ and $L = n + 1$. Now we have

$$\begin{aligned}
\prod_{i=1}^{n+1}\prod_{j=1}^{m} a_{ij} &= \Big(\prod_{i=1}^{n}\prod_{j=1}^{m} a_{ij}\Big) \star \prod_{j=1}^{m} a_{n+1,j} = \Big(\prod_{k=2}^{m+n}\prod_{i=r}^{l} a_{i,k-i}\Big) \star \prod_{j=1}^{m} a_{n+1,j} \\
&= \Big(\prod_{k=2}^{m+n}\prod_{i=r}^{l} a_{i,k-i}\Big) \star \Big(\prod_{k=n+2}^{m+n} a_{n+1,k-n-1}\Big) \star a_{n+1,m} \\
&= \Big(\prod_{k=2}^{n+1}\prod_{i=r}^{l} a_{i,k-i}\Big) \star \Big(\prod_{k=n+2}^{m+n}\prod_{i=r}^{n} a_{i,k-i}\Big) \star \Big(\prod_{k=n+2}^{m+n} a_{n+1,k-n-1}\Big) \star a_{n+1,m} \\
&= \Big(\prod_{k=2}^{n+1}\prod_{i=r}^{l} a_{i,k-i}\Big) \star \Big(\prod_{k=n+2}^{m+n}\Big[\Big(\prod_{i=r}^{n} a_{i,k-i}\Big) \star a_{n+1,k-n-1}\Big]\Big) \star a_{n+1,m} \\
&= \Big(\prod_{k=2}^{n+1}\prod_{i=r}^{L} a_{i,k-i}\Big) \star \Big(\prod_{k=n+2}^{m+n}\prod_{i=r}^{n+1} a_{i,k-i}\Big) \star a_{n+1,m} \\
&= \Big(\prod_{k=2}^{m+n}\prod_{i=r}^{L} a_{i,k-i}\Big) \star a_{n+1,m} = \prod_{k=2}^{m+n+1}\prod_{i=r}^{L} a_{i,k-i} = \prod_{k=2}^{m+n+1}\prod_{i+j=k} a_{ij}.
\end{aligned}$$

■

Notation. We often denote by

$$\prod_{i,j} a_{ij}$$

the common element of the second and third equalities in the above theorem.

Remark. An interesting consequence of the above theorem is that it enables us to compute the following *telescoping product*. Let $\star$ be an associative and commutative binary operation on S. Then for invertible elements $a_1, \ldots, a_n \in S$ we have

$$\begin{aligned}
\prod_{i=2}^{n}(a_i \star a_{i-1}^{-1}) &= \prod_{i=2}^{n} a_i \star \prod_{i=2}^{n} a_{i-1}^{-1} = \prod_{i=2}^{n-1} a_i \star a_n \star \Big(\prod_{i=n}^{2} a_{i-1}\Big)^{-1} \\
&= a_n \star \prod_{i=2}^{n-1} a_i \star \Big(\prod_{i=2}^{n} a_{i-1}\Big)^{-1} = a_n \star \prod_{i=2}^{n-1} a_i \star \Big(\prod_{i=1}^{n-1} a_i\Big)^{-1} \\
&\qquad\qquad \text{(We changed } i-1 \text{ to } i \text{ in the last term.)} \\
&= a_n \star \prod_{i=2}^{n-1} a_i \star \Big(a_1 \star \prod_{i=2}^{n-1} a_i\Big)^{-1} = a_n \star \prod_{i=2}^{n-1} a_i \star \Big(\prod_{i=2}^{n-1} a_i\Big)^{-1} \star a_1^{-1} \\
&= a_n \star a_1^{-1}.
\end{aligned}$$

Note that here we have used both the generalized associativity and the generalized commutativity.

Definition A.69. Suppose $\star$ is an associative binary operation on S and $a \in S$. We define the **powers** a^n for positive integer n to be $\prod_{i=1}^{n} a$, i.e.

$$a^1 := a, \quad \dots \quad a^n := a^{n-1} \star a.$$

If there exists an identity e we define $a^0 := e$, and if a has an inverse a^{-1} we define

$$a^{-n} := (a^{-1})^n.$$

Theorem A.70. *Suppose $\star$ is an associative binary operation on S. Then for all $a, b \in S$ we have*

(i) *If a commutes with b, then a^n commutes with b^m, for all $m, n \geq 0$. If one or both of a, b are invertible, we can allow n and/or m to be negative too.*

(ii) *If a is invertible, then a^n is also invertible for all $n \in \mathbb{Z}$, and*

$$(a^n)^{-1} = a^{-n} = (a^{-1})^n.$$

(iii) *$a^n \star a^m = a^{n+m}$ for all $m, n \geq 0$. If a is invertible, we can allow m, n to be negative too.*

(iv) *$(a^n)^m = a^{nm}$ for all $m, n \geq 0$. If a is invertible, we can allow m, n to be negative too.*

(v) *If a, b commute, we have $a^n \star b^n = (a \star b)^n$ for all $n \geq 0$. If a, b are invertible, we can allow n to be negative too.*

Proof. The proof is the same as of Theorem A.13. ■

Definition A.71. Suppose $\star$ and $+$ are binary operations on S. We say $\star$ is **distributive** over $+$ if for all $a, b, c \in S$ we have

$$a \star (b + c) = (a \star b) + (a \star c),$$
$$(b + c) \star a = (b \star a) + (c \star a).$$

Generalized Distributivity. *Suppose $\star, +$ are associative binary operations on S, and $\star$ is distributive over $+$. Then for all $a_{ij} \in S$ and $n_j \in \mathbb{N}$ we have*

$$\Big(\sum_{i_1=1}^{n_1} a_{i_1 1}\Big) \star \cdots \star \Big(\sum_{i_k=1}^{n_k} a_{i_k k}\Big) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} (a_{i_1 1} \star \cdots \star a_{i_k k}).$$

Proof. First suppose $k = 2$ and n_1 or n_2 is 1. We have to show that

$$a_{11} \star \Big(\sum_{i=1}^{n} a_{i2}\Big) = \sum_{i=1}^{n} a_{11} \star a_{i2}, \qquad \Big(\sum_{i=1}^{n} a_{i1}\Big) \star a_{12} = \sum_{i=1}^{n} a_{i1} \star a_{12}.$$

This can be easily done by induction on n. Now for the general case, we proceed by induction on k. Suppose the conclusion holds for $k-1$. Then we have

$$\begin{aligned}
&\Big(\sum_{i_1=1}^{n_1} a_{i_1 1}\Big) \star \Big(\sum_{i_2=1}^{n_2} a_{i_2 2}\Big) \star \cdots \star \Big(\sum_{i_k=1}^{n_k} a_{i_k k}\Big) \\
&\qquad = \sum_{i_1=1}^{n_1} \left[a_{i_1 1} \star \Big(\sum_{i_2=1}^{n_2} a_{i_2 2}\Big) \star \cdots \star \Big(\sum_{i_k=1}^{n_k} a_{i_k k}\Big) \right] \\
&\qquad = \sum_{i_1=1}^{n_1} \left[\Big(\sum_{i_2=1}^{n_2} a_{i_1 1} \star a_{i_2 2}\Big) \star \cdots \star \Big(\sum_{i_k=1}^{n_k} a_{i_k k}\Big) \right] \\
&\qquad = \sum_{i_1=1}^{n_1} \left[\sum_{i_2=1}^{n_2} \cdots \sum_{i_k=1}^{n_k} \big((a_{i_1 1} \star a_{i_2 2}) \cdots \star a_{i_k k}\big) \right] \\
&\qquad = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_k=1}^{n_k} (a_{i_1 1} \star a_{i_2 2} \cdots \star a_{i_k k}).
\end{aligned}$$

■

Remark. Note that we do not need the commutativity of $+$ in the above theorem. To make this clear let us look at the special case

$$\begin{aligned}
(a_{11} + a_{21})(a_{12} + a_{22}) &= a_{11}(a_{12} + a_{22}) + a_{21}(a_{12} + a_{22}) \\
&= a_{11}a_{12} + a_{11}a_{22} + a_{21}a_{12} + a_{21}a_{22} \\
&= \sum_{i_2=1}^{2} a_{11}a_{i_2 2} + \sum_{i_2=1}^{2} a_{21}a_{i_2 2} \\
&= \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} a_{i_1 1}a_{i_2 2}.
\end{aligned}$$

Definition A.72. A **Group** is a nonempty set G with an associative binary operation with identity, in which all elements are invertible.

Remark. As we showed in the last section, the identity is unique, and the inverse of any element is unique. Furthermore, the cancellation law and the generalized associativity hold in a group.

***Example* A.73.** Remember that a permutation is a one-to-one and onto map from $\{1, \ldots, n\}$ to itself, for some positive integer n. Permutations form a group under the operation of composition of maps. We denote this group by S_n, and call it the **symmetric group** on n letters.

Definition A.74. A group whose operation is commutative, is called an **abelian group**.

Remark. We will use the multiplicative notation for a general group, and the additive notation for an abelian group. In particular, the powers of an element a will be shown by a^n in a general group, and by na in an abelian one.

Definition A.75. The **order** of an element a in a group G with identity e, is the smallest positive integer n such that $a^n = e$. If no such number exists, we say a has infinite order.

Theorem A.76. *In a finite group G every element has finite order, and its order is a divisor of the number of elements of G.*

Proof. Let $a \in G$, and let n be the cardinality of G. Consider the sequence of $n+1$ elements $e, a, a^2, \ldots, a^n$. Then two of them must be the same, so $a^i = a^j$ for some $j > i$. Hence $a^{j-i} = e$.

Now let m be the order of a. Then $e, a, \ldots, a^{m-1}$ are distinct. If $m = n$ then we have $m|n$. Suppose $m < n$. Let

$$b_1 \in G - \{e, a, \ldots, a^{m-1}\}.$$

Then $b_1, b_1a, \ldots, b_1a^{m-1}$ are also distinct, and they are different than $e, a, \ldots, a^{m-1}$. The reason is that $b_1a^i = b_1a^j$ implies $a^{j-i} = e$, and $b_1a^i = a^j$ implies $b_1 = a^{j-i}$. By continuing this way we will have distinct elements

$$\begin{aligned} &e, a, \ldots, a^{m-1}, \\ &b_1, b_1a, \ldots, b_1a^{m-1}, \\ &\vdots \\ &b_k, b_ka, \ldots, b_ka^{m-1}. \end{aligned}$$

This process must stop at some step, since G is a finite set. Hence $m(k+1) = n$. ■

Appendix B

Factorization

B.1 Euclidean Domains

Definition B.1. Let R be an integral domain, and $a, b \in R$. If there exists $r \in R$ such that $b = ra$, then we say a **divides** b, or a is a **divisor** of b, or b is a **multiple** of a; and we write $a|b$.

An element $u \in R$ is called a **unit** if $u|1$. Two elements a, b are **associates** if $a|b$ and $b|a$.

Proposition B.2. *Suppose R is an integral domain and $a, b, c \in R$. Then*

(i) *The units of R are exactly the invertible elements of R.*
(ii) *For all $a \in R$ we have $1|a$, $a|a$, and $a|0$.*
(iii) *a, b are associates if and only if there is a unit $u \in R$ such that $a = ub$.*
(iv) *$a|b$ and $b|c$ implies $a|c$.*
(v) *$a|b$ implies $a|bc$, and $ac|bc$.*
(vi) *$a|b$ and $a|c$ imply $a|(b + c)$.*

Proof. **(i)** $a|1$ if and only if there is $r \in R$ such that $ra = 1$, i.e. a is invertible with $a^{-1} = r$.

(ii) We have $a1 = 1a = a$, and $0a = 0$.

(iii) If a, b are associates, then $a|b$ and $b|a$. Thus there are $r, s \in R$ such that $ra = b$ and $sb = a$. If $a = 0$ then $b = 0$, hence $a = 1b$. So suppose $a \neq 0$. Now we have

$$1a = a = sb = sra.$$

Thus $sr = 1$, since R is an integral domain. Therefore s is a unit, and $a = sb$. Conversely, if $a = ub$ for some unit u, then $b = u^{-1}a$. Hence $a|b$ and $b|a$.

(iv) There are $r, s \in R$ such that $b = ra$ and $c = sb$. Hence $c = sra$, so $a|c$.

(v) There is $r \in R$ such that $b = ra$. Thus $bc = rac = rca$. So $a|bc$, and $ac|bc$.

(vi) There are $r, s \in R$ such that $b = ra$ and $c = sa$. Therefore $b + c = (r + s)a$, so $a|(b + c)$. ■

***Exercise* B.3.** Suppose R is an integral domain and $a, b, c \in R$. Show that if a, b are associates, and b, c are associates, then a, c are associates too.

Proposition B.4. *Let $a, b \in \mathbb{Z}$, and suppose $b \neq 0$. If $a|b$ then $|a| \leq |b|$.*

Proof. There is $c \in \mathbb{Z}$ such that $b = ca$. Hence $|b| = |ca| = |c||a|$. But $c \neq 0$ since $b \neq 0$. Thus $|c| \geq 1$. Therefore $|b| = |c||a| \geq |a|$. ■

Definition B.5. A **Euclidean domain** is an integral domain R on which there exists a **degree function**

$$d : R - \{0\} \to \{n \in \mathbb{Z} : n \geq 0\},$$

such that

(i) For nonzero $a, b \in R$ if $a|b$ then $d(a) \leq d(b)$.

(ii) **(Division Algorithm)** For all $a \in R$ and all nonzero $b \in R$, there are $q, r \in R$ such that

$$a = bq + r, \qquad \text{where either } r = 0, \text{ or } d(r) < d(b).$$

Here q is called the **quotient** and r is called the **remainder**.

Theorem B.6. *$\mathbb{Z}$ is a Euclidean domain with the degree function $d(n) = |n|$. Furthermore, the quotient and the remainder in its division algorithm are unique, if we require the remainder to be nonnegative.*

Proof. Let $n, m \in \mathbb{Z}$, and assume $m \neq 0$. We prove the existence of a division algorithm

$$n = mq + r,$$

by induction on $|n|$. At first, we do not impose any sign restriction on the remainder. If $|n| < |m|$ we can simply put $q = 0$ and $r = n$. If $|n| = |m|$ then we have $n = \pm m$, so we can put $q = \pm 1$ and $r = 0$.

Now suppose $|n| > |m|$, and the conclusion holds for all integers with absolute value less than $|n|$. Then we have

$$|n - m| \text{ or } |n + m| < |n|,$$

where we have to choose $\pm$ according to the signs of n, m. Thus by the induction hypothesis we have

$$n \pm m = mq + r,$$

where r is either 0 or $|r| < |m|$. Therefore we have $n = m(q \mp 1) + r$ as desired.

It is easy to see that we can take r to be nonnegative. Since if $r < 0$ we have

$$n = mq + r = m(q \mp 1) + (\pm m + r).$$

Now as we have $|r| < |m|$, we can choose $\pm$ according to the sign of m, so that

$$0 < \pm m + r < |m|.$$

Next to prove the uniqueness, suppose to the contrary that we have

$$mq_1 + r_1 = n = mq_2 + r_2,$$

where $0 \leq r_i < |m|$. Then we have $m(q_1 - q_2) = r_2 - r_1$. If $r_2 - r_1 \neq 0$ then $q_1 - q_2 \neq 0$, and we have $m|r_2 - r_1$. But $|r_2 - r_1| < |m|$ and we have arrived at a contradiction. Hence $r_1 = r_2$, and therefore $q_1 = q_2$. ■

***Example* B.7.** As we saw in the last chapter, when F is a field, $F[x]$ is a Euclidean domain with the degree function $d(f) = \deg f$. In addition for nonzero polynomials f, g we have

$$g|f \implies \deg g \leq \deg f.$$

Since if $f = gh$, then $\deg f = \deg g + \deg h$. As a result, the units of $F[x]$ are precisely the nonzero constant polynomials. Because if $u \in F[x]$ is unit then it is invertible, hence it is nonzero. Thus we have $\deg u \leq \deg 1 = 0$, since $u|1$. Therefore $\deg u = 0$, and u is constant. On the other hand, nonzero constant polynomials are units, since they are invertible as F is a field.

Proposition B.8. *Suppose R is a Euclidean domain, and $a, b \in R$. If b is nonzero and it is not a unit, then for all nonzero a we have*

$$d(a) < d(ab).$$

Proof. We know that $d(b) \leq d(ab)$, since $b|ab$. Suppose to the contrary that $d(ab) = d(a)$. Then we have $a = abq + r$ where $d(r) < d(ab) = d(a)$ or $r = 0$. If $r = 0$ we have $a(1 - bq) = 0$ which is a contradiction, since b is not a unit and a is nonzero. Hence we have $r \neq 0$, and $d(r) < d(a)$. Now

$$a(1 - bq) = a - abq = r.$$

Thus $a|r$ and we must have $d(a) \leq d(r)$. This contradiction proves the result. ■

Theorem B.9. *Let $n, b \in \mathbb{N}$, and suppose $b > 1$. Then there is a unique integer $m \geq 0$, and unique integers $0 \leq r_0, r_1, \ldots, r_m < b$ with $r_m \neq 0$, such that*

$$n = r_m b^m + r_{m-1} b^{m-1} + \cdots + r_1 b + r_0.$$

Remark. This is the *base b representation* of the number n. It is sometimes denoted by $n = (r_m \dots r_0)_b$.

Proof. First we prove the existence. The proof is by induction on n. If $1 \leq n < b$, then we can put $m = 0$ and $r_0 = n > 0$. Suppose the conclusion holds for all $k < n$. We can assume that $n \geq b$. Now we have $n = bq + r$ where $0 \leq r < b$. But q cannot be nonpositive since then we would have $n \leq r < b$, contrary to our assumption. Thus we must have $q > 0$. Then $n = bq + r > q + r \geq q$. Therefore by the induction hypothesis we have

$$q = s_m b^m + \cdots + s_1 b + s_0,$$

for some $0 \leq s_i < b$, with $s_m \neq 0$. Then we have

$$n = bq + r = s_m b^{m+1} + \cdots + s_1 b^2 + s_0 b + r,$$

as desired.

For the uniqueness, again the proof is by induction on n. If $1 \leq n < b$, then the representation is obviously unique. Because $b^i \geq b > n$, so we cannot have $m > 0$. Suppose the uniqueness holds for all positive integers less than n. We can assume that $n \geq b$. Suppose that

$$s_k b^k + \cdots + s_1 b + s_0 = n = r_m b^m + \cdots + r_1 b + r_0,$$

where $0 \leq r_i, s_j < b$, and $r_m, s_k \neq 0$. Then $m, k > 0$, since $n \geq b$. Now we have

$$(s_k b^{k-1} + \cdots + s_1) b + s_0 = n = (r_m b^{m-1} + \cdots + r_1) b + r_0.$$

Therefore r_0, s_0 are the remainder in the division of n by b. Hence $r_0 = s_0$. Since the quotient is also unique, we have

$$s_k b^{k-1} + \cdots + s_1 = r_m b^{m-1} + \cdots + r_1.$$

But, in the above paragraph we showed that when $n \geq b$ then the quotient is a positive integer strictly less than n. Therefore by the induction hypothesis we have $m - 1 = k - 1$, and $r_i = s_i$ for $1 \leq i \leq m$. Hence $m = k$, and the base b representation of n is unique. ∎

Theorem B.10. *Let F be a field. Let $f, g \in F[x]$ be nonzero polynomials, and suppose $\deg g \geq 1$. Then there is a unique integer $m \geq 0$, and unique polynomials $r_0, r_1, \dots, r_m \in F[x]$ with $r_m \neq 0$, such that*

$$f = r_m g^m + r_{m-1} g^{m-1} + \cdots + r_1 g + r_0,$$

where $\deg r_i < \deg g$ for each i.

Proof. First we prove the existence. The proof is by induction on $\deg f$. If $0 \leq \deg f < \deg g$, then we can put $m = 0$ and $r_0 = f \neq 0$. Suppose the conclusion holds for all polynomials with degree less than $\deg f$. We can assume that $\deg f \geq \deg g$. Now we have $f = gq + r$ where $\deg r < \deg g$. Then $gq = f - r$, so $\deg q + \deg g = \deg(f - r)$. But $\deg r < \deg f$. Thus $\deg(f - r) = \deg f$, since subtracting r does not change the coefficient of the highest degree term in f. Hence

$$\deg q = \deg f - \deg g < \deg f.$$

Also $q \neq 0$, since otherwise we would have $f = r$, which implies $\deg f = \deg r < \deg g$. Therefore by the induction hypothesis we have

$$q = s_m g^m + \cdots + s_1 g + s_0,$$

for some $s_i \in F[x]$, with $s_m \neq 0$, and $\deg s_i < \deg g$. Then we have

$$f = gq + r = s_m g^{m+1} + \cdots + s_1 g^2 + s_0 g + r,$$

as desired.

For the uniqueness, again the proof is by induction on $\deg f$. If $0 \leq \deg f < \deg g$, then the representation is unique. Because if

$$f = r_m g^m + r_{m-1} g^{m-1} + \cdots + r_1 g + r_0$$

for some $m > 0$, where $\deg r_i < \deg g$ and $r_m \neq 0$, then

$$\begin{aligned} \deg(r_{m-1} g^{m-1} + \cdots + r_1 g + r_0) &\leq \max_{j \leq m-1} \big(\deg(r_j g^j)\big) \\ &= \max_{j \leq m-1} (\deg r_j + j \deg g) \\ &< m \deg g \leq \deg(r_m g^m). \end{aligned}$$

Hence $\deg f = \deg(r_m g^m) \geq \deg g$, which is contrary to our assumption. Thus $m = 0$, and therefore $r_0 = f$.

Now suppose the uniqueness holds for all polynomials with degree less than $\deg f$. We can assume that $\deg f \geq \deg g$. Suppose that

$$s_k g^k + \cdots + s_1 g + s_0 = f = r_m g^m + \cdots + r_1 g + r_0,$$

where $\deg r_i, \deg s_j < \deg g$, and $r_m, s_k \neq 0$. Then $m, k > 0$, since $\deg f \geq \deg g > \deg r_0, \deg s_0$. Now we have

$$(s_k g^{k-1} + \cdots + s_1) g + s_0 = f = (r_m g^{m-1} + \cdots + r_1) g + r_0.$$

Therefore r_0, s_0 are the remainder in the division of f by g. Hence $r_0 = s_0$, since the remainder and the quotient in the division of polynomials are unique. Thus we also have

$$s_k g^{k-1} + \cdots + s_1 = r_m g^{m-1} + \cdots + r_1.$$

But, in the first part of this proof we showed that when $\deg f \geq \deg g$ then the quotient is a nonzero polynomial whose degree is strictly less than $\deg f$. Therefore by the induction hypothesis we have $m - 1 = k - 1$, and $r_i = s_i$ for $1 \leq i \leq m$. Hence $m = k$, and the representation of f is unique. ■

Remark. As a special case of the above theorem we set $g(x) = x - a$, for some $a \in F$. Then $\deg r_i < \deg g = 1$, so each r_i is a constant. Hence for every $f \in F[x]$ there are unique $a_1, \ldots, a_m \in F$, where $a_m \neq 0$ when f is not constant, such that

$$f(x) = a_m(x - a)^m + \cdots + a_1(x - a) + f(a).$$

The fact that $r_0 = f(a)$ follows easily by evaluating both sides of the identity at a. Note that we can allow f to be zero by setting each $a_i = 0$.

B.2 Principal Ideal Domains

Definition B.11. Let R be a commutative ring. An **ideal** is a nonempty subset $I \subset R$ such that

(i) For all $a, b \in I$ we have $a + b \in I$.
(ii) For all $a \in I$ and $r \in R$ we have $ra \in I$.

***Example* B.12.** Let $a_1, \ldots, a_n \in R$. Then it is easy to see that the set

$$(a_1, \ldots, a_n) := \{r_1 a_1 + \cdots + r_n a_n : \text{ for all } r_i \in R\}$$

is an ideal. It is called the *ideal generated by* $a_1, \ldots, a_n$.

Definition B.13. An ideal I is called **principal** if

$$I = (a) = \{ra : r \in R\}$$

for some $a \in R$.

Proposition B.14. *Suppose R is an integral domain, and $a, b \in R$. Then*

(i) *$(a) = R$ if and only if a is a unit.*
(ii) *$(b) \subset (a)$ if and only if $a|b$, if and only if $b \in (a)$.*
(iii) *$(b) = (a)$ if and only if a, b are associates.*

Proof. **(i)** If a is unit then $1 = a^{-1}a \in (a)$. Hence for all $r \in R$ we have $r = r1 \in (a)$. Thus $R = (a)$. Conversely, if $R = (a)$ then $1 \in (a)$. Hence there is $s \in R$ such that $sa = 1$, i.e. a is a unit.

(ii) If $a|b$ then $b = sa$ for some $s \in R$. Hence $b \in (a)$. Thus for all $r \in R$ we have $rb \in (a)$, i.e. $(b) \subset (a)$. Conversely, if $(b) \subset (a)$ then $b = 1b \in (b) \subset (a)$. Therefore $b = sa$ for some $s \in R$, hence $a|b$.

(iii) $(b) = (a)$ is equivalent to $(b) \subset (a)$ and $(a) \subset (b)$. Thus $(b) = (a)$ if and only if $a|b$ and $b|a$, i.e. if and only if a, b are associates. ■

Definition B.15. An integral domain is called a **principal ideal domain (PID)**, if all of its ideals are principal.

Theorem B.16. *Every Euclidean domain is a PID.*

Proof. Let I be an ideal in R. Let $a \in I$ be a nonzero element that has the least degree among all nonzero elements of I. This is possible due to the well ordering of nonnegative integers. We claim that $I = (a)$. It is obvious that $(a) \subset I$. For the other inclusion, let b be an arbitrary element of I. We have $b = aq + r$ where either $r = 0$ or $d(r) < d(a)$. Since $r = b + (-q)a \in I$, we cannot have $d(r) < d(a)$. Thus $r = 0$, and therefore $b \in (a)$. Hence $I \subset (a)$ as desired. ■

Definition B.17. A **greatest common divisor (g.c.d)** of two elements a, b in an integral domain R, is an element $c \in R$ such that

(i) $c|a$ and $c|b$.

(ii) If $r \in R$ is a common divisor of a, b, i.e. $r|a$ and $r|b$, then $r|c$.

Remark. Note that when $R = \mathbb{Z}$, and c is a greatest common divisor of nonzero $a, b \in \mathbb{Z}$, then for any common divisor r of a, b we have $|r| \leq |c|$, since $r|c$. Note that $c \neq 0$, because 0 cannot be a divisor of a nonzero element.

Proposition B.18. *Suppose R is an integral domain, and $a, b \in R$. Then any two greatest common divisors of a, b are associates.*

Proof. Suppose c_1, c_2 are greatest common divisors of a, b. Then c_1, c_2 are both common divisors of a, b. Hence we must have $c_1|c_2$ and $c_2|c_1$, since c_1, c_2 are both greatest common divisors of a, b. Thus c_1, c_2 are associates. ■

Euclidean Algorithm. *Suppose R is a Euclidean domain, and $a, b \in R$ are nonzero. Consider the following sequence of divisions*

$$\begin{aligned} a &= bq_0 + r_0, \\ b &= r_0q_1 + r_1, \\ r_0 &= r_1q_2 + r_2, \\ &\vdots \end{aligned}$$

Then after finitely many divisions the remainder becomes zero, i.e. we have

$$r_{n-1} = r_n q_{n+1}.$$

Also, r_n is a greatest common divisor of a, b.

Proof. Note that $d(b) > d(r_0) > d(r_1) > \cdots$. Hence the process must stop at some point, since the d-values are nonnegative integers. Therefore the division must be impossible at some step, which means that the remainder in the last step is zero. Now let us show that the remainder from one step before the last step, i.e. r_n, is a g.c.d of a, b. First note that $r_n | r_{n-1}$. Thus from $r_{n-2} = r_{n-1}q_n + r_n$ we see that $r_n | r_{n-2}$. If we continue inductively we get $r_n | b$ and $r_n | a$. Hence r_n is a common divisor of a, b. Next suppose $c|a$ and $c|b$. Then $c|a - bq_0 = r_0$. Again we can show inductively that $c|r_n$. Thus r_n is a g.c.d of a, b. ■

Theorem B.19. *Suppose R is a PID, and $a, b \in R$ are nonzero. Then a, b have a greatest common divisor $c \in R$. Furthermore we have*

$$c = ra + sb,$$

for some $r, s \in R$.

Proof. The ideal generated by a, b, i.e. (a, b) is principal, since R is a PID. Thus there is $c \in R$ such that

$$(a, b) = (c).$$

We claim that c is a greatest common divisor of a, b. Since $a, b \in (a, b) = (c)$, we have $c|a$, $c|b$. On the other hand $c \in (c) = (a, b)$, so $c = ra + sb$ for some $r, s \in R$. Thus if an element $q \in R$ divides both a, b, then it will also divide both ra, sb. Hence q divides $ra + sb = c$ too. ■

Remark. Recall that when R is a Euclidean domain, the generator of an ideal is a nonzero element with the least degree in that ideal. Hence in this case, we can say that a g.c.d of a, b is a nonzero element with the least degree in the set

$$\{ra + sb : r, s \in R\}.$$

Remark. The above two theorems are in particular true in $\mathbb{Z}$, and in $F[x]$ when F is a field.

B.3 Unique Factorization Domains

Definition B.20. Let R be an integral domain. A nonzero element $r \in R$ is called **irreducible** if it is not a unit, and its only divisors are units or associates of itself, i.e.

$$r = ab \implies a, b \text{ are either unit, or associates of } r.$$

A nonzero element $p \in R$ is called **prime**, if it is not a unit, and for all $a, b \in R$ we have

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

Remark. The contrapositive of the definition of a prime element p is that

$$p \not| a \text{ and } p \not| b \implies p \not| ab.$$

By an easy induction we can show that if p is prime, then for $a_1, \ldots, a_n \in R$ we have

$$p \not| a_1 \text{ and } p \not| a_2 \text{ and } \ldots \text{ and } p \not| a_n \implies p \not| a_1 \cdots a_n.$$

Equivalently, if $p|a_1 \cdots a_n$ then $p|a_i$ for some i.

***Exercise* B.21.** Let R be an integral domain, and suppose $a, b \in R$ are associates. Show that if a is irreducible then b is also irreducible, and if a is prime then b is also prime.

Theorem B.22. *Every prime element of an integral domain is irreducible.*

Proof. Suppose p is a prime, and we have a factorization $p = ab$. Then $p|ab$ so either $p|a$ or $p|b$. On the other hand, both a, b divide p. Thus either a is an associate of p, or b is. Suppose for instance that a, p are associates. Then $a = pu$ for some unit u. Hence

$$0 = p - ab = p - pub = p(1 - ub) \implies ub = 1.$$

Therefore b is a unit. ■

Theorem B.23. *Every irreducible element of a PID is prime.*

Proof. Let r be an irreducible element of the PID R. Suppose for $a, b \in R$ we have $r|ab$, and $r \not| a$. We must show that $r|b$. We claim that the greatest common divisor of r, a is a unit. The reason is that if the g.c.d of r, a is u, then $u|r$. Hence u is either a unit or an associate of r, since r is irreducible. But $u|a$ too, so it cannot be an associate of r. Because otherwise we would have $u = rv$ for some unit element v, and therefore $r|a$, which is a contradiction.

Now we know that for some $x, y \in R$ we have $u = xr + ya$. Therefore we have

$$b = u^{-1}xbr + u^{-1}yab.$$

Since r divides the right hand side of the above equation, we get $r|b$ as desired. ■

Definition B.24. An integral domain R is called a **unique factorization domain (UFD)**, if every nonzero element of R that is not a unit, can be written as a product of irreducible elements in a unique way. In other words, for all nonzero $a \in R$ which is not a unit we have

(i) There are irreducible elements $p_1, \dots, p_n \in R$ such that a has the factorization

$$a = p_1 \cdots p_n.$$

(ii) If there is another factorization of a into irreducible elements $a = q_1 \cdots q_m$, then $m = n$, and there is a permutation $\sigma \in S_n$ such that $p_i, q_{\sigma(i)}$ are associates.

***Exercise* B.25.** Show that every irreducible element of a UFD is prime.

Theorem B.26. *Every PID is a UFD.*

Proof. The uniqueness of a factorization is a consequence of the fact that in a PID irreducible elements are prime. Suppose

$$p_1 \cdots p_n = q_1 \cdots q_m,$$

where p_i, q_j's are primes. We proceed by induction on n. When $n = 1$ we have $p_1 | q_1 \cdots q_m$. Thus $p_1 | q_k$ for some k. But q_k is irreducible and p_1 is not a unit, hence $p_1 = uq_k$ for some unit u. Therefore

$$1 = q_1 \cdots q_{k-1} u q_{k+1} \cdots q_m,$$

since the cancellation law holds in integral domains. Hence if $m > 1$, the other q_j's must be units, which is a contradiction.

Now suppose the uniqueness is true for n, and we have

$$p_1 \cdots p_n p_{n+1} = q_1 \cdots q_m.$$

Then $p_{n+1} | q_1 \cdots q_m$, and therefore $p_{n+1} | q_k$ for some k. We can argue as above and conclude that for some unit u we have

$$p_1 \cdots p_n = q_1 \cdots q_{k-1} u q_{k+1} \cdots q_m = q_1 \cdots q_{k-1} (u q_{k+1}) \cdots q_m.$$

Note that uq_{k+1} is also irreducible. Now by the induction hypothesis we have $n = m - 1$, and there is a permutation $\sigma \in S_n$ such that for $i \leq n$, $p_i, q_{\sigma(i)}$ are associates when $\sigma(i) < k$, p_i, uq_{k+1} are associates when $\sigma(i) = k$, and $p_i, q_{\sigma(i)+1}$ are associates when $\sigma(i) > k$. Let

$$\hat{\sigma}(i) := \begin{cases} \sigma(i) & i \leq n,\ \sigma(i) < k \\ k & i = n+1 \\ \sigma(i)+1 & i \leq n,\ \sigma(i) > k \end{cases}$$

be a permutation in S_{n+1}. Then $p_i, q_{\tilde{\sigma}(i)}$ are associates for all $i \leq n+1$ as desired. Note that when p_i, uq_{k+1} are associates, then p_i, q_{k+1} are associates too.

Next for the existence of a factorization, let a be a nonzero element of our PID, which is not a unit. Suppose to the contrary that a does not have a factorization into irreducible elements. Then a cannot be irreducible itself, since then we have the factorization $a = a$. Thus $a = bc$ where b, c are not unit nor an associate of a. If both b, c can be factorized into irreducible elements, then a can be factorized too. Hence at least one of them, which we call it a_1, does not have a factorization. As a_1 is not an associate of a and divides a, we have

$$(a) \subsetneq (a_1).$$

Since a_1 does not have a factorization into irreducible elements, we can argue as above and find another element a_2 such that

$$(a) \subsetneq (a_1) \subsetneq (a_2).$$

We can continue this process inductively and get

$$(a) \subsetneq (a_1) \subsetneq \cdots \subsetneq (a_n) \subsetneq \cdots .$$

Now, let $I := \bigcup_{n\geq 1}(a_n)$. It is easy to see that I is an ideal. Because if $b, c \in I$ then $b \in (a_i)$ and $c \in (a_j)$ for some i, j. Let $n = \max\{i, j\}$. Then $b, c \in (a_n)$. Hence $b + c \in (a_n) \subset I$, and also $sb \in (a_n) \subset I$ for all elements s. Thus we have $I = (r)$ for some element r. Then $r \in I$, so $r \in (a_m)$ for some m. But this means that $(r) \subset (a_m)$. In particular we have $(a_{m+1}) \subset (a_m)$, which is a contradiction. Therefore a must have a factorization into irreducible elements. ■

Second Proof. Here we give another proof for the existence of the factorization when our PID is a Euclidean domain. Let a be a nonzero element which is not a unit. The proof is by strong induction on $d(a)$, the degree of a. If a has the least d-value among the nonzero elements that are not units, then a must be irreducible. To see this suppose that $a = bc$, and c is not a unit. Note that $b, c \neq 0$ since $a \neq 0$. Then by Proposition B.8 we have $d(b) < d(bc) = d(a)$. But a has the least d-value among nonzero elements which are not units, so b must be unit. Therefore a is irreducible, and thence has a factorization into irreducible elements.

Now suppose every element with degree less than $d(a)$ has a factorization into irreducible elements. If a is irreducible, then it has a factorization. Otherwise we have $a = bc$, where b, c are nonzero and they are not units. Hence again by Proposition B.8 we have $d(b) < d(bc) = d(a)$. Similarly $d(c) < d(bc) = d(a)$. Therefore by the induction hypothesis b, c can be written as a product of irreducible elements. Now if we multiply those expressions we obtain a factorization of a into irreducible elements, as desired. ■

Fundamental Theorem of Arithmetic. *$\mathbb{Z}$ is a UFD, i.e. every nonzero integer other than ± 1 can be written uniquely as a product of prime integers.*

Proof. $\mathbb{Z}$ is a Euclidean domain, so it is a PID, hence it is a UFD. Note that ± 1 are the only units of $\mathbb{Z}$, since they are the only invertible elements of the ring $\mathbb{Z}$. ■

Theorem B.27. *Suppose F is a field. Then $F[x]$ is a UFD, i.e. every nonconstant polynomial with coefficients in a field can be written uniquely as a product of irreducible polynomials.*

Proof. $F[x]$ is a Euclidean domain, so it is a PID, hence it is a UFD. Note that the nonzero constant polynomials are precisely the units of $F[x]$, as we showed in Example B.7. ■