

# Introduction to Linear Algebra

*with Earth Science Applications*

---

Draft Version

Benjamin C. Loi



Introduction to Linear Algebra with Earth Science Applications

Copyright ©, Benjamin C. Loi, 2024

All rights reserved. Any reproduction of this work, in part or in whole, is prohibited without the written permission of the author.

# Preface

---

This is a Linear Algebra textbook specifically designed for students that are majoring in any Earth Science related subjects like Geophysics and Atmospheric Sciences, who are also interested in Mathematics. With these target readers in mind, we set out to provide an adequate treatment about Linear Algebra concepts that enables them to tackle relevant Earth Science problems. In each chapter, we focus on a selected Linear Algebra topic, motivated by Earth Science examples and supplemented with Python programming tutorials. At the end of each chapter, a number of exercises are given for elucidating the concepts and working on Earth Science projects.

Benjamin Loi



# Contents

---

<b>1</b>	<b>Introduction to Matrices and Linear Systems</b>	<b>13</b>
1.1	Definition and Operations of Matrices . . . . .	13
1.1.1	Basic Structure of Matrices . . . . .	13
1.1.2	Matrix Operations . . . . .	14
1.2	Definition of Linear Systems of Equations . . . . .	21
1.3	Elementary Row Operations . . . . .	24
1.4	Earth Science Applications . . . . .	26
1.5	Python Programming . . . . .	31
1.6	Exercises . . . . .	34
<b>2</b>	<b>Inverses and Determinants</b>	<b>39</b>
2.1	Identity Matrices and Transpose . . . . .	39
2.1.1	Identity Matrices . . . . .	39
2.1.2	Transpose . . . . .	41
2.1.3	Symmetric Matrices . . . . .	42
2.2	Inverses . . . . .	43
2.2.1	Definition and Properties of Inverses . . . . .	43
2.2.2	(Reduced) Row Echelon Form . . . . .	46
2.2.3	Finding Inverses by Gaussian Elimination . . . . .	52
2.3	Determinants . . . . .	57
2.3.1	Computing Determinants . . . . .	57
2.3.2	Properties of Determinants . . . . .	62
2.3.3	Finding Inverses by Adjugate . . . . .	66
2.4	Python Programming . . . . .	69
2.5	Exercises . . . . .	72

## *Contents*

---

<b>3</b>	<b>Solutions for Linear Systems</b>	<b>75</b>
3.1	Number of Solutions for Linear Systems . . . . .	75
3.2	Solving Linear Systems . . . . .	78
3.2.1	Solving Linear Systems by Gaussian Elimination . . . . .	79
3.2.2	Solving Linear Systems by Inverse . . . . .	88
3.3	Earth Science Applications . . . . .	89
3.4	Python Programming . . . . .	94
3.5	Exercises . . . . .	96
<b>4</b>	<b>Introduction to Vectors</b>	<b>101</b>
4.1	Definition and Operations of Geometric Vectors . . . . .	101
4.1.1	Basic Structure of Vectors in the Real $n$ -space $\mathbb{R}^n$ . . . . .	101
4.1.2	Fundamental Vector Operations . . . . .	104
4.2	Special Vector Operations . . . . .	108
4.2.1	Dot Product . . . . .	108
4.2.2	Cross Product . . . . .	116
4.3	Earth Science Applications . . . . .	121
4.4	Python Programming . . . . .	123
4.5	Exercises . . . . .	125
<b>5</b>	<b>Vector Geometry</b>	<b>129</b>
5.1	Lines and Planes . . . . .	129
5.1.1	Translating Equation Form to Vector Form . . . . .	130
5.1.2	Recovering Equation Form from Vector Form . . . . .	132
5.1.3	Generalizing to Higher Dimensions . . . . .	133
5.2	Further Geometric Applications of Dot Product . . . . .	135
5.2.1	Projection . . . . .	135
5.2.2	Distance . . . . .	137
5.3	Further Geometric Applications of Cross Product . . . . .	138
5.3.1	Area . . . . .	138
5.3.2	Volume . . . . .	139
5.4	Useful Vector Identities . . . . .	143
5.5	Earth Science Applications . . . . .	146
5.6	Python Programming . . . . .	149
5.7	Exercises . . . . .	150

<b>6 Vector Spaces and Coordinate Bases</b>	<b>155</b>
6.1 Making of the Real $n$ -space $\mathbb{R}^n$ . . . . .	155
6.1.1 $\mathbb{R}^n$ as a Vector Space . . . . .	155
6.1.2 Subspaces of $\mathbb{R}^n$ . . . . .	157
6.1.3 Span by Linear Combinations of Vectors . . . . .	161
6.1.4 Linear Independence, CR Factorization . . . . .	167
6.2 Coordinate Bases for $\mathbb{R}^n$ and its Subspaces . . . . .	182
6.2.1 Coordinate Bases for $\mathbb{R}^n$ . . . . .	182
6.2.2 Coordinate Bases for Subspaces of $\mathbb{R}^n$ . . . . .	185
6.2.3 Direct Sum Representation . . . . .	189
6.3 The Four Fundamental Subspaces Induced by Matrices . . . . .	197
6.3.1 Row Space, Column Space . . . . .	197
6.3.2 Null Space, Rank-Nullity Theorem . . . . .	202
6.4 Python Programming . . . . .	209
6.5 Exercises . . . . .	210
<b>7 More on Coordinate Bases, Linear Transformations</b>	<b>213</b>
7.1 Ideas of Linear Transformations . . . . .	213
7.1.1 Linear Maps between Vector Spaces . . . . .	213
7.1.2 One-to-one and Onto, Kernel and Range . . . . .	221
7.1.3 Composition of Linear Transformations . . . . .	228
7.1.4 Vector Space Isomorphism to $\mathbb{R}^n$ . . . . .	230
7.2 Additional Discussions about Coordinate Bases . . . . .	237
7.2.1 Linear Change of Coordinates . . . . .	237
7.2.2 Gram-Schmidt Orthogonalization, QR Decomposition	243
7.3 Python Programming . . . . .	250
7.4 Exercises . . . . .	253
<b>8 Complex Vectors/Matrices and Block Matrices</b>	<b>255</b>
8.1 Definition and Operations of Complex Numbers . . . . .	255
8.1.1 Basic Structure of Complex Numbers . . . . .	255
8.1.2 Complex Number Operations . . . . .	256
8.1.3 Geometric Meaning of Complex Numbers . . . . .	258
8.2 Complex Vectors and Complex Matrices . . . . .	262
8.2.1 Operations and Properties of Complex Vectors . . . . .	262

## *Contents*

---

8.2.2	Operations and Properties of Complex Matrices . . . . .	264
8.2.3	The Complex $n$ -space $\mathbb{C}^n$ . . . . .	269
8.3	Manipulating Block Matrices . . . . .	271
8.3.1	Block Matrix Multiplication . . . . .	272
8.3.2	Inverse and Determinant of a Block Matrix . . . . .	274
8.3.3	Restriction of a Linear Transformation, Direct Sum of a Matrix . . . . .	280
8.4	Python Programming . . . . .	288
8.5	Exercises . . . . .	291
<b>9</b>	<b>Eigenvalues and Eigenvectors</b>	<b>295</b>
9.1	Eigenvalues and Eigenvectors of a Square Matrix . . . . .	295
9.1.1	Definition of Eigenvalues and Eigenvectors . . . . .	295
9.1.2	Finding Eigenvalues and Eigenvectors with Characteristic Polynomials . . . . .	298
9.1.3	Eigenspace as an Invariant Subspace . . . . .	306
9.1.4	Cayley-Hamilton Theorem . . . . .	307
9.2	Diagonalization . . . . .	309
9.2.1	Mathematical Ideas of Diagonalization . . . . .	309
9.2.2	Diagonalization for Real Eigenvalues . . . . .	311
9.2.3	Properties of Diagonalization . . . . .	314
9.2.4	Diagonalization for Complex Eigenvalues . . . . .	317
9.3	System of Ordinary Differential Equations . . . . .	323
9.4	Earth Science Applications . . . . .	331
9.5	Python Programming . . . . .	336
9.6	Exercises . . . . .	338
<b>10</b>	<b>Orthogonal and Normal Matrices</b>	<b>341</b>
10.1	Orthogonal Matrices . . . . .	341
10.1.1	Definition of Orthogonal Matrices . . . . .	341
10.1.2	Geometric Implications of Orthogonal Matrices . . . . .	344
10.2	Orthogonal Diagonalization . . . . .	351
10.3	Orthogonal Projections and Spectral Theorem . . . . .	357
10.3.1	Projections onto a Subspace . . . . .	357
10.3.2	Orthogonal Projections . . . . .	359

10.3.3	Spectral Theorem . . . . .	365
10.4	Normal Matrices and Unitary Diagonalization . . . . .	368
10.4.1	Unitary and Normal Matrices . . . . .	368
10.4.2	Unitary Diagonalization . . . . .	369
10.5	Python Programming . . . . .	375
10.6	Exercises . . . . .	377
<b>11</b>	<b>Quadratic Forms</b>	<b>381</b>
11.1	Mathematical and Geometric Ideas of Quadratic Forms . . . . .	381
11.1.1	Definition of Quadratic Forms . . . . .	381
11.1.2	(Semi)Definiteness and Congruence . . . . .	384
11.1.3	Conic Sections . . . . .	392
11.2	Statistics with Quadratic Form . . . . .	400
11.2.1	Variance and Covariance . . . . .	400
11.2.2	Principal Component Analysis (PCA) . . . . .	409
11.3	Python Programming . . . . .	417
11.4	Earth Science Applications: Empirical Orthogonal Functions (EOFs) . . . . .	420
11.5	Exercises . . . . .	423
<b>12</b>	<b>Inner Product Spaces</b>	<b>427</b>
12.1	Definition and Properties of Inner Product Spaces . . . . .	428
12.1.1	Requirements of Inner Products . . . . .	428
12.1.2	Generalization of Length and Orthogonality via Inner Products . . . . .	430
12.1.3	Infinite-dimensional Inner Product Spaces . . . . .	432
12.2	Adjoints and Hermitian/Unitary Operators . . . . .	435
12.2.1	Definition of Adjoints . . . . .	435
12.2.2	Hermitian Operators . . . . .	438
12.2.3	Unitary Operators . . . . .	443
12.3	Revisiting Orthogonal Projections . . . . .	445
12.3.1	Orthogonal Projections for an Inner Product Space . . . . .	445
12.3.2	Revisiting Gram-Schmidt Orthogonalization . . . . .	448
12.3.3	Spectral Theorem for Hermitian Operators . . . . .	451

## *Contents*

---

12.4 Special Polynomials . . . . .	456
12.4.1 Sturm-Liouville Equations . . . . .	456
12.4.2 Generating Special Polynomials by Gram-Schmidt Orthonormalization . . . . .	461
12.5 Earth Science Applications . . . . .	464
12.6 Python Programming . . . . .	469
12.7 Exercises . . . . .	470
<b>13 Least-Square Approximation</b> . . . . .	<b>473</b>
13.1 Mathematical Ideas of Least-Square Approximation . . . . .	473
13.2 Linear Regression . . . . .	479
13.2.1 Linear Regression for One Predictor Variable . . . . .	479
13.2.2 Linear Regression for Multiple Predictor Variables . . . . .	482
13.2.3 Properties of Linear Regression . . . . .	485
13.3 Earth Science Applications . . . . .	488
13.4 Python Programming . . . . .	489
13.5 Exercise . . . . .	490
<b>14 Discrete Fourier Transform (DFT)</b> . . . . .	<b>495</b>
14.1 Mathematical Ideas of DFT . . . . .	495
14.1.1 From Fourier Series to a Prototype of DFT . . . . .	495
14.1.2 Nyquist Frequency and Real DFT . . . . .	507
14.1.3 Complex DFT . . . . .	510
14.1.4 Inverse DFT . . . . .	515
14.2 Properties of DFT . . . . .	518
14.2.1 Power Spectrum and Parseval's Theorem . . . . .	518
14.2.2 Convolution Theorem . . . . .	522
14.3 Fast Fourier Transform (FFT) . . . . .	530
14.4 Python Programming and Earth System Applications . . . . .	535
14.5 Exercise . . . . .	538
<b>Answers to Exercises</b>	<b>541</b>
<b>Index</b>	<b>561</b>

<b>Bibliography</b>	<b>565</b>
---------------------	------------



## ***Chapter 1***

# **Introduction to Matrices and Linear Systems**

---

Although the Earth System is well-known to be filled with non-linear processes, we still benefit from learning how to work with linear systems, by which many Earth Science problems can be approximated. This actually works well in a number of cases. For instance, in Atmosphere Sciences, we often consider what is called a *perturbation equation*, which assumes that deviations from the mean state are small enough to neglect quadratic terms. The most fundamental usage of Linear Algebra in Applied Sciences is to formulate, analyze and solve *linear systems of equations*. Some examples in Earth Sciences are mapping the depth of overlying soil layers underground, as well as chemical balances in various subsystems of the Earth. *Matrices* are one of the most central ingredients in Linear Algebra that can be used to described such systems, and we are going to address the basic aspects related to them in the first chapter.

## **1.1 Definition and Operations of Matrices**

### **1.1.1 Basic Structure of Matrices**

*Matrices* are rectangular arrays of numbers, the entries of which can be real or complex. For now we will work with the simpler case of real matrices first. A

matrix having  $m$  rows and  $n$  columns is called an  $m \times n$  matrix. The class of matrices with the same number of rows and columns, i.e.  $m = n$ , are known as **square matrices**. Below shows some examples of matrices.

$$\begin{bmatrix} 1.17 & 2.01 & -2.15 & 5 \\ 1.44 & 3.61 & 2.88 & -3 \end{bmatrix}$$

A  $2 \times 4$  real matrix.

$$\begin{bmatrix} \sqrt{2} - \frac{4}{\sqrt{5}}i \\ 0 \\ 1.27 \\ \sqrt{3}i \end{bmatrix}$$

A  $4 \times 1$  complex matrix.

$$\begin{bmatrix} 3 & \sqrt{2} & 9 \\ 0 & -4\pi & \frac{1}{6} \\ 5.11 & 2 & -1 \end{bmatrix}$$

A  $3 \times 3$  real, square matrix.

Given any matrix  $A$ , its entry at row  $i$  and column  $j$  will be denoted as  $A_{ij}$ . For example,

$$A = \left[ \begin{array}{c|ccc} \text{Col 1} & 2 & 1 & 7 & \frac{8}{9} \\ \hline \text{Row 2} & 5 & -\frac{1}{3} & 5 & 0 \\ \hline -3 & 4.38 & 6 & -1.66 \end{array} \right] \quad A_{21} = 5$$

Short Exercise: Find  $A_{13}$ ,  $A_{22}$ ,  $A_{34}$  and  $A_{42}$ .<sup>1</sup>

## 1.1.2 Matrix Operations

### Addition and Subtraction

Addition and subtraction between two matrices  $A$  and  $B$  are carried out *entry-wise*, which means that if  $C = A \pm B$ , then  $C_{ij} = A_{ij} \pm B_{ij}$ . This implies that

---

<sup>1</sup>  $A_{13} = 7$ ,  $A_{22} = -\frac{1}{3}$ ,  $A_{34} = -1.66$ ,  $A_{42}$  does not exist.

the two matrix operands must be of the same shape, and addition/subtraction is not defined for two matrices with different shapes. For instance, if we have

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 8.5 \\ 1 & -7 \end{bmatrix}$$

Then

$$\begin{aligned} A + B &= \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 8.5 \\ 1 & -7 \end{bmatrix} \\ &= \begin{bmatrix} 1+1 & 2+1 \\ 3+0 & 4+8.5 \\ 5+1 & 6+(-7) \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 \\ 3 & 12.5 \\ 6 & -1 \end{bmatrix} \end{aligned}$$

Short Exercise: Find  $A - B$ .<sup>2</sup>

### Scalar Multiplication

Multiplying a matrix by a number (*scalar*) constitutes a ***scalar multiplication***, in which all entries are multiplied by that scalar. It is illustrated by the example below.

$$\begin{aligned} A &= \begin{bmatrix} 2 & -5.3 & 6 \\ -1 & 4.1 & -3 \end{bmatrix} \\ 3A &= 3 \begin{bmatrix} 2 & -5.3 & 6 \\ -1 & 4.1 & -3 \end{bmatrix} \end{aligned}$$

---


$${}^2A - B = \begin{bmatrix} 0 & 1 \\ 3 & -4.5 \\ 4 & 13 \end{bmatrix}$$

$$\begin{aligned}
 &= \begin{bmatrix} 3(2) & 3(-5.3) & 3(6) \\ 3(-1) & 3(4.1) & 3(-3) \end{bmatrix} \\
 &= \begin{bmatrix} 6 & -15.9 & 18 \\ -3 & 12.3 & -9 \end{bmatrix}
 \end{aligned}$$

Short Exercise: Find  $\frac{1}{4}A$ .<sup>3</sup>

### Matrix Multiplication/Matrix Product

Meanwhile, multiplication between two matrices, commonly referred to as ***matrix multiplication/matrix product***, is not entry-wise. It can be only carried out if the number of columns of the first matrix  $A$  equals to the number of rows of the second matrix  $B$ , let's say  $r$ . In other words, they need to be of the shapes  $m \times r$  and  $r \times n$  respectively. The resulting matrix  $AB$  will then have the shape  $m \times n$ , which means that the number of rows/columns of the output matrix follows the first/second input matrix respectively. The following two examples explain this requirement.

$$A = \begin{bmatrix} 1 & 2.1 & 2 \\ 1 & 3 & 5 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 5 \\ \sqrt{7} \end{bmatrix}$$

Since the shapes of  $A$  and  $B$  are  $2 \times 3$  and  $3 \times 1$  so that the number of columns in  $A$  and the number of rows in  $B$  are both 3, the matrix product  $AB$  is possible. The resulting matrix will be of size  $2 \times 1$ . On the other hand,  $BA$  is not defined if we reverse the order of the matrix product. Meanwhile, for

$$C = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 0 & 6 & 1 \end{bmatrix} \quad D = \begin{bmatrix} 3.44 & 1.07 \\ 0 & 5.96 \\ -4.3 & 2.75 \end{bmatrix}$$

as the number of columns in  $C$  is 4, which is not equal to the number of rows in  $D$  (3), the matrix product  $CD$  is undefined in this case. (However,  $DC$  is just

---


$$\frac{1}{4}A = \begin{bmatrix} \frac{1}{2} & -1.325 & \frac{3}{2} \\ -\frac{1}{4} & 1.025 & -\frac{3}{4} \end{bmatrix}$$

valid, and what will be its shape?<sup>4</sup>) Now we are ready to see how the entries in matrix product is exactly computed.

**Definition 1.1.1** (Matrix Product/Matrix Multiplication). Given an  $m \times r$  matrix  $A$  and another  $r \times n$  matrix  $B$ , we denote the matrix product between  $A$  and  $B$  as  $AB$  that will have the shape of  $m \times n$ . To calculate any entry in  $AB$  at row  $i$  and column  $j$ , we select row  $i$  from the first matrix  $A$  and column  $j$  from the second matrix  $B$ . Subsequently, take the products within each of the  $r$  pairs of numbers from that row and column. Their sum will then be the required value of the element, i.e.

$$\begin{aligned}(AB)_{ij} &= A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i3}B_{3j} + \dots + A_{ir}B_{rj} \\ &= \sum_{k=1}^r A_{ik}B_{kj}\end{aligned}$$

again,  $r$  is the number of columns/rows in the first/second matrix.

**Example 1.1.1.** Calculate the matrix product  $C = AB$ , where

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

*Solution.* The output will be a  $2 \times 2$  matrix. Using the definition above, we have

$$\begin{aligned}C_{11} &= (AB)_{11} = A_{11}B_{11} + A_{12}B_{21} + A_{13}B_{31} \\ &= (1)(1) + (3)(2) + (5)(3) = 22 \\ C_{12} &= (AB)_{12} = A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32} \\ &= (1)(4) + (3)(5) + (5)(6) = 49\end{aligned}$$

Hence the entries along the first row of  $C$  will be 22 and 49. The remaining entries at the second row can be found in a similar way, and the readers are

---

<sup>4</sup> $DC$  will be a  $3 \times 4$  matrix.

encouraged to do this themselves. You should be able to get

$$C = \begin{bmatrix} 22 & 49 \\ 28 & 64 \end{bmatrix}$$

□

Matrix product has some important properties, listed as follows.

**Properties 1.1.2.** If  $A$ ,  $B$ ,  $C$  are some matrices having compatible shapes (*conformable*) so that the matrix multiplication operations below are valid, then

$\underbrace{A \cdots A}_{k \text{ times}} = A^k$	$k$ -th power of a (square) matrix
$(AB)C = A(BC) = ABC$	Associative Property
$(A \pm B)C = AC \pm BC$	Distributive Property
$A(B \pm C) = AB \pm AC$	Distributive Property

Another important observation is that, in general  $AB \neq BA$  even if the matrix products  $AB$  and  $BA$  are both well-defined, so they are not *commutative*. However, there are some exceptions to this.<sup>5</sup>

**Example 1.1.2.** Calculate  $-2A + 3B$ , where

$$A = \begin{bmatrix} 1 & 6 & 9 \\ 4 & 4 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 8 & 6 \\ -5 & 0 & 3 \end{bmatrix}$$

*Solution.*

$$\begin{aligned} -2A + 3B &= -2 \begin{bmatrix} 1 & 6 & 9 \\ 4 & 4 & 6 \end{bmatrix} + 3 \begin{bmatrix} 4 & 8 & 6 \\ -5 & 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} -2 & -12 & -18 \\ -8 & -8 & -12 \end{bmatrix} + \begin{bmatrix} 12 & 24 & 18 \\ -15 & 0 & 9 \end{bmatrix} \end{aligned}$$

---

<sup>5</sup>A trivial exception is that  $A = B$ .

$$= \begin{bmatrix} 10 & 12 & 0 \\ -23 & -8 & -3 \end{bmatrix}$$

□

**Example 1.1.3.** Compute  $(A + 3B)(2A - B)$ , where

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \quad B = \begin{bmatrix} -2 & 0 \\ 4 & -1 \end{bmatrix}$$

*Solution.* Using the distributive property in Properties 1.1.2, the expression can be expanded to

$$\begin{aligned} (A + 3B)(2A - B) &= A(2A - B) + (3B)(2A - B) \\ &= A(2A) + A(-B) + (3B)(2A) + (3B)(-B) \\ &= 2A^2 - AB + 6BA - 3B^2 \end{aligned}$$

Bear in mind that  $AB \neq BA$ . We calculate each of the terms, which gives

$$\begin{aligned} A^2 &= \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \\ &= \begin{bmatrix} (1)(1) + (2)(3) & (1)(2) + (2)(5) \\ (3)(1) + (5)(3) & (3)(2) + (5)(5) \end{bmatrix} \\ &= \begin{bmatrix} 7 & 12 \\ 18 & 31 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} AB &= \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 4 & -1 \end{bmatrix} \\ &= \begin{bmatrix} (1)(-2) + (2)(4) & (1)(0) + (2)(-1) \\ (3)(-2) + (5)(4) & (3)(0) + (5)(-1) \end{bmatrix} \\ &= \begin{bmatrix} 6 & -2 \\ 14 & -5 \end{bmatrix} \end{aligned}$$

Similarly, it is not difficult to obtain

$$BA = \begin{bmatrix} -2 & -4 \\ 1 & 3 \end{bmatrix} \quad B^2 = \begin{bmatrix} 4 & 0 \\ -12 & 1 \end{bmatrix}$$

Hence the final answer will be

$$\begin{aligned} & 2A^2 - AB + 6BA - 3B^2 \\ &= 2 \begin{bmatrix} 7 & 12 \\ 18 & 31 \end{bmatrix} - \begin{bmatrix} 6 & -2 \\ 14 & -5 \end{bmatrix} + 6 \begin{bmatrix} -2 & -4 \\ 1 & 3 \end{bmatrix} - 3 \begin{bmatrix} 4 & 0 \\ -12 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 14 & 24 \\ 36 & 62 \end{bmatrix} - \begin{bmatrix} 6 & -2 \\ 14 & -5 \end{bmatrix} + \begin{bmatrix} -12 & -24 \\ 6 & 18 \end{bmatrix} - \begin{bmatrix} 12 & 0 \\ -36 & 3 \end{bmatrix} \\ &= \begin{bmatrix} -16 & 2 \\ 64 & 82 \end{bmatrix} \end{aligned}$$

□

Alternatively, one can evaluate  $C = A+3B$  and  $D = 2A-B$  first, and subsequently calculate the matrix dot product  $CD$ . (This is actually easier and more efficient.) The readers should try this as an exercise.

## Matrix Equation Manipulation

For any matrix equation, one can do addition, subtraction and multiplication on both sides of the equation. However, one important note is that multiplying a matrix to an equation requires that the same matrix to be inserted to the left (or right) on both sides, respecting the order. So, for a matrix equation like (assuming the shapes of matrices are compatible),

$$AB - C = DE + F \tag{1.1}$$

if we want to multiply the equation by some matrix  $G$ , then two possibilities are

$$G(AB - C) = G(DE + F)$$

$$(AB - C)G = (DE + F)G$$

but we have, in general

$$\begin{aligned} G(AB - C) &\neq (DE + F)G \\ (AB - C)G &\neq G(DE + F) \end{aligned}$$

Doing successive matrix multiplications follows the same principle, step by step. Using the same example of Equation (1.1), given another matrix  $H$ , we note some valid outcomes.

$$\begin{aligned} HG(AB - C) &= HG(DE + F) \\ (AB - C)GH &= (DE + F)GH \\ GH(AB - C) &= GH(DE + F) \\ H(AB - C)G &= H(DE + F)G \\ G(AB - C)H &= G(DE + F)H \end{aligned}$$

However, be careful that cancellation at both sides may not be correct. If  $AB = AC$ , then we cannot conclude that  $B = C$  for certain. Nevertheless, in the next chapter we will see one of the scenarios where cancellation actually works.

## 1.2 Definition of Linear Systems of Equations

The prime application of matrices is to deal with *linear systems (of equations)* as mentioned in the introduction. To understand what a linear system is, we first have to know the definition of a *linear equation* (in multiple variables, let's say  $x_1, x_2, \dots$ , or  $x, y, \dots$ ). In a linear equation, for any additive term, there is at most one variable (unknown), with a power of one, times some constant coefficient, like  $x, -\sqrt{5}x, -y, 2.33y$ . This means that there are no cross-product terms such as  $1.68xy$ , variables with a power that is not one, like  $x^3$ , or non-linear functions, including  $\sin x, e^y$ . For  $n$  variables, a linear equation has the following form.

**Definition 1.2.1** (Linear Equation). A linear equation is an equation in the form of

$$\sum_{j=1}^n a_j x_j = a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_n x_n = h$$

where  $x_1, x_2, \dots, x_n$  are the unknowns, while  $a_1, a_2, \dots, a_n$  and  $h$  are some constants. If  $h = 0$ , then it is known as a **homogeneous linear equation**.

Short Exercise: Determine whether the equations below are (a) linear, and if they are linear, then (b) homogeneous or not. The unknowns are  $x, y, z$ .<sup>6</sup>

1.  $3x + 4.7y = 2\sqrt{2}$
2.  $\cos x + \ln y = 0$
3.  $7\pi x - z = 2$
4.  $x^2 + 3.8y^{-3/2} = 1$
5.  $1.05x + 3.17y + 6.44z = 0$
6.  $xyz = 8$

A system of linear equations are then simply a family of  $m$  linear equations in a set of some unknowns,  $m \geq 1$ .

**Definition 1.2.2** (Linear System of Equations). A linear system of size  $m \times n$ , i.e.  $m$  linear equations in  $n$  unknowns  $(x_1, x_2, \dots, x_n)$ , has the form of

$$\begin{cases} \sum_{j=1}^n a_j^{(1)} x_j = a_1^{(1)} x_1 + a_2^{(1)} x_2 + a_3^{(1)} x_3 + \cdots + a_n^{(1)} x_n & = h^{(1)} \\ \sum_{j=1}^n a_j^{(2)} x_j = a_1^{(2)} x_1 + a_2^{(2)} x_2 + a_3^{(2)} x_3 + \cdots + a_n^{(2)} x_n & = h^{(2)} \\ \vdots \\ \sum_{j=1}^n a_j^{(m)} x_j = a_1^{(m)} x_1 + a_2^{(m)} x_2 + a_3^{(m)} x_3 + \cdots + a_n^{(m)} x_n & = h^{(m)} \end{cases}$$

If  $h^{(1)}, h^{(2)}, \dots, h^{(m)}$  on R.H.S. are all zeros, i.e. all the equations are homoge-

---

<sup>6</sup>Linear/Inhomogeneous, Non-linear, Linear/Inhomogeneous, Non-linear, Linear/Homogeneous, Non-linear.

neous, then the system is called a ***homogeneous linear system (of equations)***.

It is not hard to see that for any homogeneous linear system, it always has the trivial solution of  $x_j = 0$  for  $j = 1, 2 \dots, n$ , or expressed as  $\vec{x} = \mathbf{0}$ . However, such trivial solution may not be the only solution to the system, as we shall see in Chapter 3. Below shows some examples of linear systems.

$$\begin{cases} 3.3x + 4y = 5 \\ 7x + 9.7y = 13.1 \end{cases}$$

A  $2 \times 2$  linear system with two equations, two unknowns.

$$\begin{cases} x + 2y - 4z = 3 \\ x - y + 3z = -4 \end{cases} \quad (1.2)$$

A  $2 \times 3$  linear system with two equations, three unknowns.

$$\begin{cases} x + 2.2y + 3z = 0 \\ 2x + 3z = 0 \\ 4x - 5.6y = 0 \end{cases} \quad (1.3)$$

A  $3 \times 3$  homogeneous linear system (homogeneous as the constants on R.H.S. are all zeros), notice that the coefficients of  $y$  and  $z$  in the second/third equations are zeros as well and do not appear explicitly.

The above formulation of a linear system closely resembles a tabular structure. Therefore, we are motivated to represent such systems with the language of matrices, which have an appearance of tabular arrays. Indeed, it is possible to rewrite an  $m \times n$  linear system as  $A\vec{x} = \vec{h}$ , where  $A$  is an  $m \times n$  matrix with entries copied from the coefficients in front of the variables, arranged like in Definition 1.2.2. In this book sometimes we will call it a *coefficient matrix*. Meanwhile,  $\vec{x}$  is a *column vector* (an  $n \times 1$  matrix) holding the  $n$  unknowns, and  $\vec{h}$  is another column vector (an  $m \times 1$  matrix) that contains the  $m$  constants on R.H.S. of the linear system.

**Properties 1.2.3.** For a linear system defined as in Definition 1.2.2, it can be rewritten as  $A\vec{x} = \vec{h}$ , where  $A_{ij} = a_j^{(i)}$ ,  $\vec{x} = x_j$ , and  $\vec{h} = h^{(i)}$ .

Using the second example (Equation (1.2)) above as an illustration, we can easily verify that

$$\begin{cases} x + 2y - 4z = 3 \\ x - y + 3z = -4 \end{cases}$$

can be expressed as (you should check it by expanding the matrix product)

$$\begin{bmatrix} 1 & 2 & -4 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

An even simpler representation is the *augmented matrix* which omits the unknowns and concatenates  $A$  and  $\vec{h}$ .

$$\left[ \begin{array}{ccc|c} 1 & 2 & -4 & 3 \\ 1 & -1 & 3 & -4 \end{array} \right]$$

Short Exercise: Write down the augmented matrix for the linear system in Equation (1.3).<sup>7</sup>

## 1.3 Elementary Row Operations

When we construct a matrix, it is natural to think about how to manipulate its structure. *Elementary row operations* provide such possibility in three ways, outlined in the following definition.

---


$$^7 \left[ \begin{array}{ccc|c} 1 & 2.2 & 3 & 0 \\ 2 & 0 & 3 & 0 \\ 4 & -5.6 & 0 & 0 \end{array} \right]$$

**Definition 1.3.1** (Elementary Row Operations). Denote the  $p$ -th row of a matrix as  $R_p$ . The three types of elementary row operations are

1. Multiplying a row  $R_p$  by any non-zero constant  $c \neq 0$ ;
2. Adding another row  $R_q$  times any non-zero constant  $c \neq 0$ , to a row  $R_p$ , such that the new  $p$ -th row becomes  $R_p + cR_q$ ;
3. Swapping a row  $R_p$  with another row  $R_q$ .

To facilitate computation, we denote these three kinds of operations using the following notations.

1.  $cR_p \rightarrow R_p$ ,
2.  $R_p + cR_q \rightarrow R_p$ ,
3.  $R_p \leftrightarrow R_q$

For example, the matrix  $A$

$$\begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 11 \end{bmatrix}$$

can be transformed to a new matrix  $A'$

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 3 & 5 \end{bmatrix}$$

if we apply the elementary row operation of subtracting  $2R_1$  from  $R_2$  (i.e.  $R_2 - 2R_1 \rightarrow R_2$ ).

Short Exercise: Find out the resulting matrix  $A''$  if we multiply the first row of  $A'$  by 3 and then subtract the second row from the first row.<sup>8</sup>

Attentive readers may have noticed that these three types of elementary row operations resemble what we have been always doing to the equations when solving a

---

<sup>8</sup> $\begin{bmatrix} 0 & 3 & 4 \\ 3 & 3 & 5 \end{bmatrix}$

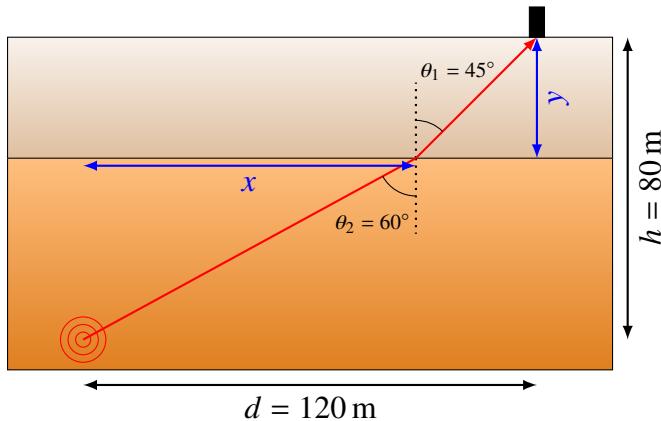


Figure 1.1: The underground schematic for the seismic ray in Example 1.4.1.

linear system as taught in high school. We re-introduce them as elementary row operations here first as they allow a systematic treatment of linear systems and matrices in later chapters.

## 1.4 Earth Science Applications

**Example 1.4.1.** Seismic wave follows *Snell's Law* like a light ray when it comes to refraction. Assuming the ground can be modelled as a two-layer system (see Figure 1.1), and we know a particular train of seismic wave generated from an underground source that reaches the ground receiver travels at an angle of  $\theta_1 = 45^\circ/\theta_2 = 60^\circ$  to the vertical at the top/bottom layer. ( $\theta_1$  can be found by analyzing the seismic waveform, and then  $\theta_2$  can be estimated by Snell's Law given we know about the densities of the respective layers.) Given that the horizontal and vertical distance between the seismic source and the surface receiver are  $d = 120 \text{ m}$  and  $h = 80 \text{ m}$ , construct a linear system for this situation in two unknowns: the depth of the top layer  $y$  and the horizontal displacement  $x$  (in meters) where the wave reaches at the interface relative to the source.

*Solution.* We can deduce two equations from the given information. Consider the upper portion of the seismic ray, from basic trigonometry, we know that

$$\begin{aligned}\frac{d-x}{y} &= \tan \theta_1 \\ d-x &= (\tan \theta_1)y \\ x + (\tan \theta_1)y &= d\end{aligned}$$

Similarly, for the lower portion of the seismic ray, we have

$$\begin{aligned}\frac{x}{h-y} &= \tan \theta_2 \\ x &= (\tan \theta_2)h - (\tan \theta_2)y \\ x + (\tan \theta_2)y &= (\tan \theta_2)h\end{aligned}$$

The corresponding linear system is

$$\begin{cases} x + (\tan \theta_1)y = d \\ x + (\tan \theta_2)y = (\tan \theta_2)h \end{cases}$$

where  $x$  and  $y$  are the unknowns to be solved.  $d$ ,  $h$ ,  $\theta_1$  and  $\theta_2$  (and hence  $\tan \theta_1$  and  $\tan \theta_2$ ) are constants. Expressing the system in matrix form, we have

$$\begin{bmatrix} 1 & \tan \theta_1 \\ 1 & \tan \theta_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} d \\ (\tan \theta_2)h \end{bmatrix}$$

Substituting the provided values for the constants ( $\tan \theta_1 = \tan(45^\circ) = 1$ ,  $\tan \theta_2 = \tan(60^\circ) = \sqrt{3}$ ), we have

$$\begin{bmatrix} 1 & 1 \\ 1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 120 \\ 80\sqrt{3} \end{bmatrix}$$

□

**Example 1.4.2.** The radiation transfer across the atmosphere of any planet (including the Earth) in the Solar system can be compared to a *multi-layer model*

with fully absorbing layers (note that it is a very simplistic approach). Assume there are  $N$  such layers and the total rate of incident Solar radiation reaching the surface is  $E_{in}$ . Each of the layers also emits radiation to the other two layers directly above/below it. The rate of radiative emission for the  $j$ -th layer that has a temperature  $T_j$  is  $E_j = \sigma T_j^4$  according to the *Stefan–Boltzmann Law*, with  $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ . The overall scenario can be seen in Figure 1.2. Formulate a linear system that represents the energy equilibrium (incoming radiation = outgoing radiation) of all layers and the surface, with  $E_j$  being the unknowns, over  $j = 1, 2, \dots, N, N + 1$ .

*Solution.* Considering the energy equilibrium for the first (topmost) layer, we have

$$-2E_1 + E_2 = 0$$

Going down to the second layer, it is

$$E_1 - 2E_2 + E_3 = 0$$

In general, for the  $j$ -th layer in the middle, where  $j$  runs from 2 to  $N$ , we can similarly obtain

$$E_{j-1} - 2E_j + E_{j+1} = 0$$

Finally, for the surface (the  $N + 1$ -th layer), we have

$$\begin{aligned} E_N - E_{N+1} + E_{in} &= 0 \\ E_N - E_{N+1} &= -E_{in} \end{aligned}$$

Summarizing all the  $N + 1$  equations, they can be expressed in matrix form as

$$\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 \\ \vdots & & & \ddots & & \vdots & \\ 0 & 0 & 0 & & -2 & 1 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{N-1} \\ E_N \\ E_{N+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ -E_{in} \end{bmatrix}$$

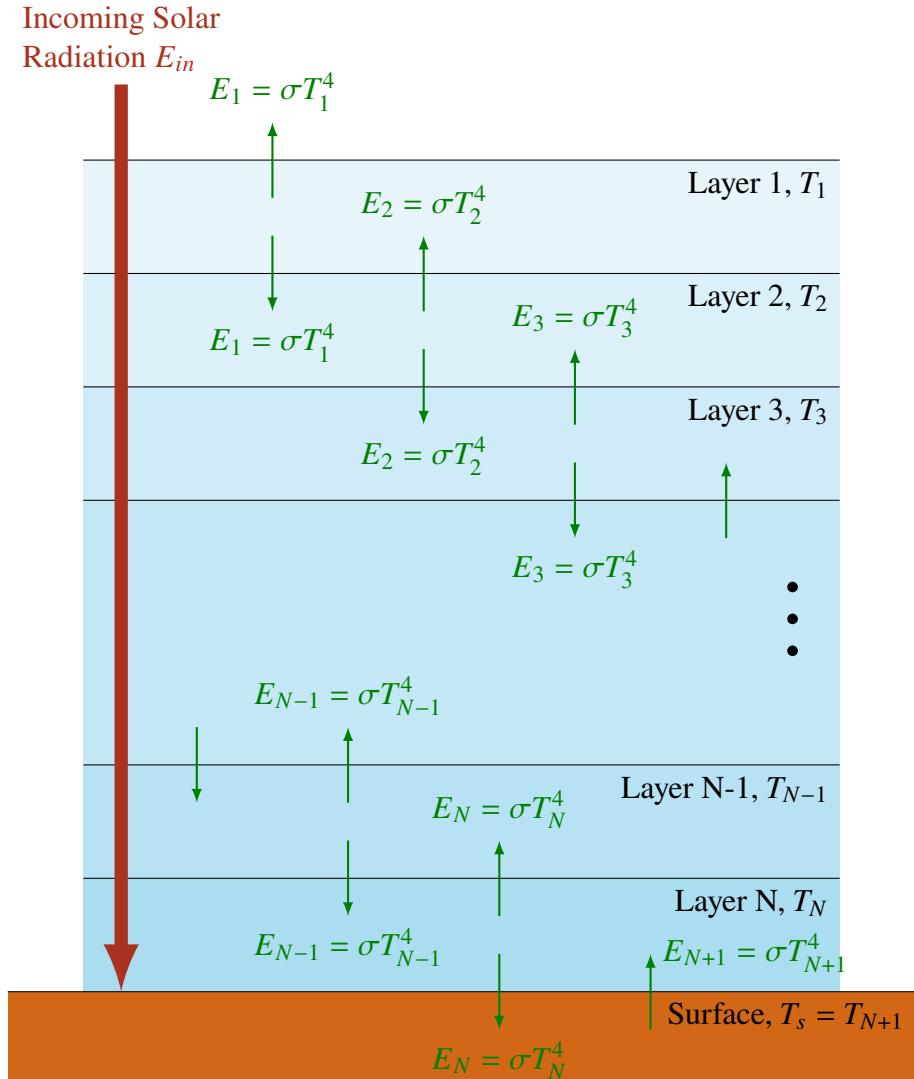


Figure 1.2: The atmospheric profile with multiple ( $N$ ) absorbing layers in Example 1.4.2. The surface is treated as an extra  $N+1$ -th layer.

Particularly, for  $N = 4$ , it is

$$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -E_{in} \end{bmatrix}$$

□

**Example 1.4.3.** The seawater in oceans contains a variety of dissolved salts in the form of ions. Most of them are sodium ( $\text{Na}^+$ ), magnesium ( $\text{Mg}^{2+}$ ), chlorine ( $\text{Cl}^-$ ) and sulphate ( $\text{SO}_4^{2-}$ ). Consider a sample of seawater and assume the concentration of other ions are negligible. Their are two major constraints over the individual concentrations of each type of ions ( $n = [\text{Na}^+]$ ,  $m = [\text{Mg}^{2+}]$ ,  $c = [\text{Cl}^-]$ ,  $s = [\text{SO}_4^{2-}]$ ). First, the overall charge of the seawater has to be neutral. Second, their concentrations should add up to the measured salinity (the total mass concentration of salts, inferred by electrical conductivity). It is given that the salinity of the seawater sample is 34 psu (1 psu =  $1 \text{ g kg}^{-1}$  which is the unit preferred in oceanography). Write down the corresponding linear system that is consisted of two equations for this situation.

(WIP)

**Example 1.4.4.** There are four weather stations in proximity. Each of them measures the local air temperature  $T_i$ , where  $i = 0, 1, 2, 3$ . Assume that the spatial pattern of temperature over the region approximately follows a linear gradient such that both  $\partial T / \partial x$  and  $\partial T / \partial y$  can be treated as constants. Assign the location of the first station to be the origin  $(0, 0)$ , and the relative locations of the second/third/fourth station are  $(10, 20)$ ,  $(25, 15)$ ,  $(-10, 5)$  (in km). The measured temperature of the four stations at some time are  $27.1^\circ\text{C}$ ,  $27.3^\circ\text{C}$ ,  $27.4^\circ\text{C}$ ,  $26.9^\circ\text{C}$ . Set up a linear system for finding  $\partial T / \partial x$  and  $\partial T / \partial y$ .

*Solution.* Since the temperature gradients  $\partial T / \partial x$  and  $\partial T / \partial y$  are assumed to be constants, we have, by Taylor expansion in both  $x$  and  $y$ ,

$$T_i = T_0 + \frac{\partial T}{\partial x}(\Delta x) + \frac{\partial T}{\partial y}(\Delta y)$$

for  $i = 1, 2, 3$  where  $\Delta x$  and  $\Delta y$  are the  $x/y$ -distances relative to the station at the origin. Substituting the provided data, we get

$$\begin{cases} 27.3 &= 27.1 + \frac{\partial T}{\partial x}(10) + \frac{\partial T}{\partial y}(20) \\ 27.4 &= 27.1 + \frac{\partial T}{\partial x}(25) + \frac{\partial T}{\partial y}(15) \\ 26.9 &= 27.1 + \frac{\partial T}{\partial x}(-10) + \frac{\partial T}{\partial y}(5) \end{cases}$$

Reorganizing them gives

$$\begin{cases} 10\frac{\partial T}{\partial x} + 20\frac{\partial T}{\partial y} &= 0.2 \\ 25\frac{\partial T}{\partial x} + 15\frac{\partial T}{\partial y} &= 0.3 \\ -10\frac{\partial T}{\partial x} + 5\frac{\partial T}{\partial y} &= -0.2 \end{cases}$$

The matrix form of this linear system will then be

$$\begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \begin{bmatrix} \frac{\partial T}{\partial x} \\ \frac{\partial T}{\partial y} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix}$$

□

We will talk about how to solve the linear systems in these four examples in Section 3.3.

## 1.5 Python Programming

We will use the package `numpy` and `scipy` throughout the book to solve linear algebra problems via *Python* programming. First, we can define a 2D `numpy` array that works as a matrix.

```
import numpy as np
myMatrix1 = np.array([[1, 4], [5, 3]])
print(myMatrix1)
```

which gives

```
[[1 4]
 [5 3]]
```

representing the matrix

$$\begin{bmatrix} 1 & 4 \\ 5 & 3 \end{bmatrix}$$

We can similarly define another matrix:

```
myMatrix2 = np.array([[1, 3], [5, 6]])
```

Addition, subtraction, and scalar multiplication are straight-forward.

```
myMatrix3 = 3*myMatrix1 - 4*myMatrix2
print(myMatrix3)
```

The above code produces

```
[[ -1   0]
 [-5 -15]]
```

and you can verify the answer by hand. Meanwhile, matrix product is done by the function `np.matmul()`.

```
myMatrix4 = np.matmul(myMatrix1, myMatrix2) # or equivalently
                                              myMatrix1 @ myMatrix2
print(myMatrix4)
```

gives

```
[[21 27]
 [20 33]]
```

To select a specific entry, use indexing by square brackets. The first index/second index represents row/column. Beware that each index starts at zero in *Python*. So putting the number 1 in the first/second index actually means the second row/column. So

```
print(myMatrix4[1,0])
```

refers to the entry at row 2, column 1 of `myMatrix4` which is 20. Also, we can select the  $i$ -th row (or the  $j$ -th column) by `<Matrix>[i-1, :]` (`<Matrix>[:, j-1]`), where the colon `:` implies selecting along the entire row (column). For example,

```
print(myMatrix3[0,:])
print(myMatrix4[:,1])
```

gives `[ -1 0 ]` and `[ 27 33 ]` respectively. Now let's see how to perform elementary row operations. It will be easier and less error-prone if we copy the array before performing the operations.

```
myMatrix5 = np.copy(myMatrix4)
myMatrix5[0,:] = myMatrix5[0,:]/3
print(myMatrix5)
```

The lines above, when executed, divide the second row of `myMatrix5` (which is a copy of `myMatrix4`) by 3, and give

```
[[ 7   9]
 [20  33]]
```

Meanwhile, the subsequent lines below

```
myMatrix5[1,:] = myMatrix5[1,:] - 2*myMatrix5[0,:]
print(myMatrix5)
```

proceed to subtract 2 times the first row from the second row, and produce

```
[[ 7   9]
 [ 6  15]]
```

Row interchange is a bit more tricky.

```
myMatrix6 = np.copy(myMatrix4)
myMatrix6[[0, 1],:] = myMatrix6[[1, 0],:]
```

This swaps the first and second row. (You can swap columns in a similar way.) Printing out the new matrix by `print(myMatrix6)` shows

```
[[20  33]
 [21  27]]
```

An important pitfall is that, since our inputs to `np.array` are all integers, the previous arrays will automatically have a data type of `int` (integer). This may produce unexpected errors when the calculation leads to decimals/fractions. If it is the case, then we can avoid such bugs by declaring the array with the keyword `dtype=float` to use *floating point numbers*, like

```
myMatrix1 = np.array([[1, 4], [5, 3]], dtype=float)
```

when printed out via `print(myMatrix1)` it gives

```
[[1. 4.]  
 [5. 3.]]
```

Notice the newly appeared decimal points after the original integers. Alternatively, we can add decimal points to the integer entries during the array initialization, as

```
myMatrix1 = np.array([[1., 4.], [5., 3.]])
```

## 1.6 Exercises

**Exercise 1.1** Let

$$A = \begin{bmatrix} 1 & 2 \\ 5 & -1 \end{bmatrix} \quad B = \begin{bmatrix} -4 & 3 \\ -2 & 7 \end{bmatrix}$$

Find:

(a)  $A + B$ ,

(b)  $2A - \frac{3}{2}B$ ,

(c)  $AB$ ,

(d)  $BA$ .

**Exercise 1.2** Let

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

Find:

- (a)  $AB$ ,
- (b)  $BA$ .

**Exercise 1.3** Let

$$A = \begin{bmatrix} 4 & 6 \\ 3 & 3 \end{bmatrix}$$

$$B = \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$$

$$C = \begin{bmatrix} 3 & 9 & 1 \\ 4 & 3 & -1 \end{bmatrix}$$

Find:

- (a)  $(A + B)C$ ,
- (b)  $AC + BC$ ,
- (c)  $(AB)C$ ,
- (d)  $A(BC)$ .

**Exercise 1.4** Let

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 7 & 2 & -1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 5 & -2 \\ 4 & 3 & 1 \\ 0 & 2 & 3 \end{bmatrix}$$

Find:

- (a)  $(A + B)(2A - B)$ ,
- (b)  $(\frac{3}{2}A - B)(-A + \frac{1}{2}B)$ .

**Exercise 1.5** Let

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 6 \\ 5 & 2 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 3 & 8 \\ 4 & 0 & 7 \end{bmatrix}$$

Find:

- (a)  $A^2$ ,
- (b)  $B^2$ ,
- (c)  $AB$ ,
- (d)  $BA$ .

**Exercise 1.6** Rewrite the following system of linear equations in matrix form.

$$\begin{cases} 3y - 4z = 6 \\ 5x - y + 2z = 13 \\ 6x + z = 8 \end{cases}$$

**Exercise 1.7** For the following matrix,

$$\begin{bmatrix} 2 & 3 & 5 & 7 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 6 & 10 \end{bmatrix}$$

Find the results if the following elementary row operations are applied on it:

- (a) Multiplying the third row by a factor of 2, and then subtracting the third row by the second row,

- (b) Adding the first row by 3 times the third row, and then interchanging the first and second row, and finally subtract the third row by 2 times the first row.

**Exercise 1.8** The *dry adiabatic lapse rate*, which is the rate of decrease in air temperature when an unsaturated air parcel rises, is about  $\Gamma_{dry} = 9.8 \text{ }^{\circ}\text{C km}^{-1}$ . When the temperature of the air parcel falls below the *dew point*, the air saturates and condensation occurs. Typically, dew point temperature of an air parcel will decrease at a rate of roughly  $\Gamma_{dew} = 2 \text{ }^{\circ}\text{C km}^{-1}$ . Now, an air parcel with an initial air temperature/dew point temperature of  $T_{a,ini} = 25.4 \text{ }^{\circ}\text{C} / T_{dew,ini} = 17.8 \text{ }^{\circ}\text{C}$  at the ground starts to rise. Let  $z_{cd}$  and  $T_{cd}$  be the height above the ground (in km) and temperature (in  $^{\circ}\text{C}$ ) of the air parcel when condensation occurs. Construct a linear system with  $z_{cd}$  and  $T_{cd}$  as the unknowns to represent this situation.

**Exercise 1.9** In some ancient Chinese Mathematics texts, the problem of *Chickens and Rabbits in the Same Cage* was posed. "Now there are some chickens and rabbits placed in the same cage, with a total number of 35 heads and 94 legs. How many chickens and rabbits are there respectively?" Given the fact that a chicken (rabbit) has two (four) legs (and obviously only one head), write down the corresponding linear system in terms of the numbers of chickens  $x$  and rabbits  $y$ .



## Chapter 2

# Inverses and Determinants

---

In this chapter, we are going to discuss two important concepts about matrices, which are their *inverses* and *determinants*. They will appear from time to time in the remaining parts of this book. To derive them, we need to introduce some prerequisite ideas first, including the *identity matrix*, *transpose*, and the methods of *Gaussian Elimination* and *cofactor expansion*.

## 2.1 Identity Matrices and Transpose

### 2.1.1 Identity Matrices

One important class of matrices is the *identity matrices*. They are  $n \times n$  square matrices, where  $n$  can be any positive integer, with entries along the *main diagonal* (where index of row = column) being 1 and other off-diagonal elements being 0. Usually, they are denoted by  $I_n$ , or simply  $I$ .

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Identity matrices of size  $2 \times 2$  and  $3 \times 3$  with the main diagonal 1s highlighted.

**Definition 2.1.1** (Identity Matrix). An identity matrix of the square shape  $n \times n$   $I_n$  is defined as  $[I_n]_{ij} = 1$ , for  $i = j$ , and  $[I_n]_{ij} = 0$ , for  $i \neq j$ , where  $1 \leq i, j \leq n$ .

Short Exercise: Explicitly write down  $I_5$ .<sup>1</sup>

One important property of identity matrices is

**Properties 2.1.2.** Matrix product between any matrix  $A$  with an identity matrix  $I$  always returns  $A$  whenever the matrix product is defined. If  $A$  is of the shape  $m \times n$ , then  $AI_n = I_m A = A$ . If  $A$  is now a square matrix such that  $m = n$  (and  $I_m = I_n = I$ ), then we have  $AI = IA = A$ .

In other words, the identity  $I$  can be regarded to be the "1" in the world of matrices. This is one of the cases that  $AB = BA$  commutes (if both of them are square and either one of them is the identity matrix). Using the matrix

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

as an example, the readers can try to compute  $AI_3$  and  $I_2A$  to see if the results are  $A$  itself.<sup>2</sup>

$${}^1I_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} {}^2 AI_3 &= \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} (a)(1) + (b)(0) + (c)(0) & (a)(0) + (b)(1) + (c)(0) & (a)(0) + (b)(0) + (c)(1) \\ (d)(1) + (e)(0) + (f)(0) & (d)(0) + (e)(1) + (f)(0) & (d)(0) + (e)(0) + (f)(1) \end{bmatrix} \\ &= \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} = A \end{aligned}$$

The calculation of  $I_2A = A$  is similar.

## 2.1.2 Transpose

**Transpose** of a matrix, denoted by adding the superscript  $T$ , is formed by interchanging its rows and columns, that is, flipping the elements about the main diagonal.

**Definition 2.1.3** (Transpose). The transpose of an  $m \times n$  matrix  $A$ , denoted as  $A^T$ , is formed according to the relation  $[A^T]_{pq} = A_{qp}$ ,  $1 \leq p \leq n$ ,  $1 \leq q \leq m$ , i.e. swapping the row and column indices. Now  $A^T$  is an  $n \times m$  matrix.

Two examples are given below to show the effect of applying transpose on matrices.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 4 & 3 \\ 2 & -2 & 0 \\ -3 & 1 & 4 \end{bmatrix} \quad B^T = \begin{bmatrix} 1 & 2 & 3 \\ -4 & -2 & 1 \\ 3 & 0 & 4 \end{bmatrix}$$

Particularly, in the second example, we have outlined the main diagonal of  $B$  (as well as  $B^T$ ) and how the elements flip about it when transpose is carried out. Some useful properties about transpose are listed as follows.

**Properties 2.1.4.** For two matrices  $A$  and  $B$ , we have

1.  $(cA)^T = cA^T$ , where  $c$  is any constant;
2.  $(A^T)^T = A$ , i.e. transposing twice returns the original matrix (which is obvious);
3.  $(A \pm B)^T = A^T \pm B^T$ , if  $A$  and  $B$  have the same shape;
4.  $(AB)^T = B^T A^T$ , if  $A$  and  $B$  are conformable;

5.  $A_{kk} = A_{kk}^T$  for any  $k$  that  $A_{kk}$  is defined, i.e. the main diagonal is unaffected by transpose.

Short Exercise: Show that  $(ABC)^T = C^T B^T A^T$  if the matrices have compatible shapes for the matrix multiplication.<sup>3</sup>

### 2.1.3 Symmetric Matrices

A **symmetric matrix** has its elements mirrored about the main diagonal. Taking transpose of such a matrix will leave it unchanged. Implicitly, it is required to be a square matrix.

**Definition 2.1.5** (Symmetric Matrix). If an  $n \times n$  square matrix  $A$  and its transpose  $A^T$  are equal, i.e.  $A_{pq} = [A^T]_{pq} = A_{qp}$  for all  $1 \leq p, q \leq n$ , or simply  $A = A^T$ , then  $A$ , and also  $A^T$ , are symmetric.

As an example,

$$\begin{bmatrix} 1 & 2 & -2 \\ 2 & 0 & 4 \\ -2 & 4 & 3 \end{bmatrix}$$

is a  $3 \times 3$  symmetric matrix.

Short Exercise: Show that  $Y = XX^T$  and  $Z = X^TX$  are symmetric for any matrix  $X$ .<sup>4</sup>

In contrast, we also have **skew-symmetric matrices** such that  $A^T = -A$ . This automatically requires elements along the main diagonal to be all zeros.

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

A  $3 \times 3$  skew-symmetric matrix.

---

<sup>3</sup>By (4),  $(ABC)^T = ((AB)(C))^T = C^T(AB)^T = C^T B^T A^T$

<sup>4</sup>By Properties 2.1.4,  $Y^T = (XX^T)^T = (X^T)^T(X)^T = XX^T = Y$ , similar goes for  $Z = X^TX$ .

## 2.2 Inverses

### 2.2.1 Definition and Properties of Inverses

**Inverse** of a square matrix, denoted by appending the superscript  $^{-1}$ , is another square matrix such that the matrix product between these two matrices (in either order) yields an identity matrix.

**Definition 2.2.1** (Inverse). An  $n \times n$  square matrix  $B$  is said to be the inverse of another  $n \times n$  square matrix  $A$  if  $AB = BA = I_n$ . This inverse matrix is denoted as  $B = A^{-1}$ , and the relation becomes  $AA^{-1} = A^{-1}A = I$ . The opposite direction also holds, i.e.  $A$  is the inverse of  $A^{-1}$ . Hence, we say that  $A$  and  $A^{-1}$  are the inverse of each other.

If there exists an inverse  $A^{-1}$  for the square matrix  $A$ , then both  $A$  and  $A^{-1}$  are called **invertible**. Otherwise,  $A$  is said to be **singular**. This is another situation in which a matrix product  $AB = BA$  (if  $B = A^{-1}$ ) can commute.<sup>5</sup>

In the last chapter, we only define addition, subtraction, and multiplication for matrices, omitting division like it is an elephant in the room. The inverse serves as a remedy for this by acting as the reciprocal in the world of matrices. This allows us to "divide" on both sides of a matrix equation provided the relevant inverse exists. Remember, in the last chapter, we mentioned that cancellation may not work for something like  $AB = AC$ . But if  $A^{-1}$  exists, then by multiplying it to the left of both sides of the equation, we can effectively "divide by  $A$ "

$$AB = AC$$

---

<sup>5</sup> $AA^{-1} = I$  implies  $A^{-1}A = I$  and vice versa. However, while looking innocent, showing this is actually not trivial and prone to circular logic. A heuristic way to "prove" it is to note that

$$\begin{aligned} AA^{-1} &= I \\ AA^{-1}A &= IA \\ A(A^{-1}A) &= A \end{aligned}$$

which implies that multiplying  $A$  by  $A^{-1}A$  returns  $A$  itself, so it should be reasonable to assume  $A^{-1}A = I$ .

$$\begin{aligned}
 A^{-1}AB &= A^{-1}AC \\
 (A^{-1}A)B &= (A^{-1}A)C && (\text{Properties 1.1.2}) \\
 IB &= IC && (\text{Definition 2.2.1}) \\
 B &= C && (\text{Properties 2.1.2})
 \end{aligned}$$

so that cancellation holds in this situation. Take the matrix equation  $AG = H$  as another example, if  $A$  has an inverse  $A^{-1}$ , then we may do a matrix "division" as follows:

$$\begin{aligned}
 AG &= H \\
 A^{-1}AG &= A^{-1}H \\
 ((A^{-1}A)G = IG =) G &= A^{-1}H && (\text{Properties 1.1.2 and 2.1.2, Definition 2.2.1})
 \end{aligned}$$

In addition, the inverse of a matrix, if exists, must be unique.

**Properties 2.2.2** (Uniqueness of Inverse). If  $A$  has an inverse  $A^{-1}$ , it is unique.

*Proof.* This property can be proved easily by first assuming that the invertible matrix  $A$  has two different inverses,  $B$  and  $C$ . Subsequently, by Definition 2.2.1, we have  $BA = I$  (and also  $AC = I$ ). Multiplying by  $C$  to the right on both sides gives

$$\begin{aligned}
 BAC &= IC \\
 B(AC) &= C && (\text{Properties 1.1.2 and 2.1.2}) \\
 B(I) &= C && (AC = I \text{ from assumption}) \\
 B &= C && (\text{Properties 2.1.2})
 \end{aligned}$$

So,  $B$  and  $C$  are actually the same matrix, implying that the inverse of  $A$  is unique.  $\square$

**Example 2.2.1.** Let

$$A = \begin{bmatrix} 4 & 6 \\ 3 & 5 \end{bmatrix} \quad B = \begin{bmatrix} \frac{5}{2} & -3 \\ -\frac{3}{2} & 2 \end{bmatrix}$$

Show that  $A$  and  $B$  are inverse to each other.

*Solution.*

$$\begin{aligned} AB &= \begin{bmatrix} 4 & 6 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} \frac{5}{2} & -3 \\ -\frac{3}{2} & 2 \end{bmatrix} \\ &= \begin{bmatrix} (4)(\frac{5}{2}) + (6)(-\frac{3}{2}) & (4)(-3) + (6)(2) \\ (3)(\frac{5}{2}) + (5)(-\frac{3}{2}) & (3)(-3) + (5)(2) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2 \end{aligned}$$

We leave it to the readers for showing  $BA = I$  too as an exercise. Hence,  $AB = BA = I$ ,  $A$  and  $B$  are indeed the inverse of each other.  $\square$

The followings are some properties of inverses.

**Properties 2.2.3.** If a square matrix  $A$  is invertible and has an inverse  $A^{-1}$ , then

1.  $(cA)^{-1} = \frac{1}{c}A^{-1}$ , for any constant  $c \neq 0$ ;
2.  $(A^{-1})^{-1} = A$ , i.e. the inverse of an inverse returns the original matrix;
3.  $(A^n)^{-1} = (A^{-1})^n$ , for any positive integer  $n$ ;
4.  $(AB)^{-1} = B^{-1}A^{-1}$ , provided that  $B$  is invertible too (and they are square matrices of the same size);
5.  $(A^T)^{-1} = (A^{-1})^T$ .

However,  $(A \pm B)^{-1}$  may not be equal to  $A^{-1} \pm B^{-1}$ , or even may be singular. We shall briefly prove (4) here.

*Proof.* It is given that  $A$  and  $B$  is invertible, and by Definition 2.2.1, we have  $AA^{-1} = I$ , as well as

$$BB^{-1} = I$$

Multiplying by  $A$  and  $A^{-1}$  to the left and right on both sides of above respectively yields

$$\begin{aligned} ABB^{-1}A^{-1} &= AIA^{-1} \\ AB(B^{-1}A^{-1}) &= (AI)A^{-1} = AA^{-1} \quad (\text{Properties 1.1.2 and 2.1.2}) \\ &= I \quad (\text{Definition 2.2.1}) \end{aligned}$$

This shows that multiplying  $AB$  by  $B^{-1}A^{-1}$  produces an identity matrix, and therefore  $(AB)^{-1} = B^{-1}A^{-1}$  is the unique inverse of  $AB$  by Definition 2.2.1 and Properties 2.2.2.  $\square$

Short Exercise: Show that  $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$  if  $A$ ,  $B$  and  $C$  are invertible and conformable.<sup>6</sup>

(4) of Properties 2.2.3 explicitly shows that the product  $AB$  is invertible if  $A$  and  $B$  are themselves invertible. The converse is actually true as well.<sup>7</sup> Hence

**Properties 2.2.4.** For two square matrices  $A$  and  $B$ ,  $AB$  is invertible if and only if  $A$  and  $B$  are invertible.

## 2.2.2 (Reduced) Row Echelon Form

Naturally, the next question is how to compute the inverse of any square matrix. For this, we have to understand a specific form of matrices called the (*reduced*)

---

<sup>6</sup>By (4),  $(ABC)^{-1} = ((AB)(C))^{-1} = C^{-1}(AB)^{-1} = C^{-1}B^{-1}A^{-1}$

<sup>7</sup>Let's assume  $AB$  is invertible and has an inverse  $C = (AB)^{-1}$ , hence we have  $(AB)C = I$  by Definition 2.2.1, (notice that  $A$ ,  $B$ , and  $C$  are all square matrices of the same extent) and by Properties 1.1.2,  $A(BC) = I$ . Using Definition 2.2.1 (as well as Properties 2.2.2) again, we immediately identify  $BC$  as the inverse of  $A$  and  $A$  is invertible. The case for  $B$  is similarly proved.

**row echelon form** first. A matrix is in reduced row echelon form (*rref*) when it satisfies the following requirements.

**Definition 2.2.5** ((Reduced) Row Echelon Form). A matrix is in row echelon form if

1. The first non-zero number in every row is 1, which is known as the "*Leading 1*" (sometimes referred to as a *pivot*),
2. "*Leading 1*" of a lower row must appear farther to the right than that of any higher row,
3. Any row consisted of all zeros is placed at the bottom;
4. If additionally, any column containing a leading 1 (sometimes called a *pivotal column*) have zeros elsewhere in that column, then it is in *reduced* row echelon form.

It is apparent that all identity matrices are in (reduced) row echelon form. Examples of row echelon form (but not *reduced*), with the leading 1s highlighted are

$$\begin{array}{ll} A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} & B = \begin{bmatrix} 1 & 3 & 1 & 2 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ C = \begin{bmatrix} 1 & 4 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} & D = \begin{bmatrix} 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Meanwhile, examples of *reduced* row echelon form are

$$\begin{array}{ll} G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} & H = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

The following matrices are *not* in row echelon form. (why?)<sup>8</sup>

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

Short Exercise: Decide if the following matrices are in (reduced) row echelon form or not.<sup>9</sup>

$$X = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We have studied about elementary row operations in the last chapter, which now can be used to transform matrices into their reduced row echelon form. The procedure is comprised of two major parts, the *forward phase*, converting the matrix to row echelon form first, and the *backward phase*, eventually transforming it into reduced row echelon form. The first phase is also named **Gaussian Elimination**, and together they are called **Gauss-Jordan Elimination**<sup>10</sup>. We demonstrate the entire procedure using an example.

**Example 2.2.2.** Carry out Gauss-Jordan Elimination on the following matrix to make it become reduced row echelon form.

$$A = \begin{bmatrix} 2 & 0 & 4 & 6 \\ 3 & 3 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

*Solution.* At each step of the forward phase, the strategy is to look at on the leftmost column that has at least one non-zero entries (any column consisting of full zeros is ignored). Along that column, we either find an existing leading 1, or create a leading 1 via multiplying some row having a starting entry  $a$  that

---

<sup>8</sup> $P$  violates (2) and  $Q$  does not satisfy (1) and (3) of Definition 2.2.5.

<sup>9</sup>Yes, Yes (reduced), No.

<sup>10</sup>Often we just write Gaussian Elimination in place of Gauss-Jordan Elimination.

is as large as possible in magnitude, by the constant  $1/a$ . (The leading entry selected by this algorithm is commonly called the *pivot*, and the process is called *pivoting*.) The row holding the leading 1 is subsequently put at the top, by an interchanging of rows if needed. In this example, such rows will be highlighted in red.

$$\begin{bmatrix} 2 & 1 & 4 & 6 \\ 3 & 3 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 3 & 3 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix} \quad \frac{1}{2}R_1 \rightarrow R_1$$

We have picked the first row  $R_1$  for the leading 1 through multiplying it by a factor of  $\frac{1}{2}$  here, but a leading 1 can be obtained from the other two rows as well. Subsequently, we make all the elements below the leading 1 along that *pivotal column* become zero, by adding the top row (which holds the leading 1), times  $-a_i$  (where  $a_i$  is the corresponding leading entry of row  $i$ ) to the other rows. Those zeros produced in this way will be highlighted in blue.

$$\begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 3 & 3 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & \frac{3}{2} & -5 & -9 \\ 1 & 2 & 3 & 4 \end{bmatrix} \quad R_2 - 3R_1 \rightarrow R_2$$

$$\rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & \frac{3}{2} & -5 & -9 \\ 0 & \frac{3}{2} & 1 & 1 \end{bmatrix} \quad R_3 - R_1 \rightarrow R_3$$

The first iteration is finished. We now repeat the same process over the remaining submatrix made up of elements that are not yet highlighted in colour, from left to right recursively.

$$\begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & \frac{3}{2} & -5 & -9 \\ 0 & \frac{3}{2} & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & -\frac{10}{3} & -6 \\ 0 & \frac{3}{2} & 1 & 1 \end{bmatrix} \quad \frac{2}{3}R_2 \rightarrow R_2$$

$$\rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & -\frac{10}{3} & -6 \\ 0 & 0 & 6 & 10 \end{bmatrix} \quad R_3 - \frac{3}{2}R_2 \rightarrow R_3$$

$$\rightarrow \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & -\frac{10}{3} & -6 \\ 0 & 0 & 1 & \frac{5}{3} \end{bmatrix} \quad \frac{1}{6}R_3 \rightarrow R_3$$

Now, all entries below every leading 1 are zeros, and the forward phase is completed. We have obtained the row echelon form as an intermediate product. The backward phase is done similarly but in a bottom-up fashion, from right to left. By adding appropriate multiples of lower rows to higher rows, we turn all the non-zero elements above the leading 1 along every pivotal column into zeros. Non-pivotal columns (the last column here) are ignored.

$$\begin{array}{l}
 \left[ \begin{array}{cccc} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & -\frac{10}{3} & -6 \\ 0 & 0 & 1 & \frac{5}{3} \end{array} \right] \rightarrow \left[ \begin{array}{cccc} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{array} \right] \quad R_2 + \frac{10}{3}R_3 \rightarrow R_2 \\
 \rightarrow \left[ \begin{array}{cccc} 1 & \frac{1}{2} & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{array} \right] \quad R_2 - 2R_3 \rightarrow R_1 \\
 \rightarrow \left[ \begin{array}{cccc} 1 & 0 & 0 & -\frac{1}{9} \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{array} \right] \quad R_1 - \frac{1}{2}R_2 \rightarrow R_1
 \end{array}$$

The matrix is now in reduced row echelon form as required. The amount of leading 1s in the rref of the matrix is known as its *rank*, which equals to 3 here.  $\square$

Short Exercise: Repeat the example above but start by interchanging  $R_1$  and  $R_3$ . <sup>11</sup>

From the short exercise above, we can see that even if we apply different elementary row operations (particularly for the creation of leading 1s) during Gauss-Jordan Elimination, we will acquire the same reduced echelon form in the end. In fact,

**Theorem 2.2.6** (Uniqueness of Reduced Row Echelon Form). Reduced row echelon form of a matrix is unique.

We shall omit the proof here. The following properties further reveal how elementary row operations are associated with reduced row echelon form.

---

<sup>11</sup>For checking, after the first iteration, it will be (WIP) and the end result will be the same.

**Properties 2.2.7.** If a matrix can be transformed into another matrix by elementary row operations, they are said to be *row equivalent*.

Since for any pair of row equivalent matrices, either of them can be transformed into the other one by elementary row operations, and hence can be further transformed into the reduced row echelon form of the other matrix, by Theorem 2.2.6, the uniqueness of rref implies that

**Properties 2.2.8.** Row equivalent matrices have the same reduced row echelon form. Particularly, they are row equivalent to this rref. If two matrices have different reduced row echelon forms, then they are not row equivalent, and vice versa.

Let's go through one more simple example about Gauss-Jordan Elimination.

**Example 2.2.3.** Transform the following matrix into reduced row echelon form.

$$A = \begin{bmatrix} 2 & 2 & 1 \\ 6 & 4 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{bmatrix}$$

*Solution.* One possible way to do the forward elimination is

$$\begin{array}{c}
 \left[ \begin{array}{ccc} 2 & 2 & 1 \\ 6 & 4 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc} 6 & 4 & 1 \\ 2 & 2 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{array} \right] \\
 R_1 \leftrightarrow R_2
 \end{array}$$
  

$$\begin{array}{c}
 \rightarrow \left[ \begin{array}{ccc} 1 & \frac{2}{3} & \frac{1}{6} \\ 2 & 2 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{array} \right] \\
 \frac{1}{6}R_1 \rightarrow R_1
 \end{array}$$
  

$$\begin{array}{c}
 \rightarrow \left[ \begin{array}{ccc} 1 & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & 0 \\ 0 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & -\frac{1}{3} \end{array} \right] \\
 R_2 - 2R_1 \rightarrow R_2 \\
 R_3 - 2R_1 \rightarrow R_3 \\
 R_4 - 2R_1 \rightarrow R_4
 \end{array}$$

$$\begin{aligned} & \rightarrow \begin{bmatrix} 1 & \frac{2}{3} & \frac{1}{6} \\ 0 & 1 & 1 \\ 0 & \frac{5}{3} & \frac{5}{3} \\ 0 & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} & \frac{3}{2}R_2 \rightarrow R_2 \\ & \rightarrow \begin{bmatrix} 1 & \frac{2}{3} & \frac{1}{6} \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & R_3 - \frac{5}{3}R_2 \rightarrow R_3 \\ & & R_4 + \frac{1}{3}R_2 \rightarrow R_4 \end{aligned}$$

The backward elimination is straight-forward.

$$\begin{bmatrix} 1 & \frac{2}{3} & \frac{1}{6} \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -\frac{1}{2} \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad R_1 - \frac{2}{3}R_2 \rightarrow R_1$$

The rank of the matrix can be readily seen to be 2. □

### 2.2.3 Finding Inverses by Gaussian Elimination

With Gaussian Elimination, obtaining the inverse  $A^{-1}$  of any invertible matrix  $A$  is now possible. We start by writing out an identity matrix  $I$  of the same shape and concatenate this identity matrix to the right of  $A$ , leading to an augmented form of  $[A|I]$ . Then we carry out elementary row operations simultaneously on both sides of  $[A|I]$  such that the matrix to the left, originally as  $A$ , is reduced to the identity matrix  $I$  by Gaussian Elimination. The identity matrix to the right will then be transformed into the desired inverse by the same set of elementary operations, such that the concatenated matrix will appear as  $[I|A^{-1}]$ ,

**Example 2.2.4.** Find the inverse of

$$A = \begin{bmatrix} 1 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 1 & 1 \end{bmatrix}$$

by Gaussian Elimination

*Solution.* Appending an  $3 \times 3$  identity matrix to the right, we have

$$\begin{array}{c}
 \left[ \begin{array}{ccc|ccc} 1 & 4 & 5 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 4 & 5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & -2 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array} \right] \quad R_2 - 2R_3 \rightarrow R_2 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 4 & 5 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{array} \right] \quad R_2 \leftrightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 4 & 5 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{array} \right] \quad R_2 - R_3 \rightarrow R_2 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 4 & 0 & 1 & -5 & 10 \\ 0 & 1 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{array} \right] \quad R_1 - 5R_3 \rightarrow R_1 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -1 & -2 \\ 0 & 1 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{array} \right] \quad R_1 - 4R_2 \rightarrow R_1
 \end{array}$$

Hence the required inverse is

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

□

Short Exercise: Verify the inverse of  $A^{-1}$  above is just  $A$  by the same method.  
[12](#)

The underlying reason why the above procedure can produce the inverse matrix is the equivalence between elementary row operations and multiplication by appropriate *elementary matrices*.

---

<sup>12</sup>You should be able to retrieve the matrix  $A$  back. The first column of  $A^{-1}$  already holds a leading 1 and elements below which are zeros. A possible next step is to multiply  $R_2$  by  $-1$  and then subtract  $R_3$  by  $R_2$ .

**Properties 2.2.9** (Elementary Matrices). Any elementary row operation on an  $m \times n$  matrix can be represented by multiplying it to the left with a suitable *elementary matrix*. Such a matrix is essentially the one appeared after applying that particular elementary row operation on an identity matrix. For the three types of elementary row operations described in Definition 1.3.1:

1.  $cR_p \rightarrow R_p, c \neq 0,$
2.  $R_p + cR_q \rightarrow R_p,$
3.  $R_p \leftrightarrow R_q$

their corresponding elementary matrices  $E$  are square ( $m \times m$ ), and *invertible* (see the following remark) in which

1.  $E_{kk} = 1$  for any  $k$ , except  $E_{pp} = c$ ;
2.  $E_{kk} = 1$  for all  $k$ , with  $E_{pq} = c$ ;
3.  $E_{kk} = 1$  for any  $k$ , except  $E_{pp} = 0$  and  $E_{qq} = 0$ , with  $E_{pq} = E_{qp} = 1$ .

Entries not mentioned are all zeros.

Since it is quite abstract, it is useful to have some actual examples.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Multiplying  $R_2$  by a factor of 2:  $2R_2 \rightarrow R_2$

$$\begin{bmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Adding 3 times  $R_2$  to  $R_1$ :  $R_1 + 3R_2 \rightarrow R_1$

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

Swapping  $R_1$  and  $R_3$ :  $R_1 \leftrightarrow R_3$

Any elementary row operation can be apparently undone by an inverse elementary row operation (addition vs subtraction, multiplication vs division ( $c \neq 0$ ),

swapping twice). Accordingly, any elementary matrix has another corresponding elementary matrix as its inverse, and the readers are invited to think about their forms in the exercise below.

**Short Exercise:** Write down the inverses of the three example elementary matrices above.<sup>13</sup>

For instance, consider a matrix

$$\begin{bmatrix} 1 & 4 & 3 \\ 2 & 5 & 1 \\ -1 & 0 & 2 \end{bmatrix}$$

then the action of subtracting  $R_2$  from  $R_3$ ,  $R_3 - R_2 \rightarrow R_3$ . can be expressed as

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

**Short Exercise:** Find out the  $3 \times 3$  elementary matrix for subtracting 2 times the third row from the first row. What happens when we apply this elementary matrix to the left of the matrix above?<sup>14</sup>

Now we are ready to see why finding inverses by Gaussian Elimination works.

**Theorem 2.2.10.** If a matrix  $A$  can be converted to an identity matrix  $I$  as its reduced row echelon form by Gaussian Elimination, then it is invertible since the same steps can in turn be applied on  $I$ , producing its inverse  $A^{-1}$ .

Using the language of Properties 2.2.8, the matrix  $A$  has to be row equivalent to  $I$  for  $A^{-1}$  to exist. This also means if Gaussian Elimination fails to reduce

$$^{13} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$^{14} \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} : \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 4 & 3 \\ 2 & 5 & 1 \\ -3 & -5 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 14 & 1 \\ 2 & 5 & 1 \\ -3 & -5 & 1 \end{bmatrix}$$

$A$  to  $I$  (i.e. the reduced row echelon form of  $A$  is some matrix other than the identity), then  $A^{-1}$  does not exist.

*Proof.* Assume  $A$  is invertible and hence  $AA^{-1} = I$  (Definition 2.2.1). From Properties 2.2.9, When doing Gaussian Elimination over  $A$ , the  $i$ -th elementary row operation executed can be represented by an elementary matrix, denoted as  $E_i$ , for  $i = 1, 2, \dots, n$  where  $n$  is the total number of steps. If we multiply these  $E_i$  successively to the left on both sides of the equation  $AA^{-1} = I$ , we have

$$\begin{aligned} E_n \cdots E_3 E_2 E_1 A A^{-1} &= E_n \cdots E_3 E_2 E_1 I \\ (E_n \cdots E_3 E_2 E_1 A) A^{-1} &= E_n \cdots E_3 E_2 E_1 I \quad (\text{Properties 1.1.2}) \\ (I) A^{-1} &= E_n \cdots E_3 E_2 E_1 I \\ A^{-1} &= E_n \cdots E_3 E_2 E_1 I \quad (\text{Properties 2.1.2}) \end{aligned}$$

from the second line to the third line, we have  $E_n \cdots E_3 E_2 E_1 A = I$  because the elementary row operations during Gaussian Elimination, represented by  $E_i$ ,  $i = 1, 2, \dots, n$ , reduce  $A$  to  $I$  as we demand in the assumption. With  $A^{-1} = E_n \cdots E_3 E_2 E_1 I$ , we immediately see that the same set of elementary matrices and hence elementary row operations can also transform  $I$  into  $A^{-1}$ , explicitly showing that  $A$  is invertible.  $\square$

As a corollary, because we have  $E_n \cdots E_3 E_2 E_1 A = I$  from above, and all  $E_i$  are invertible by Properties 2.2.9, we can multiply their inverses  $E'_i = E_i^{-1}$  (which are also elementary matrices), to the left on both sides successively, where  $i$  runs backwards from  $n$  to 1. This leads to

$$\begin{aligned} E_1^{-1} E_2^{-1} E_3^{-1} \cdots E_n^{-1} E_n \cdots E_3 E_2 E_1 A &= E_1^{-1} E_2^{-1} E_3^{-1} \cdots E_n^{-1} I \\ A &= E'_1 E'_2 E'_3 \cdots E'_n \end{aligned}$$

as each of the pairs  $E_n^{-1} E_n$ ,  $E_{n-1}^{-1} E_{n-1}$ ,  $\dots$ ,  $E_2^{-1} E_2$ ,  $E_1^{-1} E_1$  cancels out to produce  $I$ , and hence

**Properties 2.2.11.** All invertible matrices can be written as a product of some sequence of elementary matrices.

## 2.3 Determinants

### 2.3.1 Computing Determinants

The **determinant** of a *square* matrix  $A$ , denoted by  $\det(A)$  or  $|A|$ , is a number associated with certain intrinsic properties of the matrix which can help us to find its inverse (Determinant of non-square matrices is undefined). Determinant of a  $1 \times 1$  matrix is equal to the matrix's only entry. Determinants of  $2 \times 2$  and  $3 \times 3$  matrices can be calculated by a trick called **Sarrus' Rule**.

#### Sarrus' Rule

**Properties 2.3.1** (Sarrus' Rule). Determinants of size  $2 \times 2$  and  $3 \times 3$  matrices can be found by the Sarrus' Rule. For a  $2 \times 2$  matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Its determinant is computed by

$$\left| \begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right|$$

$$|A| = a_{11}a_{22} - a_{21}a_{12}$$

which is the product of elements crossed by the red arrow, minus the blue one. Similarly, for a  $3 \times 3$  matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Its determinant can be found by

$$\left| \begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{array} \right|$$

$$|A| = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) \\ - (a_{31}a_{22}a_{13} + a_{32}a_{23}a_{11} + a_{33}a_{21}a_{12})$$

**Example 2.3.1.** Find the determinant of the following matrix.

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

*Solution.* By Sarrus's Rule (Properties 2.3.1), we have

$$\begin{aligned} |A| &= \begin{vmatrix} 1 & 2 & 4 \\ -5 & 0 & -3 \\ 4 & 3 & 1 \end{vmatrix} \\ &= ((1)(0)(1) + (2)(-3)(4) + (4)(-5)(3)) \\ &\quad - ((4)(0)(4) + (3)(-3)(1) + (1)(-5)(2)) \\ &= (0 - 24 - 60) - (0 - 9 - 10) \\ &= -65 \end{aligned}$$

□

## Cofactor Expansion

Another commonly used method to calculate determinants is *Cofactor Expansion*, also known as *Laplace Expansion*. Before discussing cofactor expansion, it is necessary to know what *cofactors* are.

**Definition 2.3.2** (Cofactor and Minor). The **cofactor**  $C_{ij}$  at the  $(i, j)$  position of a matrix  $A$  is simply the determinant of the submatrix formed by deleting the  $i$ -th row and  $j$ -th column of  $A$ :  $M_{ij}$  (called the **minor** at  $(i, j)$ ), times the factor of  $(-1)^{i+j}$ , that is,  $C_{ij} = (-1)^{i+j} M_{ij}$ .

The  $(-1)^{i+j}$  factor can be visualized as a checkerboard pattern like

$$\begin{bmatrix} + & - & + & \dots \\ - & + & - & \\ + & - & + & \\ \vdots & & & \ddots \end{bmatrix}$$

So, for a matrix like

$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \\ 3 & 5 & 7 \end{bmatrix}$$

Its cofactor at  $(2, 1)$  is

$$\begin{aligned} C_{21} &= (-1)^{(2+1)} \begin{vmatrix} 3 & 5 \\ 5 & 7 \end{vmatrix} && \text{(Definition 2.3.2)} \\ &= (-1)((3)(7) - (5)(5)) && \text{(Properties 2.3.1)} \\ &= 4 \end{aligned}$$

Short Exercise: Find  $C_{13}$  and  $C_{32}$  for the matrix above.<sup>15</sup>

With **Cofactor (Laplace) Expansion**, the determinant of a matrix is computed as the sum of products between each entry and the corresponding cofactor along a picked row/column of it.

---

<sup>15</sup>  $C_{13} = (-1)^{1+3} \begin{vmatrix} 2 & 4 \\ 3 & 5 \end{vmatrix} = (1)((2)(5) - (3)(4)) = -2$ , similarly  $C_{32} = 4$ .

**Properties 2.3.3** (Cofactor/Laplace Expansion). The determinant of a  $n \times n$  square matrix  $A$ ,  $|A|$ , can be found by selecting either a fixed row  $i$ , or column  $j$ , and adding up the products of every element-cofactor pair along that row/column. For the former case (selected the  $i$ -th row), the determinant is computed as

$$\begin{aligned}|A| &= A_{i1}C_{i1} + A_{i2}C_{i2} + \cdots + A_{in}C_{in} \\ &= \sum_{k=1}^n A_{ik}C_{ik}\end{aligned}$$

For the latter case (fixed the  $j$ -th column), the determinant is similarly found by

$$\begin{aligned}|A| &= A_{1j}C_{1j} + A_{2j}C_{2j} + \cdots + A_{nj}C_{nj} \\ &= \sum_{k=1}^n A_{kj}C_{kj}\end{aligned}$$

where each of the cofactors  $C_{ij}$  is defined as in Definition 2.3.2. **Important:** regardless of which row or column is chosen, the result is always the same.

**Example 2.3.2.** Again, for the matrix

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

Find its determinant via cofactor expansion.

*Solution.* According to Properties 2.3.3, if we choose the first row to be expanded, its determinant is

$$\begin{aligned}|A| &= A_{11}C_{11} + A_{12}C_{12} + A_{13}C_{13} \\ &= (1)((-1)^{1+1} \begin{vmatrix} 4 & 6 \\ 5 & 7 \end{vmatrix}) + (3)((-1)^{1+2} \begin{vmatrix} 2 & 6 \\ 3 & 7 \end{vmatrix}) \\ &\quad + (5)((-1)^{1+3} \begin{vmatrix} 2 & 4 \\ 3 & 5 \end{vmatrix})\end{aligned}\tag{Definition 2.3.2}$$

$$= (1)(-2) + (3)(4) + (5)(-2) = 0 \quad (\text{Properties 2.3.1})$$

□

Short Exercise: Confirm the answer by carrying out cofactor expansion on another row or column.<sup>16</sup>

**Example 2.3.3.** Find the determinant of

$$A = \begin{bmatrix} 1 & 4 & 4 & 4 \\ 2 & 0 & 4 & 6 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 3 & 1 \end{bmatrix}$$

*Solution.* It is a  $4 \times 4$  matrix and we have to apply cofactor expansion. We can choose row or column that contains some zero(s) to reduce the computation. Here we pick the second column and by Properties 2.3.3, we have

$$\begin{aligned} |A| &= (4)(-1)^{1+2} \begin{vmatrix} 2 & 4 & 6 \\ 2 & 1 & 0 \\ 6 & 3 & 1 \end{vmatrix} + (0)(-1)^{2+2} \begin{vmatrix} 1 & 4 & 4 \\ 2 & 1 & 0 \\ 6 & 3 & 1 \end{vmatrix} \\ &\quad + (1)(-1)^{3+2} \begin{vmatrix} 1 & 4 & 4 \\ 2 & 4 & 6 \\ 6 & 3 & 1 \end{vmatrix} + (2)(-1)^{4+2} \begin{vmatrix} 1 & 4 & 4 \\ 2 & 4 & 6 \\ 2 & 1 & 0 \end{vmatrix} \end{aligned}$$

By Sarrus' Rule (Properties 2.3.1), we can calculate each of the four  $3 \times 3$  determinants (the detailed calculations are omitted, notice that we don't need to actually compute the second determinant) and obtain

$$|A| = (-4)(-6) + 0 + (-1)(50) + (2)(18) = 10$$

□

Finally, we can derive two simple results about determinants from the perspective of cofactor expansion.

---

<sup>16</sup>You should be able to get  $|A| = 0$ , no matter which row/column is selected.

**Properties 2.3.4.** If a matrix have a row/column with full zeros, or two identical/proportional rows/columns, then it has a determinant of zero.

The first case is trivial (just do the expansion along the row/column with full zeros) and we will show the second case alongside the introduction of the properties of determinants in the upcoming subsection.

### 2.3.2 Properties of Determinants

There are some notable properties about determinants. First of all, it is very easy to see that determinants for any  $n \times n$  identity matrix  $I_n$  is just 1. Second, there is a close relation between elementary row operations/elementary matrices and (their effects on) determinants, noted as follows.

**Properties 2.3.5.** The three types of elementary row operations in Definition 1.3.1, when applied on some square matrix  $A$ ,

1.  $cR_p \rightarrow R_p, c \neq 0,$
2.  $R_p + cR_q \rightarrow R_p,$
3.  $R_p \leftrightarrow R_q,$

change the determinant of  $A$  by a factor of  $c$ , 1 (unchanged), and  $-1$  (switching the sign), respectively.

**Properties 2.3.6.** The three types of elementary matrices  $E$  in Properties 2.2.9 that correspond to the elementary row operations in Definition 1.3.1,

1.  $E_{kk} = 1$  for any  $k$ , except  $E_{pp} = c$  ( $cR_p \rightarrow R_p, c \neq 0$ ),
2.  $E_{kk} = 1$  for all  $k$ , with  $E_{pq} = c$  ( $R_p + cR_q \rightarrow R_p$ ),
3.  $E_{kk} = 1$  for any  $k$ , except  $E_{pp} = 0$  and  $E_{qq} = 0$ , with  $E_{pq} = E_{qp} = 1$  ( $R_p \leftrightarrow R_q$ ),

have a determinant of  $c$ , 1, and  $-1$ , respectively.

Since the determinants of elementary matrices, by Properties 2.3.6, coincide exactly with the factors by how the determinant of a square matrix  $A$  changes when the corresponding elementary row operations are applied on  $A$  (represented by multiplication to the left of  $A$  by these elementary matrices) as shown in Properties 2.3.5, we conclude that

**Theorem 2.3.7.** For any elementary matrix  $E$  and a square matrix  $A$ , we have

$$\det(EA) = \det(E) \det(A)$$

This theorem will be of use when we later prove other properties of determinant. However, before doing so, we will demonstrate how to utilize Properties 2.3.5 (or equivalently 2.3.6) to ease the calculation of determinants.

**Example 2.3.4.** Re-do Example 2.3.3 utilizing Properties 2.3.5.

*Solution.* We can factor out the 2 in second row and subtract 3 times the third row from the fourth row. By Properties 2.3.5, we have

$$\begin{aligned} |A| &= \begin{vmatrix} 1 & 4 & 4 & 4 \\ 2 & 0 & 4 & 6 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 3 & 1 \end{vmatrix} = 2 \begin{vmatrix} 1 & 4 & 4 & 4 \\ 1 & 0 & 2 & 3 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 3 & 1 \end{vmatrix} \\ &= 2 \begin{vmatrix} 1 & 4 & 4 & 4 \\ 1 & 0 & 2 & 3 \\ 2 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{vmatrix} \end{aligned}$$

The determinant in the last line can be computed by doing cofactor expansion along the fourth row which now contains two zeros. With Properties 2.3.3 and 2.3.1, it is

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

$$= 0 + (-1)(9) + 0 + (1)(14) = 5$$

and hence  $|A| = 2(5) = 10$ . □

With Theorem 2.3.7, we can unearth the relation between invertibility of a square matrix and its determinant.

**Properties 2.3.8.** An invertible matrix has a non-zero determinant. Otherwise, a singular matrix has a determinant of zero.

*Proof.* Let's denote the matrix in question as  $A$ . For the case in which  $A$  is invertible, by Properties 2.2.11 it can be written as the product of some elementary matrices  $E_1, E_2, \dots, E_{n-1}, E_n$ , i.e.

$$A = E_1 E_2 \cdots E_{n-1} E_n$$

Taking the determinant of both sides, we have

$$\det(A) = \det(E_1 E_2 \cdots E_{n-1} E_n)$$

By repetitively using Theorem 2.3.7, we have

$$\begin{aligned}\det(A) &= \det(E_1(E_2 \cdots E_{n-1} E_n)) \\ &= \det(E_1) \det(E_2 \cdots E_{n-1} E_n) \\ &= \det(E_1) \det(E_2) \det(\cdots E_{n-1} E_n) \\ &= \det(E_1) \det(E_2) \cdots \det(E_{n-1}) \det(E_n)\end{aligned}$$

Since by Properties 2.3.6, all elementary matrices have a non-zero determinant (particularly we have required  $c \neq 0$  when multiplying a row), i.e.  $\det(E_i) \neq 0$  for all  $i$ , we have  $\det(A) \neq 0$ . We will not go through the details for singular matrices, which are put in the footnote below only for reference.<sup>17</sup> □

Other properties of determinants include:

---

<sup>17</sup>By Theorem 2.2.10, singular matrices have reduced row echelon forms that are not the identity. Observe that all other square rrefs that are not the identity must have at least one row of full zeros, and by Properties 2.3.4 has a determinant of zero.

**Properties 2.3.9.** For any  $n \times n$  square matrices  $A$  and  $B$ , we have

1.  $\det(A^T) = \det(A)$ ,
2.  $\det(kA) = k^n \det(A)$ , for any constant  $k$ ,
3.  $\det(AB) = \det(A) \det(B)$ , and
4.  $\det(A^{-1}) = \frac{1}{\det(A)}$ , if  $A$  is invertible.

By extension,  $\det(A_1 A_2 \cdots A_n) = \det(A_1) \det(A_2) \cdots \det(A_n)$ .

For instance, if

$$A = \begin{bmatrix} 2 & 3 \\ 5 & 9 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 5 \\ 1 & 0 \end{bmatrix}$$

then

$$|A| = (2)(9) - (3)(5) = 3 \quad |B| = (4)(0) - (5)(1) = -5$$

$$\begin{aligned} AB &= \begin{bmatrix} 2 & 3 \\ 5 & 9 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} (2)(4) + (3)(1) & (2)(5) + (3)(0) \\ (5)(4) + (9)(1) & (5)(5) + (9)(0) \end{bmatrix} \\ &= \begin{bmatrix} 11 & 10 \\ 29 & 25 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} |AB| &= (11)(25) - (10)(29) \\ &= -15 = (3)(-5) = |A||B| \end{aligned}$$

So we can see in this case,  $\det(AB) = \det(A) \det(B)$  indeed. We put the formal proof for (3) of Properties 2.3.9 in the footnote for reference.<sup>18</sup>

---

<sup>18</sup>There are two cases to consider,  $A$  being invertible or singular. If  $A$  is singular, then by Properties 2.2.4,  $AB$  is also singular. And by Properties 2.3.8, both  $\det(A)$  and  $\det(AB)$

Short Exercise: Prove (4) of Properties 2.3.9.<sup>19</sup>

### 2.3.3 Finding Inverses by Adjugate

An alternative method to compute the inverse of a matrix is by using its *adjugate*, which is the transpose of its cofactor matrix associated to it.

**Definition 2.3.10** (Adjugate). For a matrix  $A$ , its adjugate is defined as

$$[\text{adj}(A)]_{pq} = (C_{pq})^T = C_{qp}$$

where  $C_{pq}$  is the cofactor of  $A$  at  $(p, q)$ , formulated as in Definition 2.3.2.

**Properties 2.3.11.** The inverse of a matrix  $A$  can be computed from its adjugate

will be zero, and the equality holds trivially. Otherwise, if  $A$  is invertible, then we can follow the idea in the proof of Properties 2.3.8, and let  $A = E_1 E_2 \cdots E_{n-1} E_n$  as a sequence of elementary matrices. By using Theorem 2.3.7 back and forth, we have

$$\begin{aligned} \det(AB) &= \det(E_1 E_2 \cdots E_{n-1} E_n B) \\ &= \det(E_1) \det(E_2) \cdots \det(E_{n-1}) \det(E_n) \det(B) && (\text{Theorem 2.3.7}) \\ &= (\det(E_1) \det(E_2) \cdots \det(E_{n-1}) \det(E_n)) \det(B) \\ &= \det(E_1 E_2 \cdots E_{n-1} E_n) \det(B) && (\text{Theorem 2.3.7}) \\ &= \det(A) \det(B) \end{aligned}$$

So the equality is true in both cases.

<sup>19</sup>Consider  $A^{-1} A = I$ , and take determinant on both sides. By (3), we have

$$\begin{aligned} \det(A^{-1} A) &= \det(I) \\ \det(A^{-1}) \det(A) &= 1 && (\text{The identity always has a determinant of 1}) \\ \det(A^{-1}) &= \frac{1}{\det(A)} \end{aligned}$$

by

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

From this formula, it is obvious that singular matrices, having a determinant of zero, does not have an inverse.

**Example 2.3.5.** For a  $2 \times 2$  matrix

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

It is not difficult to see that the determinant is  $ad - bc$ , and the adjugate matrix is

$$\begin{bmatrix} d & -c \\ -b & a \end{bmatrix}^T = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

So the inverse, if  $ad - bc \neq 0$ , is

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

**Example 2.3.6.** Find the inverse of the following matrix by evaluating its adjugate.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 5 \\ 1 & 4 & 11 \end{bmatrix}$$

*Solution.* First of all, by Sarrus' Rule (Properties 2.3.1)

$$\begin{aligned} |A| &= ((1)(3)(11) + (2)(5)(1) + (3)(1)(4)) \\ &\quad - ((3)(3)(1) + (1)(5)(4) + (2)(1)(11)) \\ &= (33 + 10 + 12) - (9 + 20 + 22) \end{aligned}$$

$$= 4$$

The adjugate matrix is

$$\begin{aligned} \text{adj}(A) &= \begin{bmatrix} \left| \begin{array}{cc} 3 & 5 \\ 4 & 11 \end{array} \right| & -\left| \begin{array}{cc} 1 & 5 \\ 1 & 11 \end{array} \right| & \left| \begin{array}{cc} 1 & 3 \\ 1 & 4 \end{array} \right|^T \\ -\left| \begin{array}{cc} 2 & 3 \\ 4 & 11 \end{array} \right| & \left| \begin{array}{cc} 1 & 3 \\ 1 & 11 \end{array} \right| & -\left| \begin{array}{cc} 1 & 2 \\ 1 & 4 \end{array} \right| \\ \left| \begin{array}{cc} 2 & 3 \\ 3 & 5 \end{array} \right| & -\left| \begin{array}{cc} 1 & 3 \\ 1 & 5 \end{array} \right| & \left| \begin{array}{cc} 1 & 2 \\ 1 & 3 \end{array} \right| \end{bmatrix} \\ &= \begin{bmatrix} 13 & -6 & 1 \\ -10 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix}^T = \begin{bmatrix} 13 & -10 & 1 \\ -6 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix} \end{aligned}$$

(be careful of not forgetting the transpose!) Putting the pieces together according to the formula in Properties 2.3.11, we have

$$\begin{aligned} A^{-1} &= \frac{1}{\det(A)} \text{adj}(A) \\ &= \frac{1}{4} \begin{bmatrix} 13 & -10 & 1 \\ -6 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{13}{4} & -\frac{5}{2} & \frac{1}{4} \\ -\frac{3}{2} & 2 & -\frac{1}{2} \\ \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \end{bmatrix} \end{aligned}$$

□

A summarizing point to be emphasized is that

**Theorem 2.3.12** (Equivalence Statements). For a square matrix  $A$ , the followings are equivalent:

- (a)  $A$  is invertible, i.e.  $A^{-1}$  exists,
- (b)  $\det(A) \neq 0$ ,

(c) The reduced row echelon form of  $A$  is  $I$ .

which is just a rephrasing of Properties 2.3.8 and Theorem 2.2.10. Particularly, invertibility is equivalent to a non-zero determinant. We will see the expansion of these equivalence statements in later chapters.

## 2.4 Python Programming

To create an identity matrix of size  $n$ , we use `np.identity(n)`. For example,

```
import numpy as np
I4 = np.identity(4)
print(I4)
```

returns

```
[[1.  0.  0.  0.]
 [0.  1.  0.  0.]
 [0.  0.  1.  0.]
 [0.  0.  0.  1.]]
```

Applying transpose on a matrix is simple where we just add `.T` after the array variable, like

```
myMatrix1 = np.array([[1.,  0.,  3.],
                      [1.,  4.,  1.],
                      [-1., 2.,  4.]])
print(myMatrix1)
print(myMatrix1.T)
```

yields

```
[[ 1.  0.  3.]
 [ 1.  4.  1.]
 [-1.  2.  4.]]
 [[ 1.  1. -1.]
 [ 0.  4.  2.]
 [ 3.  1.  4.]]
```

Finding the inverse of a matrix requires the `scipy.linalg` library and call the `inv` function.

```
from scipy import linalg
myMatrix2 = linalg.inv(myMatrix1)
print(myMatrix2)
print(myMatrix1@myMatrix2) # Check: should give the identity
```

gives the expected results of

```
[[ 0.4375   0.1875  -0.375   ]
 [-0.15625  0.21875  0.0625  ]
 [ 0.1875   -0.0625  0.125   ]]
[[1.  0.  0.]
 [0.  1.  0.]
 [0.  0.  1.]]
```

Meanwhile, we can use the `det` function to calculate the determinant of a matrix as follows. First,

```
print(linalg.det(myMatrix1))
```

gives the expected output of `32.0`. As another example,

```
myMatrix3 = np.array([[3.,  1.,  3.,  2.],
                      [0., -1., -3.,  1.],
                      [1., -1., -2.,  0.],
                      [2.,  0.,  1.,  0.]])
print(linalg.det(myMatrix3))
```

produces an extremely small value of `1.1102230246251562e-16`. In fact, the matrix

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

has a determinant of exactly zero. It is an artifact of numerical error when using floating point numbers. If we keep going ahead and computes its inverse by `linalg.inv(myMatrix3)`, we will obtain an absurd output of

```
[[ 1.200959e+15 -2.401919e+15  3.602879e+15 -3.602879e+15]
 [ 6.004799e+15 -1.200959e+16  1.801439e+16 -1.801439e+16]
 [-2.401919e+15  4.803839e+15 -7.205759e+15  7.205759e+15]
 [-1.200959e+15  2.401919e+15 -3.602879e+15  3.602879e+15]]
```

that have entries of extremely large magnitude. This phenomenon is due to the extremely small "determinant", through Properties 2.3.11, magnifies the adjugate by being in the denominator. (The actual computation does not use Properties 2.3.11 directly but this is a heuristic perspective to view the problem.) To prevent this, we can add a `if` condition to look for singularity, defining a function like

```
def safe_inv(matrix):
    if np.abs(linalg.det(matrix)) < np.finfo(float).eps:
        print("Warning: The matrix is highly singular!")
        return(np.nan)
    else:
        return(linalg.inv(matrix))
```

where `np.finfo(float).eps` gives the so-called *machine epsilon*  $\epsilon$  (the order of relative round-off error) of `float` and we want the absolute value of the determinant be larger than that. Subsequently, calling `safe_inv(myMatrix3)` will print a warning. Finally, we note that we can use `sympy` to acquire the reduced row echelon form of a matrix. Let's use the matrix in Example 2.2.3 for demonstration.

```
import sympy

myMatrix4 = np.array([[2., 2., 1.],
                     [6., 4., 1.],
                     [2., 3., 2.],
                     [2., 1., 0.]])
myMatrix4_sympy = sympy.Matrix(myMatrix4) # Convert the numpy
                                         array to a sympy matrix
print(myMatrix4_sympy.rref())
```

then returns two objects

```
(Matrix([
[1, 0, -0.5],
[0, 1, 1.0],
[0, 0, 0],
[0, 0, 0]]), (0, 1))
```

The first one is the reduced row echelon form we want, and the second is a tuple which keeps the column indices of the pivots. `sympy` also does *zero testing* such that

```
myMatrix3_sympy = sympy.Matrix(myMatrix3)
print(myMatrix3_sympy**(-1))
```

raises properly the error of

```
NonInvertibleMatrixError("Matrix det == 0; not invertible.")
sympy.matrices.common.NonInvertibleMatrixError: Matrix det
== 0; not invertible.
```

## 2.5 Exercises

**Exercise 2.1** Find the determinant of the matrix below by inspection.

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 & 9 \\ 0 & 0 & 10 & 11 & 12 \\ 0 & 0 & 0 & 13 & 14 \\ 0 & 0 & 0 & 0 & 15 \end{bmatrix}$$

By the same logic, derive a general formula for the determinant of any upper(lower)-triangular matrix.<sup>20</sup>

**Exercise 2.2** Let

$$A = \begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 6 \\ 0 & 1 \end{bmatrix}$$

Verify:

- (a)  $(AB)^T = B^T A^T$ ,
- (b)  $(AB)^{-1} = B^{-1} A^{-1}$ , and
- (c)  $\det(AB) = \det(A) \det(B)$ .

---

<sup>20</sup>An upper(lower)-triangular matrix is a matrix who elements below (above) the main diagonal are all zeros.

for this particular case.

**Exercise 2.3** If

$$A = \begin{bmatrix} 3 & 2 & 9 \\ 1 & 2 & 3 \\ 4 & 0 & 4 \end{bmatrix}$$

Find its inverse by

- (a) Gaussian Elimination, and
- (b) Determinant and adjugate.

**Exercise 2.4** Let

$$A = \begin{bmatrix} 0 & 2 & 5 \\ 0 & 4 & 9 \\ 1 & 2 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 2 & 3 & 4 \\ 2 & 4 & 6 \\ 3 & 5 & 8 \end{bmatrix}$$

Verify:

- (a)  $(AB)^T = B^T A^T$ ,
- (b)  $(AB)^{-1} = B^{-1} A^{-1}$ , and
- (c)  $\det(AB) = \det(A) \det(B)$ .

for this particular case.

**Exercise 2.5** Show that

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

is singular.

**Exercise 2.6** Given

$$A = \begin{bmatrix} 1 & 9 & 1 & 4 \\ 0 & 6 & 2 & 8 \\ 1 & 9 & 3 & 9 \\ 0 & 9 & 0 & 1 \end{bmatrix}$$

Find its determinant, inverse, and determinant of the inverse.

**Exercise 2.7** For the following matrix,

$$A = \begin{bmatrix} p & 1 & 2 \\ 0 & 2 & p \\ 4 & -2 & 0 \end{bmatrix}$$

Find the values of  $p$  such that  $A$  is invertible.

**Exercise 2.8** Show that for any square matrix  $A$ ,  $A + A^T$  is symmetric and  $A - A^T$  is skew-symmetric. Hence show with an explicit formula that any square matrix  $A$  can be written as the sum of a symmetric matrix and a skew-symmetric matrix.

**Exercise 2.9** Prove that if  $A$  is an invertible  $n \times n$  matrix,  $|A| \neq 0$ , then we have

$$\det(\text{adj}(A)) = (\det(A))^{n-1}$$

using Properties 2.3.9 and 2.3.11.

## **Chapter 3**

# **Solutions for Linear Systems**

---

The last chapter has introduced the necessary machinery for solving linear systems of equations and now we are going to see how to apply them under suitable circumstances. Remember, in the first chapter, we have formulated some problems about linear systems of equations appearing in Earth Science, and they will be solved accordingly.

## **3.1 Number of Solutions for Linear Systems**

Before tackling any linear system, we may like to know there are how many solutions. In fact, there are only three possibilities.

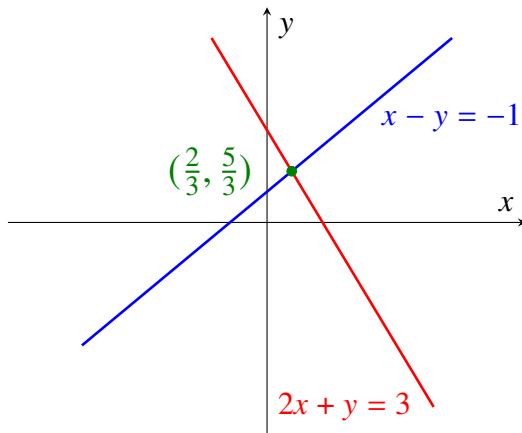
**Theorem 3.1.1** (Number of Solutions for a Linear System). For a system of linear equations  $A\vec{x} = \vec{h}$  (recall Definition 1.2.2 and Properties 1.2.3), it has either:

1. No solution,
2. An unique solution, or
3. Infinitely many solutions.

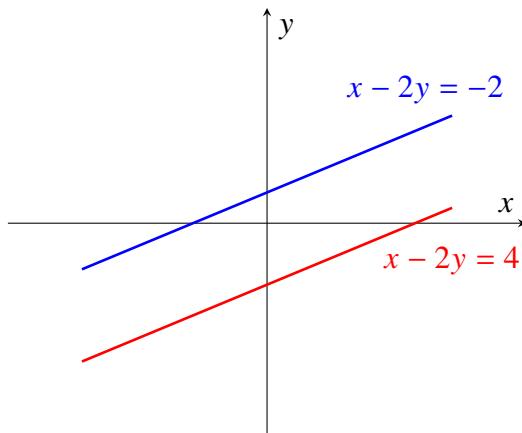
for the unknowns  $\vec{x}$ .

This can be illustrated by considering a linear system with two equations and two unknowns, with each equation representing a line. There are three types of scenarios.

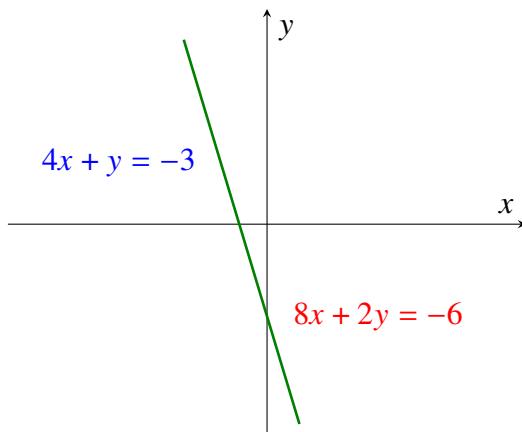
$$\begin{cases} a_1x + b_1y = h_1 \\ a_2x + b_2y = h_2 \end{cases}$$



One solution: Two non-parallel lines (red/blue) intersecting at one point (green).

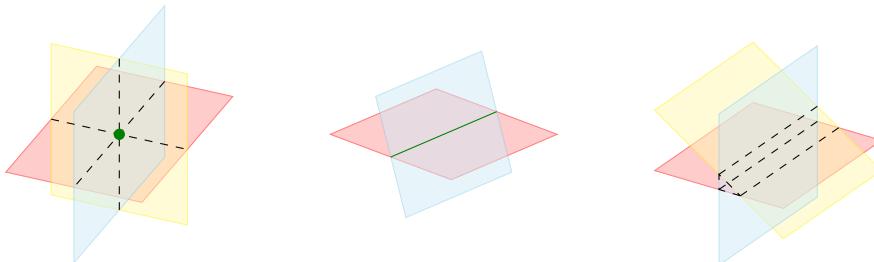


No solution: Two parallel lines never touch each other.



Infinitely many solutions: Two parallel lines overlap each other.

It goes similarly for any linear system of three unknowns in which equations represent planes instead, and the intersection of two non-parallel planes will be a line. We show three possible scenarios below. The readers can try to imagine and drawing out other possibilities.



One solution (left): Three planes (red/yellow/blue) intersecting at one point (green). Infinitely many solutions (middle): Two planes intersecting along a straight line. No solution (right): Three planes intersecting pair-wise along three non-intersecting parallel lines.

In fact, this theorem about the existence of solutions is true for any number of variables and equations.<sup>1</sup> If there is any solution, then the system is called

---

<sup>1</sup>Some readers may think if there can be finitely many solutions only. Unfortunately, it is

**consistent.** Otherwise, if no solution exists, then it is known as **inconsistent**.

Clearly, the next task is about how to find out which case the linear system belongs to. The following theorem reveals the relationship between the number of solutions for a *square* linear system and the determinant of its coefficient matrix.

**Theorem 3.1.2.** For a square linear system  $A\vec{x} = \vec{h}$ , if the coefficient matrix  $A$  is invertible, i.e.  $\det(A) \neq 0$ , then there is always only one unique solution. However, if  $A$  is singular,  $\det(A) = 0$ , then it has either no solution, or infinitely many solutions.

As a consequence, if the homogeneous linear system  $A\vec{x} = \mathbf{0}$  has as singular coefficient matrix with  $\det(A) = 0$ , since it always has a trivial solution of  $\vec{x} = \mathbf{0}$ , the above theorem implies that the homogeneous system must have infinitely many solutions (since it will not have no solution). We defer the arguments for Theorem 3.1.2, as well as the discussion about non-square systems, until we start actually solving linear systems in the next subsection.

Short Exercise: By inspection, determine the number of solutions for the following linear systems.<sup>2</sup>

$$\begin{bmatrix} 2 & 1 & 6 \\ 3 & 0 & 4 \\ 1 & 1 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 4 & 3 \\ 1 & 5 & 2 \\ 1 & 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

## 3.2 Solving Linear Systems

Now it is the time to get down to solving linear systems (preferably written in matrix form), and we have two methods to choose:

---

impossible. Assume there are two distinct solutions  $\vec{x}_1, \vec{x}_2$  to the system  $A\vec{x} = \vec{h}$ , then it is easy to show by construction all  $\vec{x}_t = t\vec{x}_1 + (1 - t)\vec{x}_2$  for any  $t$  will be valid solutions which are infinitely many.

<sup>2</sup>These two homogeneous linear system has a determinant of  $-1$  and  $0$ , and hence by Theorem 3.1.2 the first system has a unique solution and the second one has infinitely many solutions.

1. By Gaussian Elimination, for linear system in any shape, or
2. By Inverse, which is apparently only applicable for square, invertible coefficient matrices.

### 3.2.1 Solving Linear Systems by Gaussian Elimination

Like in Section 2.2.3, applying Gaussian Elimination on the augmented matrix (introduced at the end of Section 1.2) of a linear system can yield the solution to the right. The principle involving elementary row operations is the same as that in Properties 2.2.9 and Theorem 2.2.10, but with  $A\vec{x} = \vec{h}$  instead of  $AA^{-1} = I$ . Let  $A_{\text{rref}}$  be the reduced row echelon form of  $A$ , and  $E_1, E_2, \dots, E_n$  be the elementary matrices used in the Gaussian Elimination process to arrive at the rref. For any solution  $\vec{x}$  to the system  $A\vec{x} = \vec{h}$ , we multiply the elementary matrices one by one to the left on both sides of the equation, leading to

$$(E_n \cdots E_2 E_1) A \vec{x} = (E_n \cdots E_2 E_1) \vec{h}$$

$$(E_n \cdots E_2 E_1 A) \vec{x} = A_{\text{rref}} \vec{x} = (E_n \cdots E_2 E_1) \vec{h}$$

hence  $\vec{x}$  will be the solution to  $A_{\text{rref}} \vec{x} = \vec{h}$  at the same time where  $\vec{h} = E_n \cdots E_2 E_1 \vec{h}$ . Therefore, the solutions of  $A\vec{x} = \vec{h}$  and  $A_{\text{rref}} \vec{x} = \vec{h}$  coincide, which can be inferred more directly from the latter system. In addition, the coefficient matrix  $A$  can be non-square, but we will look at the easier case of a square coefficient matrix first.

#### Square Systems

**Example 3.2.1.** Solve the following linear system by Gaussian Elimination.

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 4 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 10 \\ 8 \end{bmatrix}$$

*Solution.* We rewrite the system in augmented form and then apply Gaussian Elimination, with the aim to reduce the coefficient matrix on the left to the identity matrix.

$$\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 1 & 1 & 4 & 10 \\ 2 & 0 & 3 & 8 \end{array} \rightarrow \begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 3 & 7 \\ 0 & 0 & 1 & 2 \end{array} \quad \begin{array}{l} R_2 - R_1 \rightarrow R_2 \\ R_3 - 2R_1 \rightarrow R_3 \end{array}$$

$$\rightarrow \begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{array} \quad \begin{array}{l} R_2 - 2R_3 \rightarrow R_2 \\ R_1 - R_3 \rightarrow R_1 \end{array}$$

which translates to

$$\begin{cases} x = 1 \\ y = 1 \\ z = 2 \end{cases} \quad \text{or} \quad \vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

Note that we have successfully converted the coefficient matrix to the identity along the way, which by Theorem 2.3.12 (c) to (a) implies that the coefficient matrix is invertible. This explains the first part of Theorem 3.1.2 as every unknown is now associated to a single leading 1 in the corresponding column of the identity matrix acquired from the reduction process and a unique solution can be derived.  $\square$

**Example 3.2.2.** Solve the linear system

$$\begin{bmatrix} 3 & 7 & 2 \\ 1 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 8 \\ 2 \\ 2 \end{bmatrix}$$

if possible.

*Solution.* Again, we apply Gaussian Elimination on the augmented matrix to obtain

$$\begin{array}{ccc|c} 3 & 7 & 2 & 8 \\ 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & 2 \end{array} \rightarrow \begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 3 & 7 & 2 & 8 \\ 0 & 2 & 1 & 2 \end{array} \quad R_1 \leftrightarrow R_2$$

$$\begin{aligned}
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 4 & 2 & 2 \\ 0 & 2 & 1 & 2 \end{array} \right] \quad R_2 - 3R_1 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 2 & 1 & 2 \end{array} \right] \quad \frac{1}{4}R_2 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{array} \right] \quad R_3 - 2R_2 \rightarrow R_3
 \end{aligned}$$

The last row corresponds to  $0 = 1$  which is contradictory. As a consequence, the system is inconsistent, i.e. no solution exists.  $\square$

**Example 3.2.3.** Find all solutions for the following linear system.

$$\left[ \begin{array}{ccc} 1 & 2 & 1 \\ 2 & 5 & 3 \\ 0 & 1 & 1 \end{array} \right] \left[ \begin{array}{c} x \\ y \\ z \end{array} \right] = \left[ \begin{array}{c} 1 \\ 2 \\ 0 \end{array} \right]$$

*Solution.* Gaussian Elimination leads to

$$\begin{aligned}
 \left[ \begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 2 & 5 & 3 & 2 \\ 0 & 1 & 1 & 0 \end{array} \right] & \rightarrow \left[ \begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{array} \right] \quad R_2 - 2R_1 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 - R_2 \rightarrow R_3 \quad R_1 - 2R_2 \rightarrow R_1
 \end{aligned}$$

Now, the last row corresponds to  $0 = 0$ , which is vacuous and implies that one equation is spurious. This also means we can assign an unknown as a **free variable (parameter)** for expressing other variables. We will choose unknown(s) that is/are not linked to any pivot in the reduced coefficient matrix. As the variables  $x$  and  $y$  already correspond to the two pivots in the first/second column,

we can let  $z = t$  where  $t$  represents a free parameter. Then the first/second row gives  $x = 1 + t$ ,  $y = -t$  respectively, and

$$\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1+t \\ -t \\ t \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

with  $-\infty < t < \infty$ . The first column vector appearing alone

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

is the so-called **particular solution**, which can be any vector  $\vec{x} = \vec{x}_p$  that satisfies the inhomogeneous part of the system  $A\vec{x} = \vec{h}$ . Meanwhile, the second column vector multiplied by the free parameter  $t$

$$t \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

is known as the **complementary solution**, the family of all vectors  $\vec{x} = \vec{x}_c$  that satisfy the homogeneous part  $A\vec{x} = \mathbf{0}$ . Combined together, they form the **general solution**  $\vec{x}_g = \vec{x}_p + \vec{x}_c$  as the complete set of solutions to the linear system.  $\square$

Short Exercise: Try plugging in any number  $t$  to the general solution and verify the consistency.<sup>3</sup>

---

<sup>3</sup>Let's say  $t = 1$  and  $\tilde{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (1) \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ , then clearly  $A\tilde{x} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 3 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$ .

$\tilde{x}$  can become a new particular solution by noting that the original solution form can be rewritten as

$$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (1) \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + (t-1) \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} + t' \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \tilde{x} + t' \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

where we "extract"  $\tilde{x}$  from shifting the free parameter via  $t' = t - 1$ , and it represents the same general solution as the original expression.

The complementary solution encompasses all possible solutions to the homogeneous part  $A\vec{x} = \mathbf{0}$  of the linear system  $A\vec{x} = \vec{h}$ . For broader situations, it can contain more than one pairs of free parameter and column vector (or none, if the homogeneous part only permits the trivial solution of all zeros), and the complementary solution becomes a *linear combination* of multiple *linearly independent* column vectors who satisfy  $A\vec{x} = \mathbf{0}$  on their own. (We will clarify about these concepts in Chapter 6.) The amount of free variables is decided by the number of columns in the coefficient matrix (unknowns), minus the number of pivots (constraints) in its reduced row echelon form. This quantity is called *nullity* and in the last example it equals to 1. In case of multiple free variables, we assign the corresponding number of free parameters to non-pivotal unknowns and apply the same procedure as in the example above to acquire a set of complementary solution. Any column vector that constitutes the complementary solution (followed by a free parameter) can be scaled by any non-zero factor as we desire.<sup>4</sup>

Meanwhile, the particular solution can be set to any valid solution to the linear system (the choice does not affect the structure of its complementary part, see the footnote to the last short exercise). If the linear system is itself homogeneous, then the zero vector  $\mathbf{0}$  can always be chosen as a particular solution which does not appear explicitly.

We have seen in the previous two examples that if the reduced row echelon form of the square coefficient matrix has some row of full zeros, then

---

<sup>4</sup>Using the last example as a demonstration,

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \frac{t}{2} \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$$

where we use  $s = \frac{t}{2}$  as a new free parameter. Notice that the old column vector  $\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$  in the original expression of complementary solution, and the newly generated column vector  $\begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$  that is double in length, both satisfy the homogeneous part  $A\vec{x} = 0$ .

it either leads to no solution (if inconsistent) or infinitely many solutions (if consistent). Since such a matrix at the same time has a determinant of zero (by Properties 2.3.4) and is singular, this establishes the second part of Theorem 3.1.2.

For non-square coefficient matrices, two cases occur.

1. There are more equations (rows) than unknowns (columns). The system is **overdetermined**. The rref of coefficient matrix then must have at least one row of full zeros. If any one of them is inconsistent, then contradiction will arise just like in Example 3.2.2 and there will be no solution. However, if all zero rows are consistent (i.e.  $0 = 0$ ), then there still can be a unique solution or infinitely many of them.
2. There are fewer equations (rows) than unknowns (columns). The system is said to be **underdetermined**. There must be unknown(s) that is/are non-pivot(s) in the reduced row echelon form of the coefficient matrix. Hence free variables, and infinitely many solutions ensue if there is no inconsistent row (if there is at least one row of  $0 = k$  where  $k$  is a non-zero constant, then there is no solution). The calculation is similar to that in Example 3.2.3.

Let's see some examples for non-square linear systems.

### Overdetermined Systems

**Example 3.2.4.** Find the solution to the following overdetermined system, if any.

$$\begin{bmatrix} 1 & 4 & 0 \\ 2 & 2 & 3 \\ 1 & 1 & 2 \\ 0 & 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ 3 \\ 5 \end{bmatrix}$$

*Solution.*

$$\begin{array}{c}
 \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 2 & 2 & 3 & 8 \\ 1 & 1 & 2 & 3 \\ 0 & 3 & 1 & 5 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & -6 & 3 & 0 \\ 0 & -3 & 2 & -1 \\ 0 & 3 & 1 & 5 \end{array} \right] \quad R_2 - 2R_1 \rightarrow R_2 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & -3 & 2 & -1 \\ 0 & 3 & 1 & 5 \end{array} \right] \quad R_3 - R_1 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & \frac{5}{2} & 5 \end{array} \right] \quad -\frac{1}{6}R_2 \rightarrow R_2 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & \frac{5}{2} & 5 \end{array} \right] \quad R_3 + 3R_2 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & \frac{5}{2} & 5 \end{array} \right] \quad R_4 - 3R_2 \rightarrow R_4 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 10 \end{array} \right] \quad 2R_3 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 10 \end{array} \right] \quad R_4 - \frac{5}{2}R_3 \rightarrow R_4
 \end{array}$$

The last row is inconsistent and hence the overdetermined system has no solution.  $\square$

**Example 3.2.5.** Show that there are infinitely many solution to the following overdetermined system.

$$\left[ \begin{array}{ccc} 1 & 1 & 2 \\ 1 & 2 & 5 \\ 2 & 1 & 1 \\ 1 & 0 & -1 \end{array} \right] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 3 \\ 1 \end{bmatrix}$$

*Solution.*

$$\begin{array}{c}
 \left[ \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 1 & 2 & 5 & 3 \\ 2 & 1 & 1 & 3 \\ 1 & 0 & -1 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & 3 & 1 \\ 0 & -1 & -3 & -1 \\ 0 & -1 & -3 & -1 \end{array} \right] \quad R_2 - R_1 \rightarrow R_2 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 - 2R_1 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_4 - R_1 \rightarrow R_4 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_4 + R_2 \rightarrow R_4 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_1 - R_2 \rightarrow R_1
 \end{array}$$

The two rows of full zeros indicate that two out of the four equations are redundant and there are effectively two constraints only, over the three variables. We can let the non-pivotal unknown  $z = t$  be a free variable like in Example 3.2.3, and obtain  $x = 1 + t$ ,  $y = 1 - 3t$  from the first two rows. Thus the general solution is

$$\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1+t \\ 1-3t \\ t \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix}$$

where  $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$  is a particular solution and the nullity is 1. □

## Underdetermined Systems

**Example 3.2.6.** Solve the following underdetermined system.

$$\left[ \begin{array}{cccc} 1 & 1 & 2 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 0 & 3 & 2 \end{array} \right] \left[ \begin{array}{c} x \\ y \\ u \\ v \end{array} \right] = \left[ \begin{array}{c} 0 \\ 1 \\ -1 \end{array} \right]$$

*Solution.*

$$\begin{aligned} \left[ \begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 3 & 2 & -1 \end{array} \right] &\rightarrow \left[ \begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & -1 & 1 & 1 & -1 \end{array} \right] && R_2 - R_1 \rightarrow R_2 \\ &\rightarrow \left[ \begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] && R_3 - R_1 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cccc|c} 1 & 0 & 3 & 2 & -1 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] && R_3 + R_2 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cccc|c} 1 & 0 & 3 & 2 & -1 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] && R_1 - R_2 \rightarrow R_1 \end{aligned}$$

The last zero row is consistent. The first two columns are pivotal and we can let the remaining two unknowns that are not associated to any leading 1,  $u = s$  and  $v = t$  be free variables. From the first two equations, we retrieve  $x = -1 - 3s - 2t$  and  $y = 1 + s + t$ , and therefore the general solution is

$$\vec{x} = \begin{bmatrix} x \\ y \\ u \\ v \end{bmatrix} = \begin{bmatrix} -1 - 3s - 2t \\ 1 + s + t \\ s \\ t \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} -3 \\ 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -2 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

with  $\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$  as a particular solution and the nullity being 2. □

### 3.2.2 Solving Linear Systems by Inverse

For a square linear system  $A\vec{x} = \vec{h}$ , if  $A$  has a non-zero determinant and is invertible, then we can apply its inverse to recover the solution. Remember that multiplying a matrix by its inverse returns an identity matrix, hence it is possible to multiply the inverse  $A^{-1}$  (or heuristically, "dividing" by  $A$ ) to the left on both sides of the equation  $A\vec{x} = \vec{h}$  to cancel out the  $A$  on L.H.S., which reads

$$\begin{aligned} A^{-1}A\vec{x} &= (A^{-1}A)\vec{x} = A^{-1}\vec{h} \\ \vec{x} &= I\vec{x} = A^{-1}\vec{h} \quad (\text{Definition 2.2.1 and Properties 2.1.2}) \end{aligned} \quad (3.1)$$

This solution is unique, guaranteed by Theorem 3.1.2.

**Example 3.2.7.** Solve the linear system  $A\vec{x} = \vec{h}$  by the inverse method, where

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

*Solution.* It can be checked that the inverse of the coefficient matrix is

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

The readers are encouraged to derive the inverse by themselves. Subsequently, we have the solution to the linear system as

$$\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = A^{-1}\vec{h} = \begin{bmatrix} -\frac{3}{2} & -\frac{1}{2} & \frac{5}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{3}{2} & -\frac{1}{2} & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

□

Doing Gaussian Elimination to find the inverse and then compute the solution by  $\vec{x} = A^{-1}\vec{h}$  (Equation (3.1)) is, at first sight somehow the same as using Gaussian Elimination directly to solve the linear system as sketched in Section 3.2.1. However, in computer, calculation of inverse can be unstable (see Section 2.4) and there are some other practical reasons not to take the first approach, which will be discussed in Section 3.4.

Besides, Theorem 2.3.12 can be extended as below by incorporating Theorem 3.1.2:

**Theorem 3.2.1** (Equivalence Statement, ver. 2). For a square matrix  $A$ , the followings are equivalent:

- (a)  $A$  is invertible, i.e.  $A^{-1}$  exists,
- (b)  $\det(A) \neq 0$ ,
- (c) The reduced row echelon form of  $A$  is  $I$ ,
- (d) The linear system  $A\vec{x} = \vec{h}$  has a unique solution for any  $\vec{h}$ , particularly  $A\vec{x} = \mathbf{0}$  has only the trivial solution  $\vec{x} = \mathbf{0}$ .

### Cramer's Rule

## 3.3 Earth Science Applications

Now we are going to revisit and find the solutions to the two linear system problems in Section 1.4.

**Example 3.3.1.** Solve for the horizontal displacement  $x$  and depth of top layer  $y$  in the seismic ray problem of Example 1.4.1.

*Solution.* The linear system is

$$\begin{bmatrix} 1 & 1 \\ 1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 120 \\ 80\sqrt{3} \end{bmatrix}$$

Since it is just a  $2 \times 2$  coefficient matrix, we can directly use the expression in Example 2.3.5 to find its inverse, which is

$$\frac{1}{\sqrt{3}-1} \begin{bmatrix} \sqrt{3} & -1 \\ -1 & 1 \end{bmatrix} = \frac{1+\sqrt{3}}{2} \begin{bmatrix} \sqrt{3} & -1 \\ -1 & 1 \end{bmatrix}$$

and solve the system by multiplying the inverse following the method demonstrated in Section 3.2.2, leading to

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1+\sqrt{3}}{2} \begin{bmatrix} \sqrt{3} & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 120 \\ 80\sqrt{3} \end{bmatrix} = \begin{bmatrix} 60 + 20\sqrt{3} \\ 60 - 20\sqrt{3} \end{bmatrix}$$

Therefore the required horizontal displacement and depth of top layer are about 94.6 m and 25.4 m respectively.  $\square$

**Example 3.3.2.** Find the radiative loss  $E_j$  and hence temperature  $T_j$  in each layer of the multi-layer model in Example 1.4.2. In particular, what is the temperature at the surface ( $j = N + 1$ )?

*Solution.* The linear system is

$$\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 \\ \vdots & & & \ddots & & \vdots & \\ 0 & 0 & 0 & & -2 & 1 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{N-1} \\ E_N \\ E_{N+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ -E_{in} \end{bmatrix}$$

where  $N$  is any positive integer. Since  $N$  can be arbitrarily large, we may wish to avoid the direct computation of a massive inverse. Instead, we resort to a

tactful way of row reduction to reveal the pattern of  $R_j$ . Rather than starting the reduction at the top as usual, we build up at the bottom, subtracting the lower row from the row directly above it and then moving up a row, repeated until we reach the top.

$$\begin{array}{l}
 \left[ \begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & \vdots & & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array} \right] \\
 \rightarrow \left[ \begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & \vdots & & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array} \right] \quad R_N + R_{N+1} \rightarrow R_N \\
 \rightarrow \left[ \begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & \vdots & & \vdots \\ 0 & 0 & 0 & & -1 & 0 & 0 & -E_{in} \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array} \right] \quad R_{N-1} + R_N \rightarrow R_{N-1} \\
 \rightarrow \quad : \quad (\text{Keep going up})
 \end{array}$$

$$\begin{aligned}
 & \rightarrow \left[ \begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & & 0 & 0 & 0 & -E_{in} \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 0 & 0 & -E_{in} \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -R_{in} \end{array} \right] \\
 & \rightarrow \left[ \begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & & 0 & 0 & 0 & -E_{in} \\ 0 & 1 & -1 & & 0 & 0 & 0 & -E_{in} \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 0 & 0 & -E_{in} \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array} \right] \quad R_2 + R_3 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccccccc|c} -1 & 0 & 0 & \cdots & 0 & 0 & 0 & -E_{in} \\ 1 & -1 & 0 & & 0 & 0 & 0 & -E_{in} \\ 0 & 1 & -1 & & 0 & 0 & 0 & -E_{in} \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 0 & 0 & -E_{in} \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array} \right] \quad R_1 + R_2 \rightarrow R_1
 \end{aligned}$$

From the first row, we readily obtain  $E_1 = E_{in}$ . The second row yields the equation

$$\begin{aligned}
 E_1 - E_2 &= -E_{in} \\
 E_2 &= E_1 + E_{in} = E_{in} + E_{in} = 2E_{in}
 \end{aligned}$$

Similarly, the subsequent rows are all in the form of  $E_j = E_{j-1} + E_{in}$ , and inductively we have  $E_j = jE_{in}$ .  $E_1 = E_{in}$  is the emission of radiation from the Earth as a whole as viewed from the space, and the corresponding *emission temperature* is  $T_e = T_1 = \sqrt[4]{E_1/\sigma} = \sqrt[4]{E_{in}/\sigma}$  by Stefan–Boltzmann Law. The surface releases terrestrial radiation at the rate of  $E_{N+1} = (N+1)E_{in}$  and has a temperature of  $T_{N+1} = \sqrt[4]{E_{N+1}/\sigma} = \sqrt[4]{(N+1)E_{in}/\sigma} = (N+1)^{1/4} \sqrt[4]{E_{in}/\sigma} =$

$(N + 1)^{1/4}T_e$ , i.e. the surface temperature is  $(N + 1)^{1/4}$  times the emission temperature. Our earth has an emission temperature of 255 K and a surface temperature of 288 K on average (notice that we have to use Kelvin instead of degree Celsius!), which leads to an effective number of absorbing layers  $N = (288/255)^4 - 1 = 0.627$ .  $\square$

**Example 3.3.3.** Show that the overdetermined system about air temperature measurements in Example 1.4.4 has no solution.

*Solution.* The overdetermined system is

$$\begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \begin{bmatrix} \frac{\partial T}{\partial x} \\ \frac{\partial T}{\partial y} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix}$$

or in augmented form,

$$\left[ \begin{array}{cc|c} 10 & 20 & 0.2 \\ 25 & 15 & 0.3 \\ -10 & 5 & -0.2 \end{array} \right]$$

Applying Gaussian Elimination, we have

$$\begin{aligned} \left[ \begin{array}{cc|c} 10 & 20 & 0.2 \\ 25 & 15 & 0.3 \\ -10 & 5 & -0.2 \end{array} \right] &\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 25 & 15 & 0.3 \\ -10 & 5 & -0.2 \end{array} \right] && \frac{1}{10}R_1 \rightarrow R_1 \\ &\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & -35 & -0.2 \\ 0 & 25 & 0 \end{array} \right] && R_2 - 25R_1 \rightarrow R_2 \\ &\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 25 & 0 \\ 0 & -35 & -0.2 \end{array} \right] && R_3 + 10R_1 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 1 & 0 \\ 0 & -35 & -0.2 \end{array} \right] && R_2 \leftrightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 1 & 0 \\ 0 & -35 & -0.2 \end{array} \right] && \frac{1}{25}R_2 \rightarrow R_2 \end{aligned}$$

$$\rightarrow \left[ \begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 1 & 0 \\ 0 & 0 & -0.2 \end{array} \right] \quad R_3 + 35R_2 \rightarrow R_3$$

so the last row is inconsistent and the linear system has no solution. There are two possibilities for this scenario. First, the starting assumption that the temperature gradients are linear is obviously an idealized one and there can be deviations in reality. Moreover, even if the physical temperature gradients  $\partial T / \partial x$  and  $\partial T / \partial y$  are truly constant, there may be measurement errors or limitation in precision for the thermometers. Therefore, it is almost impossible to obtain an exact solution in this type of problems but often a "good enough" approximation is still useful. We will learn more about how to compute such an optimized solution in Chapter 13.  $\square$

## 3.4 Python Programming

For solving square linear systems in the form of  $A\vec{x} = \vec{h}$ , we can again import the `scipy.linalg` library and call the `solve` function with the coefficient matrix  $A$  as the first argument and  $\vec{h}$  placed in the second one.

```
import numpy as np
from scipy import linalg

A = np.array([[1., 0., 1.],
              [2., 2., 3.],
              [1., 2., 0.]])
h = np.array([0., -1., 1.])
x = linalg.solve(A, h)
```

This corresponds to the linear system

$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 2 & 3 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

which has a solution of

$$\vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

`print(x)` then gives the correct output of [ 1. -0. -1.]. However, if  $A$  is a singular matrix like the one shown in Section 2.4

```
A = np.array([[3., 1., 3., 2.],
             [0., -1., -3., 1.],
             [1., -1., -2., 0.],
             [2., 0., 1., 0.]]) # "myMatrix3" in the last
                      chapter
h = np.array([0., 1., 1., -1.])
x = linalg.solve(A,h)
print(x)
```

raises a warning and an unreasonable output of

```
LinAlgWarning: Ill-conditioned matrix
(rcond=3.42661e-18): result may not be accurate.
x = linalg.solve(A,h)
[ 4.803839e+15  2.401919e+16 -9.607679e+15 -4.803839e+15]
```

Again, we can use the `sympy` package for the rescue as follows.

```
import sympy

A_sympy = sympy.Matrix(A)
h_sympy = sympy.Matrix(h)
A_sympy.solve(h_sympy)
```

which raises the same "not invertible" error as in Section 2.4. We note that, unfortunately, **there is no simple way to deal with over/under-determined systems using either `scipy` or `sympy`** (Sorry this is wrong, we can use `solveset.linsolve`, WIP). Moreover, there are two questions that may come to the curious readers when reading the programming sections of these two chapters. First, which of `scipy` and `sympy` should we choose over another? Second, why we don't compute the inverse of  $A$  and solve the system by something along the line of `x = linalg.inv(A) @ h`? For the first question, we note that `scipy` is numerical while `sympy` is symbolic, which means that if we are dealing with real

data we may find `scipy` adequate and more efficient, while if we are focusing on the theoretical part of Mathematics we can obtain a more analytical solution with `sympy`. To the second question, we refer the readers to [this excellent Stack Overflow post](#) (31256252).

## 3.5 Exercises

**Exercise 3.1** Solve the following linear system.

$$\begin{cases} 5x + y + 3z = 6 \\ 2x - y + z = \frac{7}{2} \\ 3x + 2y - 4z = -\frac{13}{2} \end{cases}$$

**Exercise 3.2** Solve  $A\vec{x} = \vec{h}_k$ , where

$$A = \begin{bmatrix} 6 & 7 & 7 \\ 1 & 0 & 2 \\ 2 & 1 & 1 \end{bmatrix} \quad \vec{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\vec{h}_1 = \begin{bmatrix} -1 \\ 5 \\ 1 \end{bmatrix} \quad \vec{h}_2 = \begin{bmatrix} 19/4 \\ 1 \\ 5/4 \end{bmatrix}$$

**Exercise 3.3** Derive the solution to the following linear system.

$$\begin{cases} 3x + 4z = 2 \\ x + y + 2z = -1 \\ x - 2y = 0 \end{cases}$$

**Exercise 3.4** Solve the following linear system.

$$\begin{cases} m + n - p - 3q &= 2 \\ m - q &= 5 \\ 3m + 2n - 2p - 7q &= 9 \end{cases}$$

How about if the R.H.S. of the third equation is equal to 3 instead?

**Exercise 3.5** For the following linear system,

$$\begin{bmatrix} 1 & 0 & \alpha \\ 0 & \alpha & 0 \\ \alpha & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \alpha \end{bmatrix}$$

Find the values of  $\alpha$  so that the system has no solution, or infinitely many solutions.

**Exercise 3.6** *Ohm's law* relates voltage drop of a current due to resistance by  $V = IR$ . In addition, *Kirchhoff's Second Law* states that: The voltage gain balances the voltage drop around any closed loop (net voltage change must be zero). The clockwise convention is adopted, i.e. around a loop, a battery with its positive terminal facing the clockwise direction is considered a voltage gain, and clockwise current passing through a resistor is deemed as a voltage drop. Together with the knowledge that current at a junction must conserve (*Kirchhoff's First Law*), find  $I_1, I_2, I_3$  (assumed flowing in the direction as indicated) for the circuit in Figure 3.1.

You will obtain two equations by considering any two loops with Kirchhoff's Second Law, and one from Kirchhoff's First Law. So, there are three equations, for the three unknown currents.

**Exercise 3.7** The *shallow water equations* (see Figure 3.2) describe the evolution of gravity wave under some approximations such as *hydrostatic balance* and a

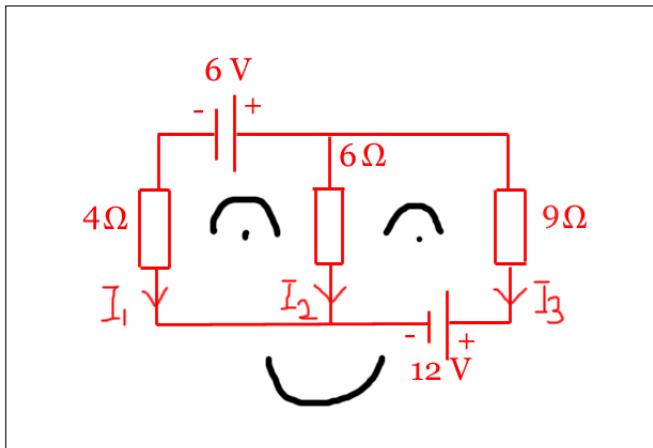


Figure 3.1: The circuit for Exercise 3.6

sufficiently shallow fluid depth, and has the form of

$$\begin{cases} \frac{\partial \eta}{\partial t} + H\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = 0 \\ \frac{\partial u}{\partial t} = -g\frac{\partial \eta}{\partial x} \\ \frac{\partial v}{\partial t} = -g\frac{\partial \eta}{\partial y} \end{cases}$$

when the Coriolis effect is ignored. By assuming a travelling wave solution

$$\begin{aligned} u &= \tilde{U} \cos(kx + ly - \omega t) \\ v &= \tilde{V} \cos(kx + ly - \omega t) \\ \eta &= \tilde{\eta} \cos(kx + ly - \omega t) \end{aligned}$$

where  $\tilde{U}$ ,  $\tilde{V}$ ,  $\tilde{\eta}$  are some constants to be determined, show that the equations become

$$\begin{cases} \omega\tilde{\eta} - kH\tilde{U} - lH\tilde{V} = 0 \\ \omega\tilde{U} - kg\tilde{\eta} = 0 \\ \omega\tilde{V} - lg\tilde{\eta} = 0 \end{cases}$$

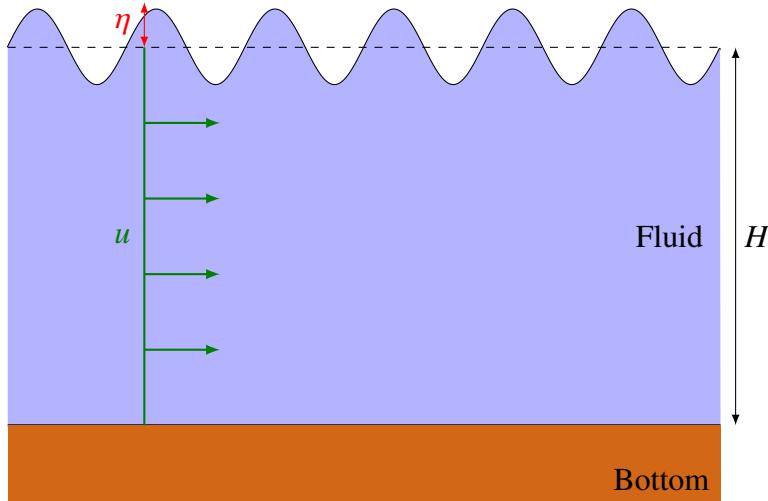


Figure 3.2: The  $x$ - $z$  cross-section of shallow water system in Exercise 3.7.  $\eta$  is the height of free surface,  $H$  is the mean depth of the fluid, and  $u$  is the fluid velocity along  $x$ -axis.

By requiring that  $\tilde{U}$ ,  $\tilde{V}$ ,  $\tilde{\eta}$  have a non-trivial solution so that they are not all zeros, derive the dispersion relation of gravity wave, which is

$$\begin{aligned}\omega^2 &= gH(k^2 + l^2) \\ \omega &= c\kappa\end{aligned}$$

where  $c = \sqrt{gH}$  is the wave speed, and  $\kappa = \sqrt{k^2 + l^2}$  is the total wavenumber.

**Exercise 3.8** Solve for the condensation height and temperature  $z_{cd}$  and  $T_{cd}$  in Exercise 1.8.

**Exercise 3.9** Solve the *Chickens and Rabbits in the Same Cage* problem in Exercise 1.9. If we now introduce a new type of mystical creature who has one head and three legs, and throw them in another cage along with some chickens and rabbits, find all possible numbers of the three species if the cage now has 48 heads and 122 legs.



## Chapter 4

# Introduction to Vectors

---

After three chapters of discussion about matrices, it is time to talk about another closely related object type in linear algebra, namely, vectors. While *vectors* and *vector spaces* have strictly mathematical definitions which make them abstract, we will take a more physical point of view with the special case of (finite-dimensional) geometric vectors first.

## 4.1 Definition and Operations of Geometric Vectors

### 4.1.1 Basic Structure of Vectors in the Real $n$ -space $\mathbb{R}^n$

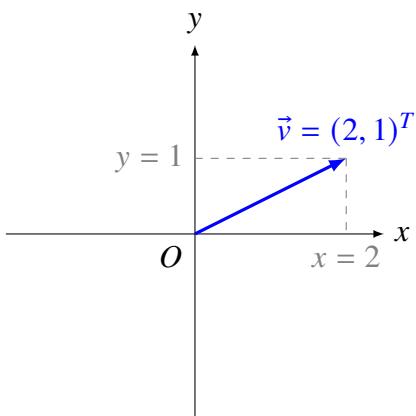
A (*geometric*) **vector** is a physical quantity represented by an ordered tuple of *components* (numbers), e.g.  $(1, 8, 7, 4)$ ,  $(1 - \iota, 1 + 3\iota, 2)$ . It has a *magnitude* (*length*) and *direction*, resembling an arrow. Some real-life examples are: two-dimensional flow velocity  $(u, v)$ , relative position of an airplane to a ground radar  $(x, y, z)$ .

**Definition 4.1.1** ( $n$ -dimensional Geometric Vector). A  $n$ -dimensional geometric vector consists of  $n$  ordered numbers called **components** and are denoted by either an arrow or boldface, like  $\vec{v}$  or  $\mathbf{v}$ . It is usually written out in two forms, as

a column vector or an ordered  $n$ -tuple:

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} = (v_1, v_2, v_3, \dots, v_n)^T$$

A  $n$ -dimensional vector can be treated as an  $n \times 1$  (**column vector**) as suggested above, or a  $1 \times n$  matrix (**row vector**) depending on the situation. The form of a column vector is taken more often than the row vector one and so the column form is assumed throughout the book if it is not further specified. That is why the superscript  $T$  is added for the  $n$ -tuple form to reflect that it is in fact a column vector despite written horizontally.



A 2D vector drawn in an x-y plane.

Movement 移動速度和方向	1-min Average Strength 一分鐘平均強度		Distance/Bearing from HK 與香港的距離和方位角
WNW 西北偏西 (288°) 18 km/h	70 kt (130 km/h)	TY (Cat. 1) 一級颱風	SSE 東南偏南 116 km
WNW 西北偏西 (289°) 20 km/h	70 kt (130 km/h)	TY (Cat. 1) 一級颱風	WSW 西南偏西 178 km

Forecast for *Typhoon Higos* (taken from [Hong Kong Weather Watch](#)). Its horizontal movement is a two-dimensional vector, even though the speed and direction are given instead of the velocities in  $x$  and  $y$ -direction (they can be

converted to each other).

Implicit in the definition of  $n$ -dimensional vectors is the  $n$ -dimensional *space* they are residing in. Assume the components of those vectors are all real, then the set of all such vectors constitutes the ***real n-space***  $\mathbb{R}^n$ .

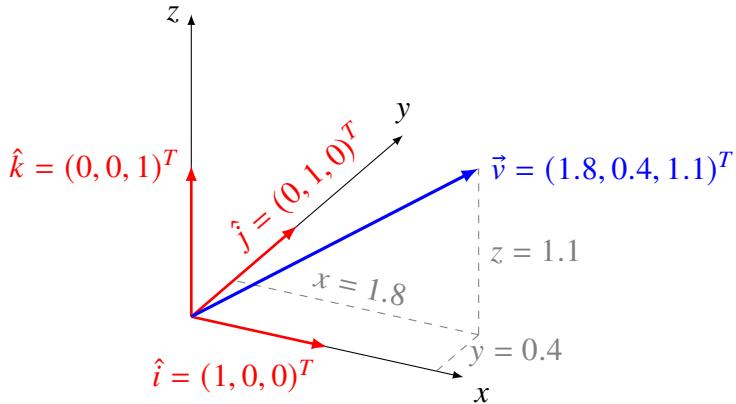
**Definition 4.1.2** (The Real  $n$ -space  $\mathbb{R}^n$ ). The real  $n$ -space  $\mathbb{R}^n$  is defined as the set of all possible  $n$ -tuples  $\vec{v} = (v_1, v_2, v_3, \dots, v_n)^T$  as defined in Definition 4.1.1, where  $v_i$  can take any *real* value, for  $i = 1, 2, 3, \dots, n$ . Such objects in  $\mathbb{R}^n$  are known as  $n$ -dimensional *real* vectors.

While we have not clearly defined what a vector space is, we note that  $\mathbb{R}^n$  fulfills the requirements of a vector space in a mathematical sense. A more detailed discussion of this aspect will be deferred to Chapter 6. Meanwhile, the complex counterpart will be explored in Chapter 8.

An  $n$ -dimensional real geometric vectors as described in Definition 4.1.1 and 4.1.2 can be written as the sum of  $n$  ***standard unit vectors*** that have a magnitude of 1 and are oriented in the positive direction along the  $p$ -th coordinates axes. They are denoted by  $\hat{e}_p$ , where  $p$  can be from 1 to  $n$ . The coordinate axes are perpendicular (or more generally, *orthogonal*, introduced later in this chapter) to each other and this coordinate system is known as the ***Cartesian (coordinate) system***. Particularly, in the three-dimensional real space  $\mathbb{R}^3$ ,  $\hat{e}_1 = \hat{i} = (1, 0, 0)^T$ ,  $\hat{e}_2 = \hat{j} = (0, 1, 0)^T$ ,  $\hat{e}_3 = \hat{k} = (0, 0, 1)^T$  correspond to "an arrow" of length 1 pointing in the positive direction of the  $x$ ,  $y$ ,  $z$  axes respectively.

**Definition 4.1.3** (Standard Unit Vector). A standard unit vector  $\hat{e}_p$  in the real  $n$ -space  $\mathbb{R}^n$  (Definition 4.1.2) has  $n$  components, consisted of 1 at the  $p$ -th entry and 0 elsewhere. Mathematically, for  $1 \leq q \leq n$ ,  $[\hat{e}_p]_q = 1$  when  $q = p$  and  $[\hat{e}_p]_q = 0$  when  $q \neq p$ .

Below is an example of a geometric vector in the three-dimensional  $xyz$  space ( $\mathbb{R}^3$ ).



$$\begin{aligned}\vec{v} &= \begin{bmatrix} 1.8 \\ 0.4 \\ 1.1 \end{bmatrix} = 1.8 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 0.4 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1.1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 1.8\hat{i} + 0.4\hat{j} + 1.1\hat{k} \\ &= (1.8, 0.4, 1.1)^T\end{aligned}$$

where we have written  $\vec{v}$  in two forms, as an  $n$ -tuple and a sum of the three standard unit vectors  $\hat{i}, \hat{j}, \hat{k}$ .

### 4.1.2 Fundamental Vector Operations

#### Addition and Subtraction

Same as their matrix counterpart, addition and subtraction between vectors are component-wise, and hence only valid for vectors of the same dimension. For  $\vec{w} = \vec{u} \pm \vec{v}$ , we have  $w_i = u_i \pm v_i$ . If

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \vec{v} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

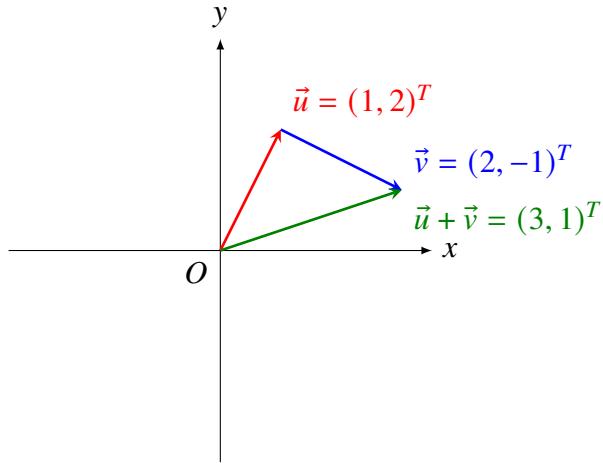
then

$$\vec{u} + \vec{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

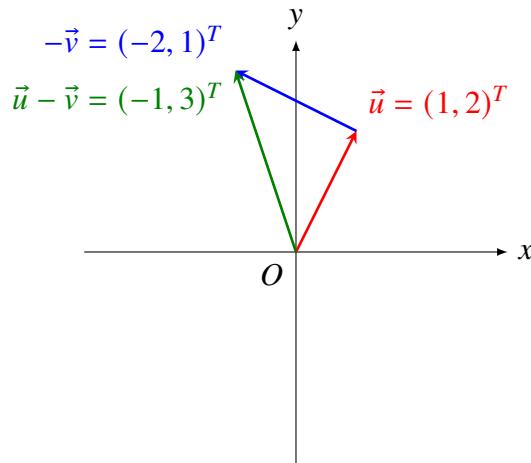
## 4.1 Definition and Operations of Geometric Vectors

---

$$\vec{u} - \vec{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$$



Addition: The tail of the blue vector is placed to the head of the red vector, and the resultant green vector is from the origin to the head of blue vector.



Subtraction: Similar to addition but with the blue vector oriented in the opposite direction.

## Scalar Multiplication

Multiplying a scalar (be it a real or complex number) to a vector means that all components are multiplied by that scalar.

$$2 \begin{bmatrix} 2 \\ 0 \\ 1 \\ 9 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 2 \\ 18 \end{bmatrix}$$

Looking back at vector subtraction, it can be viewed as addition with a factor of  $-1$  in the front.

$$\begin{bmatrix} 7 \\ 5 \\ 9 \end{bmatrix} - \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \\ 9 \end{bmatrix} + (-1) \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 7 \\ 5 \\ 9 \end{bmatrix} + \begin{bmatrix} -3 \\ -6 \\ -9 \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \\ 0 \end{bmatrix}$$

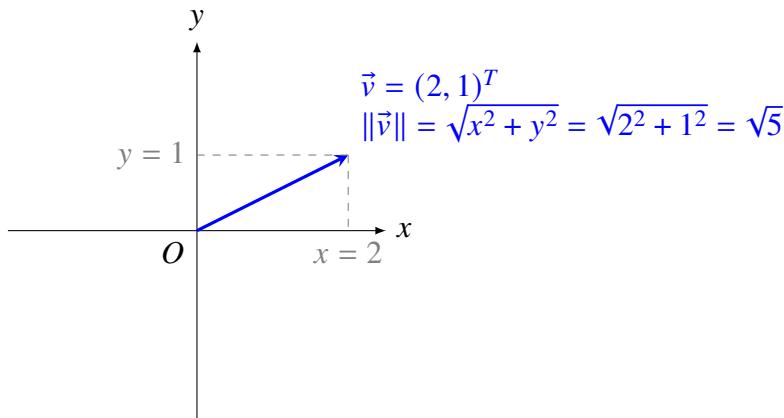
## Length and Unit Vector

**Length (magnitude)**, or more formally ***Euclidean norm***, of a vector  $\vec{v}$  is based on a generalized version of ***Pythagoras' Theorem***, and is evaluated as the square root of the sum of squares of components.

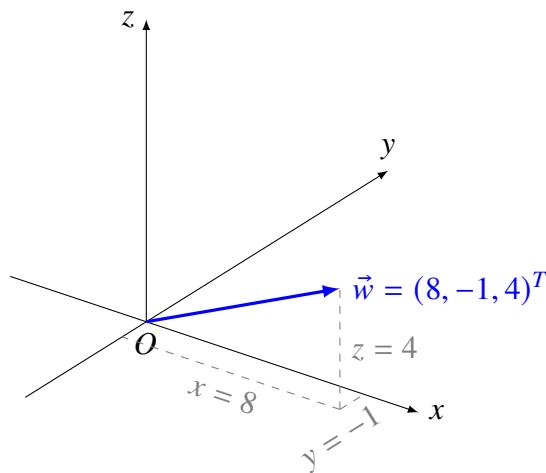
**Definition 4.1.4** (Vector Length). Length, or magnitude of a  $n$ -dimensional *real* vector  $\vec{v}$ , denoted by  $\|\vec{v}\|$ , is given by

$$\begin{aligned} \|\vec{v}\| &= \sqrt{v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2} \\ &= \sqrt{\sum_{k=1}^n v_k^2} \end{aligned}$$

For instance, the length of a two-dimensional vector follows the usual Pythagoras' Theorem as below.



Here is another example which is three-dimensional.



$$\vec{w} = \begin{bmatrix} 8 \\ -1 \\ 4 \end{bmatrix} \quad \|\vec{w}\| = \sqrt{8^2 + (-1)^2 + 4^2} = 9$$

We can create a **unit vector** that has a length of 1 from any vector  $\vec{v}$  and orients in the same direction as  $\vec{v}$ . It is simply created by dividing (normalizing)  $\vec{v}$  by its distance  $\|\vec{v}\|$ .

**Definition 4.1.5** (Unit Vector). The unit vector corresponding to a non-zero vector  $\vec{v}$  is denoted as  $\hat{v}$  and is given by

$$\hat{v} = \frac{1}{\|\vec{v}\|} \vec{v}$$

where the length  $\|\vec{v}\|$  is defined as in Definition 4.1.4.

Note that despite vectors can carry physical units, unit vectors are all physically *dimensionless* when formulated in this way.

Short Exercise: Find the unit vector for  $\vec{w} = (8, -1, 4)^T$  in the previous example, and verify that it has a length of 1.<sup>1</sup>

## 4.2 Special Vector Operations

Now we are going to introduce two special types of vector operations: *dot product*, and *cross product*.

### 4.2.1 Dot Product

(*Real*) **Dot product** (or *scalar product*) is defined for two (real) vectors that have the same number of dimension. It is the sum of products of paired components between the two vectors. In other words, it can be regarded to be the matrix product between a row vector ( $1 \times m$  matrix) and a column vector ( $m \times 1$  matrix).

**Definition 4.2.1** (Dot Product (Real)). The dot product between two  $n$ -dimensional *real* vectors  $\vec{u}$  and  $\vec{v}$  in  $\mathbb{R}^n$  are denoted as either  $\vec{u} \cdot \vec{v}$ , or by

---


$$^1 \|\vec{w}\| = 9, \hat{w} = \frac{\vec{w}}{\|\vec{w}\|} = \frac{1}{9}(8, -1, 4)^T = \left(\frac{8}{9}, -\frac{1}{9}, \frac{4}{9}\right)^T, \|\hat{w}\| = \sqrt{\left(\frac{8}{9}\right)^2 + \left(-\frac{1}{9}\right)^2 + \left(\frac{4}{9}\right)^2} = 1.$$

matrix notation  $\mathbf{u}^T \mathbf{v}$ . They are defined as

$$\begin{aligned}\vec{u} \cdot \vec{v} &= \mathbf{u}^T \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3 + \cdots + u_n v_n \\ &= \sum_{k=1}^n u_k v_k\end{aligned}$$

which is a scalar quantity.

Conversely, it can be said that entries of a matrix product are vector dot products between the corresponding rows and columns. It is emphasized that we are restricting ourselves to real entries since complex vectors introduce a bit of extra complications. Then, for two *real* matrices expressed in the form of combined row/column vectors that are  $\mathbb{R}^n$ ,

$$\begin{aligned}A &= \begin{bmatrix} \vec{u}^{(1)T} \\ \vec{u}^{(2)T} \\ \vdots \\ \vec{u}^{(p)T} \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(p)} \\ | & | & | & | \end{bmatrix}^T & B &= \begin{bmatrix} | & | & | \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \cdots & \vec{v}^{(q)} \\ | & | & | & | \end{bmatrix} \\ &= \begin{bmatrix} \vec{u}_1^{(1)} & \vec{u}_2^{(1)} & \cdots & \vec{u}_n^{(1)} \\ \vec{u}_1^{(2)} & \vec{u}_2^{(2)} & \cdots & \vec{u}_n^{(2)} \\ \vdots & \vdots & & \vdots \\ \vec{u}_1^{(p)} & \vec{u}_2^{(p)} & \cdots & \vec{u}_n^{(p)} \end{bmatrix} & &= \begin{bmatrix} \vec{v}_1^{(1)} & \vec{v}_1^{(2)} & \cdots & \vec{v}_1^{(q)} \\ \vec{v}_2^{(1)} & \vec{v}_2^{(2)} & \cdots & \vec{v}_2^{(q)} \\ \vdots & \ddots & \ddots & \vdots \\ \vec{v}_n^{(1)} & \vec{v}_n^{(2)} & \cdots & \vec{v}_n^{(q)} \end{bmatrix}\end{aligned}$$

(notice those transposes in the expression of  $A$ ) their matrix product  $AB$  can be written as

$$AB = \begin{bmatrix} \vec{u}^{(1)} \cdot \vec{v}^{(1)} & \vec{u}^{(1)} \cdot \vec{v}^{(2)} & \cdots & \vec{u}^{(1)} \cdot \vec{v}^{(q)} \\ \vec{u}^{(2)} \cdot \vec{v}^{(1)} & \vec{u}^{(2)} \cdot \vec{v}^{(2)} & \cdots & \vec{u}^{(2)} \cdot \vec{v}^{(q)} \\ \vdots & \vdots & & \vdots \\ \vec{u}^{(p)} \cdot \vec{v}^{(1)} & \vec{u}^{(p)} \cdot \vec{v}^{(2)} & \cdots & \vec{u}^{(p)} \cdot \vec{v}^{(q)} \end{bmatrix}$$

We can use dot product to express the length of a real vector.

**Properties 4.2.2.** The length of a real vector, as defined in Definition 4.1.4, can be written using its dot product between itself as

$$\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} \quad \text{or} \quad \|\vec{v}\|^2 = \vec{v} \cdot \vec{v}$$

Notice that  $\vec{v} \cdot \vec{v} = v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2 \geq 0$ . This quantity is always strictly greater than zero ( $\vec{v} \cdot \vec{v} > 0$ ) unless  $\vec{v} = \mathbf{0}$  is the zero vector (then  $\vec{v} \cdot \vec{v} = 0$ ), which makes sense physically given that it represents length.

**Example 4.2.1.** If  $\vec{u} = (1, 2, 3, 4, 5)^T$  and  $\vec{v} = (-1, 0, 2, 1, -2)^T$ , find the dot product  $\vec{u} \cdot \vec{v} = \mathbf{u}^T \mathbf{v}$ .

*Solution.*

$$\vec{u} \cdot \vec{v} = (1)(-1) + (2)(0) + (3)(2) + (4)(1) + (5)(-2) = -1$$

Alternatively,

$$\mathbf{u}^T \mathbf{v} = [1 \ 2 \ 3 \ 4 \ 5] \begin{bmatrix} -1 \\ 0 \\ 2 \\ 1 \\ -2 \end{bmatrix} = -1$$

□

Here are some properties of dot product.

**Properties 4.2.3.** For three  $n$ -dimensional real vectors  $\vec{u}$ ,  $\vec{v}$  and  $\vec{w}$ , the following properties hold.

$$\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$$

Symmetry Property

$$\vec{u} \cdot (\vec{v} \pm \vec{w}) = \vec{u} \cdot \vec{v} \pm \vec{u} \cdot \vec{w}$$

Distributive Property

$$(\vec{u} \pm \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} \pm \vec{v} \cdot \vec{w}$$

Distributive Property

$$(a\vec{u}) \cdot (b\vec{v}) = ab(\vec{u} \cdot \vec{v}) \quad \text{where } a, b \text{ are some constants}$$

Additionally, if  $A$  is an  $n \times n$  square matrix, then

$$\begin{aligned}\vec{u} \cdot (A\vec{v}) &= \mathbf{u}^T (A\mathbf{v}) = (A^T \mathbf{u})^T \mathbf{v} = (A^T \vec{u}) \cdot \vec{v} \\ (A\vec{u}) \cdot \vec{v} &= (A\mathbf{u})^T \mathbf{v} = \mathbf{u}^T (A^T \mathbf{v}) = \vec{u} \cdot (A^T \vec{v})\end{aligned}$$

where we have used Definition 4.2.1 and Properties 2.1.4.

**Example 4.2.2.** For  $\vec{u} = (1, 3, 1)^T$  and  $\vec{v} = (2, -1, 1)^T$ , find  $\|(\vec{u} + \vec{v})\|^2 = (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v})$ .

*Solution.* By Properties 4.2.3, we can rewrite the expression as

$$\begin{aligned}(\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) &= \vec{u} \cdot (\vec{u} + \vec{v}) + \vec{v} \cdot (\vec{u} + \vec{v}) \\ &= \vec{u} \cdot \vec{u} + \vec{u} \cdot \vec{v} + \vec{v} \cdot \vec{u} + \vec{v} \cdot \vec{v} \\ &= \vec{u} \cdot \vec{u} + 2\vec{u} \cdot \vec{v} + \vec{v} \cdot \vec{v}\end{aligned}$$

Subsequently,

$$\begin{aligned}&\vec{u} \cdot \vec{u} + 2\vec{u} \cdot \vec{v} + \vec{v} \cdot \vec{v} \\ &= (1, 3, 1)^T \cdot (1, 3, 1)^T + 2((1, 3, 1)^T \cdot (2, -1, 1)^T) + (2, -1, 1)^T \cdot (2, -1, 1)^T \\ &= (1^2 + 3^2 + 1^2) + 2((1)(2) + (3)(-1) + (1)(1)) + (2^2 + (-1)^2 + 1^2) \\ &= 11 + 2(0) + 6 \\ &= 17\end{aligned}$$

Alternatively, one can calculate  $\vec{w} = \vec{u} + \vec{v} = (1, 3, 1)^T + (2, -1, 1)^T = (3, 2, 2)^T$  and find  $\vec{w} \cdot \vec{w} = \|\vec{w}\|^2$  instead. (which is easier and faster)  $\square$

**Example 4.2.3.** Given  $\vec{u}$  and  $\vec{v}$  as defined in the example above, if

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

verify that  $\vec{u} \cdot (A\vec{v}) = (A^T\vec{u}) \cdot \vec{v}$ .

*Solution.*

$$\begin{aligned} A\vec{v} &= \begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} (1)(2) + (2)(-1) + (1)(1) \\ (2)(2) + (0)(-1) + (3)(1) \\ (1)(2) + (1)(-1) + (-1)(1) \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 7 \\ 0 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \vec{u} \cdot (A\vec{v}) &= (1, 3, 1)^T \cdot (1, 7, 0)^T \\ &= (1)(1) + (3)(7) + (1)(0) \\ &= 22 \end{aligned}$$

On the other hand,

$$\begin{aligned} A^T\vec{u} &= \begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} (1)(1) + (2)(3) + (1)(1) \\ (2)(1) + (0)(3) + (1)(1) \\ (1)(1) + (3)(3) + (-1)(1) \end{bmatrix} \\ &= \begin{bmatrix} 8 \\ 3 \\ 9 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
 (A^T \vec{u}) \cdot \vec{v} &= (8, 3, 9)^T \cdot (2, -1, 1)^T \\
 &= (8)(2) + (3)(-1) + (9)(1) \\
 &= 22
 \end{aligned}$$

□

### Geometric Meaning of Dot Product

The geometric meaning of dot product is embedded in the relation below.

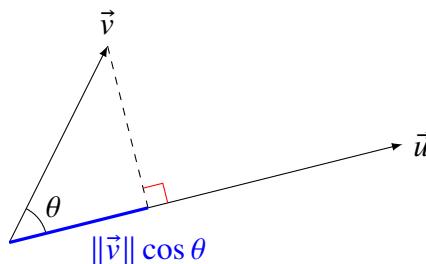
**Properties 4.2.4.** For two real vectors  $\vec{u}$  and  $\vec{v}$  that are of the same dimension, we have

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos \theta$$

where  $\theta$  is the angle between  $\vec{u}$  and  $\vec{v}$ . Furthermore, if  $\hat{u}$  and  $\hat{v}$  are unit vectors (Definition 4.1.5) such that  $\|\vec{u}\| = \|\vec{v}\| = 1$ , it reduces to

$$\hat{u} \cdot \hat{v} = \cos \theta$$

This means that the dot product between two vectors  $\vec{u}$  and  $\vec{v}$  is geometrically the signed product between  $\vec{u}$  and the parallel component (projection) of  $\vec{v}$  onto  $\vec{u}$  (or vice versa), which is illustrated in the figure below. While an angle has a clear physical meaning only in a two/three-dimensional space, such relation generalizes the idea of an angle to higher dimensions.



**Example 4.2.4.** Find the angle between  $\vec{u}$  and  $\vec{v}$  in Example 4.2.1.

*Solution.* From Example 4.2.1, we have  $\vec{u} \cdot \vec{v} = -1$ , and

$$\begin{aligned}\|\vec{u}\| &= \sqrt{1^2 + 2^2 + 3^2 + 4^2 + 5^2} = \sqrt{55} \\ \|\vec{v}\| &= \sqrt{(-1)^2 + 0^2 + 2^2 + 1^2 + (-2)^2} = \sqrt{10}\end{aligned}$$

By Properties 4.2.4, we have

$$\begin{aligned}\cos \theta &= \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|} \\ &= \frac{-1}{(\sqrt{55})(\sqrt{10})} \\ &\approx -0.0426 \\ \theta &\approx 1.613 \text{ rad} = 92.44^\circ\end{aligned}$$

□

By Properties 4.2.4, if the absolute value of the dot product  $|\vec{u} \cdot \vec{v}|$  is equal to  $\|\vec{u}\| \|\vec{v}\|$ , where  $\vec{u}$  and  $\vec{v}$  are non-zero vectors, then it implies that  $\cos \theta = \pm 1$ ,  $\theta$  is either 0 or  $\pi$ , and hence the two vectors are parallel. In contrast,

**Properties 4.2.5.** If the dot product between two real vectors  $\vec{u}$  and  $\vec{v}$  is zero ( $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u} = 0$ ), then by Properties 4.2.4,  $\cos \theta = 0$  and the angle  $\theta$  between  $\vec{u}$  and  $\vec{v}$  is  $\frac{\pi}{2}$ . In this case,  $\vec{u}$  and  $\vec{v}$  are said to be perpendicular, or *orthogonal* to each other.

From this, the concept of "*orthogonal*" becomes an extension of "perpendicular" in higher dimensions. It is easy to see that the standard unit vectors of  $\mathbb{R}^n$  are orthogonal. Note that *the zero vector is regarded to be orthogonal to any vector*, so even if  $\vec{u}$  or  $\vec{v}$  is a zero vector, this properties still hold.

Some may notice that as  $-1 \leq \cos \theta \leq 1$ , if  $|\vec{u} \cdot \vec{v}| > \|\vec{u}\| \|\vec{v}\|$ , then  $\theta$  in Properties 4.2.4 will be ill-defined. However, the *Cauchy–Schwarz Inequality* ensures this will not happen.

**Theorem 4.2.6** (Cauchy–Schwarz Inequality). Given two *real*  $n$ -dimensional vectors  $\vec{u}$  and  $\vec{v}$  ( $\vec{u}, \vec{v} \in \mathbb{R}^n$ ), the following inequality holds.

$$|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\| \|\vec{v}\|$$

$$|u_1 v_1 + u_2 v_2 + \cdots + u_n v_n| \leq \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2} \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

*Proof.* Consider  $\vec{w} = \vec{u} + t\vec{v}$ , where  $t$  is any scalar, then  $\|\vec{w}\|^2 = \vec{w} \cdot \vec{w} \geq 0$  by Properties 4.2.2. Also,  $\vec{w} \cdot \vec{w}$  can be written as a quadratic polynomial in  $t$ :

$$\vec{w} \cdot \vec{w} = (\vec{u} + t\vec{v}) \cdot (\vec{u} + t\vec{v}) = \|\vec{u}\|^2 + 2t(\vec{u} \cdot \vec{v}) + t^2 \|\vec{v}\|^2$$

Since this quantity is always greater than or equal to zero, i.e. the quadratic polynomial has no root or a repeated root, it means that the discriminant must be negative or zero. So,

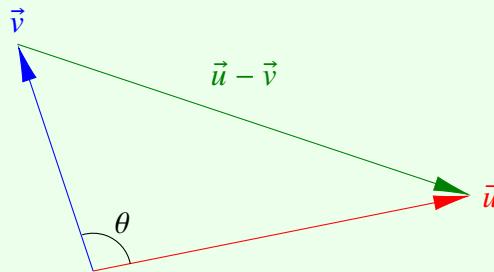
$$\begin{aligned} \Delta &= b^2 - 4ac \leq 0 \\ (2(\vec{u} \cdot \vec{v}))^2 - 4\|\vec{u}\|^2 \|\vec{v}\|^2 &\leq 0 \\ (\vec{u} \cdot \vec{v})^2 - \|\vec{u}\|^2 \|\vec{v}\|^2 &\leq 0 \\ (\vec{u} \cdot \vec{v})^2 &\leq \|\vec{u}\|^2 \|\vec{v}\|^2 \\ |\vec{u} \cdot \vec{v}| &\leq \|\vec{u}\| \|\vec{v}\| \end{aligned}$$

□

Short Exercise: Think about under what circumstances the Cauchy–Schwarz Inequality turns into an equality (i.e.  $|\vec{u} \cdot \vec{v}| = \|\vec{u}\| \|\vec{v}\|$ ).<sup>2</sup>

**Example 4.2.5.** Prove the *Cosine Law* by considering the triangle below

<sup>2</sup>When  $\vec{u}$  and  $\vec{v}$  are parallel, i.e.  $\vec{u} = k\vec{v}$  for some scalar  $k$ , or  $\vec{v} = \mathbf{0}$ .



and expanding the dot product  $\|(\vec{u} - \vec{v})\|^2 = (\vec{u} - \vec{v}) \cdot (\vec{u} - \vec{v})$ .

*Solution.* Let denote the lengths  $\|\vec{u}\|$ ,  $\|\vec{v}\|$ ,  $\|(\vec{u} - \vec{v})\|$  be  $a$ ,  $b$ ,  $c$ , then

$$\begin{aligned} c^2 &= \|(\vec{u} - \vec{v})\|^2 = (\vec{u} - \vec{v}) \cdot (\vec{u} - \vec{v}) && \text{(Properties 4.2.2)} \\ &= \vec{u} \cdot \vec{u} - \vec{u} \cdot \vec{v} - \vec{v} \cdot \vec{u} + \vec{v} \cdot \vec{v} && \text{(Properties 4.2.3)} \\ &= \|\vec{u}\|^2 - 2\vec{u} \cdot \vec{v} + \|\vec{v}\|^2 && \text{(Properties 4.2.2 and 4.2.3)} \\ &= \|\vec{u}\|^2 - 2\|\vec{u}\|\|\vec{v}\| \cos \theta + \|\vec{v}\|^2 && \text{(Properties 4.2.4)} \\ &= a^2 - 2ab \cos \theta + b^2 \end{aligned}$$

□

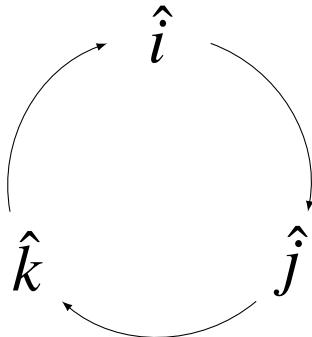
## 4.2.2 Cross Product

Another important type of vector product is the ***cross product*** (or sometimes just ***vector product***), which produces a three-dimensional real vector from two other three-dimensional real vectors. *The output vector will be orthogonal to the two input vectors*, and the direction is determined by the ***right hand rule***. Motivated by these requirements, we have the following basic definitions of cross product between the three standard unit vectors in  $\mathbb{R}^3$ .

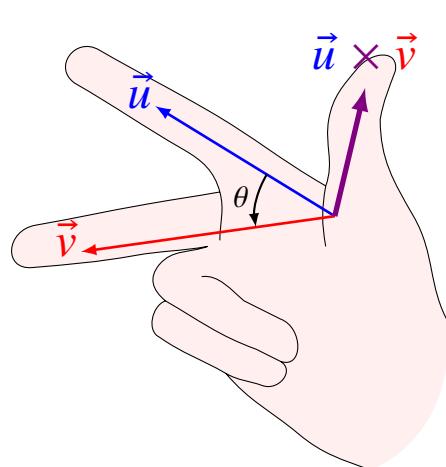
**Definition 4.2.7.** The computation of cross products (denoted by  $\times$ ) involving any two of the standard unit vectors  $\hat{i}, \hat{j}, \hat{k}$  in  $\mathbb{R}^3$  obeys the following rules.

1.  $\hat{i} \times \hat{j} = \hat{k}, \hat{j} \times \hat{i} = -\hat{k}$ ,

2.  $\hat{j} \times \hat{k} = \hat{i}$ ,  $\hat{k} \times \hat{j} = -\hat{i}$ ,
3.  $\hat{k} \times \hat{i} = \hat{j}$ ,  $\hat{i} \times \hat{k} = -\hat{j}$ , and
4.  $\hat{i} \times \hat{i} = \hat{j} \times \hat{j} = \hat{k} \times \hat{k} = \mathbf{0}$



A cyclic diagram for memorizing Definition 4.2.7. A clockwise / anti-clockwise permutation produces a positive / negative unit vector of the third.



Demonstration of the right hand rule.

The properties of cross product are noted below. One major difference setting cross product apart from the dot product is its anti-symmetric property.

**Properties 4.2.8.** For two  $\mathbb{R}^3$  vectors  $\vec{u}$  and  $\vec{v}$ , we have

$\vec{u} \times \vec{v} = -\vec{v} \times \vec{u}$	Anti-symmetry Property
$\vec{u} \times (\vec{v} \pm \vec{w}) = \vec{u} \times \vec{v} \pm \vec{u} \times \vec{w}$	Distributive Property
$(\vec{u} \pm \vec{v}) \times \vec{w} = \vec{u} \times \vec{w} \pm \vec{v} \times \vec{w}$	Distributive Property
$(a\vec{u}) \times (b\vec{v}) = ab(\vec{u} \times \vec{v})$	where $a, b$ are some constants

The calculation of cross product then follows from these rules, leading to the determinant shorthand below.

**Properties 4.2.9.** For  $\vec{u} = (u_1, u_2, u_3)^T, \vec{v} = (v_1, v_2, v_3)^T \in \mathbb{R}^3$ , their cross product  $\vec{u} \times \vec{v}$  can be written in the form of a determinant as

$$\vec{u} \times \vec{v} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

*Proof.* Starting from Definition 4.2.7 and Properties 4.2.8, we have

$$\begin{aligned} \vec{u} \times \vec{v} &= (u_1 \hat{i} + u_2 \hat{j} + u_3 \hat{k}) \times (v_1 \hat{i} + v_2 \hat{j} + v_3 \hat{k}) \\ &= u_1 v_1 (\hat{i} \times \hat{i}) + u_1 v_2 (\hat{i} \times \hat{j}) + u_1 v_3 (\hat{i} \times \hat{k}) \\ &\quad + u_2 v_1 (\hat{j} \times \hat{i}) + u_2 v_2 (\hat{j} \times \hat{j}) + u_2 v_3 (\hat{j} \times \hat{k}) \\ &\quad + u_3 v_1 (\hat{k} \times \hat{i}) + u_3 v_2 (\hat{k} \times \hat{j}) + u_3 v_3 (\hat{k} \times \hat{k}) \quad (\text{Properties 4.2.8}) \\ &= u_1 v_1 (\mathbf{0}) + u_1 v_2 (\hat{k}) - u_1 v_3 (\hat{j}) \\ &\quad - u_2 v_1 (\hat{k}) + u_2 v_2 (\mathbf{0}) + u_2 v_3 (\hat{i}) \\ &\quad + u_3 v_1 (\hat{j}) - u_3 v_2 (\hat{i}) + u_3 v_3 (\mathbf{0}) \quad (\text{Definition 4.2.7}) \\ &= (u_2 v_3 - u_3 v_2) \hat{i} + (u_3 v_1 - u_1 v_3) \hat{j} + (u_1 v_2 - u_2 v_1) \hat{k} \end{aligned}$$

Meanwhile, cofactor expansion (Properties 2.3.3) along the first row of the given determinant form

$$\begin{aligned} \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} &= \hat{i} \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} - \hat{j} \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} + \hat{k} \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \\ &= (u_2 v_3 - u_3 v_2) \hat{i} + (u_3 v_1 - u_1 v_3) \hat{j} + (u_1 v_2 - u_2 v_1) \hat{k} \end{aligned}$$

yields the identical result. □

**Example 4.2.6.** Given two  $\mathbb{R}^3$  vectors

$$\vec{u} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \quad \vec{v} = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix}$$

Find  $\vec{u} \times \vec{v}$ .

*Solution.*

$$\begin{aligned}\vec{u} \times \vec{v} &= \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 1 & 0 & 2 \\ 3 & -1 & 1 \end{vmatrix} \\ &= \hat{i} \begin{vmatrix} 0 & 2 \\ -1 & 1 \end{vmatrix} - \hat{j} \begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix} + \hat{k} \begin{vmatrix} 1 & 0 \\ 3 & -1 \end{vmatrix} \quad (\text{Cofactor expansion along the first row}) \\ &= 2\hat{i} + 5\hat{j} - \hat{k} = (2, 5, -1)^T\end{aligned}$$

□

Short Exercise: Check if  $\vec{u} \times \vec{v}$  is orthogonal to  $\vec{u}$  and  $\vec{v}$  by finding the corresponding dot products.<sup>3</sup>

Short Exercise: Following the short exercise above, show in general,  $\vec{u} \cdot (\vec{u} \times \vec{v}) = \vec{v} \cdot (\vec{u} \times \vec{v}) = 0$ .<sup>4</sup>

## Geometric Meaning of Cross Product

Similar to vector dot product, vector cross product has a geometric interpretation.

---

<sup>3</sup> $\vec{u} \cdot (\vec{u} \times \vec{v}) = (1, 0, 2)^T \cdot (2, 5, -1)^T = (1)(2) + (0)(5) + (2)(-1) = 0$ ,  $\vec{v} \cdot (\vec{u} \times \vec{v}) = (3, -1, 1)^T \cdot (2, 5, -1)^T = (3)(2) + (-1)(5) + (1)(-1) = 0$ . The zero dot product in both cases shows they are orthogonal via Properties 4.2.5.

<sup>4</sup>From the derivation of Properties 4.2.9,  $\vec{u} \times \vec{v} = (u_2 v_3 - u_3 v_2)\hat{i} + (u_3 v_1 - u_1 v_3)\hat{j} + (u_1 v_2 - u_2 v_1)\hat{k}$ , and  $\vec{u} \cdot (\vec{u} \times \vec{v}) = u_1(u_2 v_3 - u_3 v_2) + u_2(u_3 v_1 - u_1 v_3) + u_3(u_1 v_2 - u_2 v_1) = 0$  where all terms cancel out, and it is similar for  $\vec{v}$ .

**Properties 4.2.10.** Given two vectors  $\vec{u}$  and  $\vec{v}$  which are both of  $\mathbb{R}^3$ , the magnitude (length) of  $\vec{u} \times \vec{v}$  is related to the angle between  $\vec{u}$  and  $\vec{v}$  as

$$\|\vec{u} \times \vec{v}\| = \|\vec{u}\| \|\vec{v}\| \sin \theta$$

From this, we immediately know that if  $\vec{u}$  and  $\vec{v} = k\vec{u}$ , where  $k$  is some constant, are two parallel vectors, their cross product will be a zero vector as  $\theta = 0$  (or  $\pi$ ) and  $\sin \theta = 0$ . This is equivalent to the statement of  $\vec{u} \times \vec{u} = \mathbf{0}$ <sup>5</sup> (notice that it is not 0 but  $\mathbf{0}$  since it always outputs a vector!). (You can also arrive at this conclusion with Properties 4.2.8.<sup>6</sup>)

**Example 4.2.7.** If  $\vec{u} = (1, 2, 3)^T$ , and  $\vec{v} = (-1, 1, 2)^T$ , find  $(\vec{u} + 2\vec{v}) \times (\vec{u} - \vec{v})$ .

*Solution.* Observe that

$$\begin{aligned} (\vec{u} + 2\vec{v}) \times (\vec{u} - \vec{v}) &= \vec{u} \times (\vec{u} - \vec{v}) + 2\vec{v} \times (\vec{u} - \vec{v}) \\ &= \vec{u} \times \vec{u} - \vec{u} \times \vec{v} + 2\vec{v} \times \vec{u} - 2\vec{v} \times \vec{v} \\ &= \mathbf{0} - \vec{u} \times \vec{v} - 2\vec{u} \times \vec{v} - 2(\mathbf{0}) \\ &= -3\vec{u} \times \vec{v} \end{aligned}$$

where the fact that  $\vec{u} \times \vec{u} = \mathbf{0}$ ,  $\vec{v} \times \vec{v} = \mathbf{0}$  and Properties 4.2.8 are used. Now, with Properties 4.2.9, we have

$$-3\vec{u} \times \vec{v} = -3 \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 1 & 2 & 3 \\ -1 & 1 & 2 \end{vmatrix}$$

<sup>5</sup>By Properties 4.2.9,

$$\vec{u} \times \vec{u} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_1 & u_2 & u_3 \\ u_1 & u_2 & u_3 \end{vmatrix}$$

and the determinant vanishes by Properties 2.3.4 due to the identical second/third row.

<sup>6</sup>The anti-symmetric property requires  $\vec{u} \times \vec{u} = -\vec{u} \times \vec{u}$  and hence  $2(\vec{u} \times \vec{u}) = \mathbf{0}$ .

$$\begin{aligned}
 &= -3 \left( \hat{i} \begin{vmatrix} 2 & 3 \\ 1 & 2 \end{vmatrix} - \hat{j} \begin{vmatrix} 1 & 3 \\ -1 & 2 \end{vmatrix} + \hat{k} \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} \right) \quad (\text{Cofactor expansion along the first row}) \\
 &= -3(\hat{i} - 5\hat{j} + 3\hat{k}) \\
 &= -3\hat{i} + 15\hat{j} - 9\hat{k} = (-3, 15, -9)^T
 \end{aligned}$$

The readers can try the alternative of computing  $\vec{u} + 2\vec{v}$  and  $\vec{u} - \vec{v}$  first and then their cross product.  $\square$

Finally, cancellation of dot product or cross product at both sides of an equation is generally not correct, and here is a table summarizing the inputs and outputs of dot/cross product for clarification.

	Input	Output
Dot Product, or Scalar Product ( $\cdot$ )	Two real vectors of the same dimension ( $\mathbb{R}^n$ ), the order does not matter (symmetric)	A scalar
Cross Product, or Vector Product ( $\times$ )	Two three-dimensional real vectors ( $\mathbb{R}^3$ ), the order is important (anti-symmetric)	Another three-dimensional vector

## 4.3 Earth Science Applications

**Example 4.3.1.** The *Coriolis Effect* is a phenomenon describing the deflection of motion due to rotation of the Earth. It introduces an apparent force known as the *Coriolis Force* which is given by  $\vec{F}_{\text{cor}} = -2\vec{\Omega} \times \vec{v}$  where  $\vec{\Omega} = \|\vec{\Omega}\| = 7.292 \times 10^{-5} \text{ rad s}^{-1}$  represents the angular speed of Earth's rotation, and  $\vec{\Omega}$  is oriented in the direction of the North Pole. Define the local frame of reference (see Figure 4.1) with the  $x$ -direction being the zonal direction,  $y$ -direction being the meridional direction, and  $z$ -direction being the zenith direction (normal to the Earth's surface), then we have  $\vec{v} = (u, v, w) = u\hat{i} + v\hat{j} + w\hat{k}$  as the flow velocity in this local Cartesian coordinate system with unit vectors  $\hat{i}, \hat{j}, \hat{k}$  along

the  $x$ ,  $y$ ,  $z$  axes. It can be seen that  $\vec{\Omega} = (\Omega \cos \varphi)\hat{j} + (\Omega \sin \varphi)\hat{k}$  where  $\varphi$  is the latitude. Now by expanding  $\overrightarrow{F_{\text{cor}}} = -2\vec{\Omega} \times \vec{v}$  show that the components of Coriolis Force along the local  $x$ ,  $y$ ,  $z$  directions are

$$F_{\text{cor},x} = 2\Omega(v \sin \varphi - w \cos \varphi)$$

$$F_{\text{cor},y} = -2\Omega u \sin \varphi$$

$$F_{\text{cor},z} = 2\Omega u \cos \varphi$$

The *Coriolis Parameter*  $f$  is usually used to denote the factor  $2\Omega \sin \varphi$ .

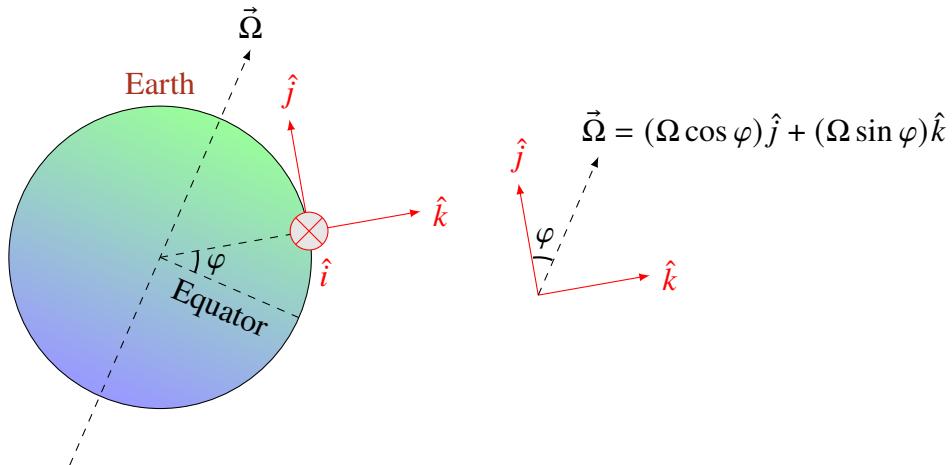


Figure 4.1: An illustration of the coordinate frame in Example 4.3.1.

*Solution.* Using Properties 4.2.9 to expand  $\overrightarrow{F_{\text{cor}}}$  gives

$$\begin{aligned} -2\vec{\Omega} \times \vec{v} &= -2((\Omega \cos \varphi)\hat{j} + (\Omega \sin \varphi)\hat{k}) \times (u\hat{i} + v\hat{j} + w\hat{k}) \\ &= -2 \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 0 & \Omega \cos \varphi & \Omega \sin \varphi \\ u & v & w \end{vmatrix} \\ &= -2[(w\Omega \cos \varphi - v\Omega \sin \varphi)\hat{i} + (u\Omega \sin \varphi)\hat{j} - (u\Omega \cos \varphi)\hat{k}] \end{aligned}$$

$$= [2\Omega(v \sin \varphi - w \cos \varphi)]\hat{i} + (-2\Omega u \sin \varphi)\hat{j} + (2\Omega u \cos \varphi)\hat{k}$$

The  $\hat{i}$ ,  $\hat{j}$ ,  $\hat{k}$  components correspond to  $F_{\text{cor},x}$ ,  $F_{\text{cor},y}$ ,  $F_{\text{cor},z}$  respectively. Assume  $w$  is negligible, then  $F_{\text{cor},x} = fv$  and  $F_{\text{cor},y} = -fu$ .  $\square$

## 4.4 Python Programming

We can use one-dimensional numpy arrays as vectors.

```
import numpy as np

myVec1 = np.array([-1., 2., 4.])
myVec2 = np.array([2., 1., 3.])
```

Addition, subtraction, and scalar multiplication works just like for matrices.

```
myVec3 = -myVec1 + 2*myVec2
print(myVec3)
```

gives the expected output of  $[5. \ 0. \ 2.]$ . We can select a component of any vector by indexing. Again, remember that indices in *Python* start from zero. `print(myVec3[1])` then returns  $0.0$ . The magnitude of a vector can be checked with `np.linalg.norm`. For example,

```
print(np.linalg.norm(myVec1))
```

produces  $4.58257569495584$  ( $\sqrt{(-1)^2 + 2^2 + 4^2} = \sqrt{21}$ ). Dot product is computed via `np.dot` as follows.

```
myDot = np.dot(myVec1, myVec2)
print(myDot)
```

which outputs  $12.0$  (as  $(-1)(2) + (2)(1) + (4)(3) = 12$ ). Similarly, cross product is found by `np.cross`.

```
myCross = np.cross(myVec1, myVec2)
print(myCross)
```

then gives

```
[ 2. 11. -5.]
```

and we can check if the cross product is orthogonal to the two input vectors.

```
# All lines below return zero.
print(np.dot(myVec1, myCross))
print(np.dot(myVec2, myCross))
print(np.dot(myVec3, myCross))
```

Dot product is defined for any two vectors with the same dimension, but cross product is only defined for three-dimensional vectors (or in some other sense two-dimensional), so

```
myVec4 = np.array([1., 3., 2., 0.])
myVec5 = np.array([2., 1., 0., -1.])
print(np.dot(myVec4, myVec5))
```

yields a valid output of 5.0, but

```
print(np.cross(myVec4, myVec5))
```

raises the error of

```
ValueError: incompatible dimensions for cross product
(dimension must be 2 or 3)
```

Finally, we note that following [this Stack Overflow post](#) (2827393), we can compute the unit vector of any given vector and angle between two vectors (based from the second observation in Properties 4.2.4,  $\theta = \cos^{-1}(\hat{u} \cdot \hat{v})$ ).

```
def unit_vector(vector):
    """ Returns the unit vector of the vector. """
    return vector / np.linalg.norm(vector)

def angle_between(v1, v2):
    """ Returns the angle in radians between vectors 'v1' and 'v2'. """
    v1_u = unit_vector(v1)
    v2_u = unit_vector(v2)
    return np.arccos(np.clip(np.dot(v1_u, v2_u), -1.0, 1.0))
```

The `np.clip` is to avoid numerical round-off error that causes the dot product of the two normalized input vectors to just fall outside (e.g. `1.0000000000000002`) the valid range  $[-1, 1]$  of  $\cos^{-1}$ . The naive way of (here the lists will be cast to one-dimensional arrays automatically during calculation.)

```
np.arccos(np.dot([1., 0, 0], [2., 0, 0]))
```

leads to the warning of

```
RuntimeWarning: invalid value encountered in arccos
nan
```

but

```
angle_between([1., 0, 0], [2., 0, 0])
```

gives  $0.0$  properly. Trying this on `myVec4` and `myVec5` with

```
print(unit_vector(myVec4))
print(angle_between(myVec4, myVec5))
```

produces a unit vector of  $[0.267 \ 0.802 \ 0.535 \ 0.]$ , and an angle of  $0.993757$  (in radians).

## 4.5 Exercises

**Exercise 4.1** For  $\vec{u} = (1, 3, 3, 7)^T$  and  $\vec{v} = (1, 2, 2, 5)^T$ , find

- (a)  $\vec{u} + \vec{v}$ ,
- (b)  $\frac{3}{2}\vec{u} - \frac{1}{2}\vec{v}$ ,
- (c)  $\vec{u} \cdot \vec{v}$ ,
- (d)  $\vec{v} \cdot \vec{u}$ ,
- (e)  $(\vec{u} - 2\vec{v}) \cdot (2\vec{u} + \vec{v})$ .

**Exercise 4.2** For  $\vec{u} = (7, 4, 1)^T$ ,  $\vec{v} = (8, 1, 1)^T$ , and

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

Verify that

- (a)  $\vec{u} \times \vec{v} = -\vec{v} \times \vec{u}$ ,
- (b)  $\vec{u} \cdot (\vec{A}\vec{v}) = (A^T \vec{u}) \cdot \vec{v}$ ,
- (c) Compute  $(3\vec{u} - 4\vec{v}) \cdot (\vec{u} \times \vec{v})$ , is the answer what you expect?

**Exercise 4.3** For  $\vec{u} = (1, -3, 9)^T$  and  $\vec{v} = (1, -2, 4)^T$ , find

- (a) Their unit vectors  $\hat{u}$  and  $\hat{v}$ ,
- (b) The angle between them, by calculating their dot product,
- (c) The cross product  $\vec{u} \times \vec{v}$ , and
- (d) Show that the vector obtained from the cross product above is orthogonal (perpendicular) to  $\vec{u}$  and  $\vec{v}$ , by calculating the corresponding dot products.

**Exercise 4.4** The following table contains incomplete data about the movement of several typhoons at some moments. Complete the table by filling in the blanks. The first one has been done as an example.

Typhoon Name	Time	Speed	Direction	Vector Form
Nuri	2008/08/22, 08:00	$13 \text{ km h}^{-1}$	$315^\circ$	$(-9.192, 9.192)$
Vicente	2012/07/24, 02:00	$18 \text{ km h}^{-1}$	$299^\circ$	
Linfa	2015/07/09, 23:00			$(-13.595, -6.339)$
Mangkhut	2018/09/16, 22:00		$288^\circ$	$(\ , 7.725)$

**Exercise 4.5** Prove the Triangular Inequality.

$$\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$$

**Exercise 4.6** Prove the Parallelogram Law. (See Figure 4.2)

$$2\|\vec{u}\|^2 + 2\|\vec{v}\|^2 = \|\vec{u} + \vec{v}\|^2 + \|\vec{u} - \vec{v}\|^2$$

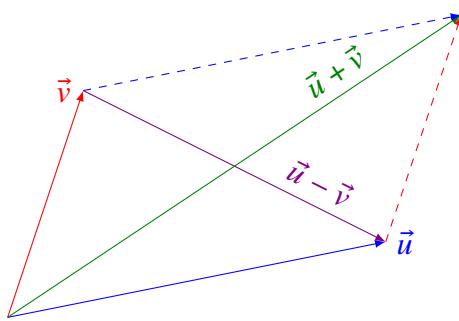


Figure 4.2: The parallelogram constructed by vectors for Exercise 4.6.

**Exercise 4.7** Show that Coriolis Force derived in Example 4.3.1 does zero work and hence is consistent with the fact that it is an apparent force and never produces/consumes energy by itself.



## Chapter 5

# Vector Geometry

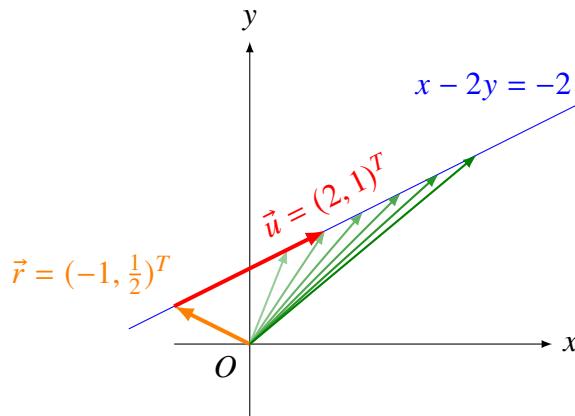
---

Vectors provides valuable assistance when it comes to describing geometric objects. In this chapter we are going to exploit the knowledge learnt in the previous chapters to solve geometry problems and inspect more deeply the intimate relationship between vectors, dot/cross products, and geometry.

## 5.1 Lines and Planes

(*Straight*) *lines* and *planes* are geometric shapes of importance in two/three-dimensional real spaces ( $\mathbb{R}^2$  and  $\mathbb{R}^3$ ) and due to their simplicity they will be frequently seen. They can be expressed either in terms of (a) an equation, and (b) vectors. We will start from the easier case of a line.

Since a straight line is a one-dimensional object, the vector form of such a line can be expressed by a fixed vector that points to its initial position, plus another vector oriented along the line's direction, times an arbitrary parameter which controls its extension or contraction, so that it traces out the line when the parameter changed continuously.



The graph of  $x - 2y = -2$  can take the vector form of  $\overrightarrow{OP} = \vec{r} + t\vec{u} = (-1, \frac{1}{2})^T + t(2, 1)^T$ . The orange/red arrow represents the initial position/direction, and the locus of green arrow is controlled by  $t$  like a slider. The cases for  $t = 0.75, 1, 1.25, 1.5, 1.75, 2$  are shown.

Short Exercise: Choose any value of  $t$  and substitute that value into the expression of  $\overrightarrow{OP}$  above to see if the  $x$  and  $y$ -components satisfy the starting equation. Also, try to increase/decrease the value of  $t$  to observe how the vector traces out the desired straight line.<sup>1</sup>

### 5.1.1 Translating Equation Form to Vector Form

The general equation form of a line on an  $x$ - $y$  plane is  $ax + by = h$ , resembling a linear system of one equation with two unknowns. From Section 3.2.1, it can be observed that it has infinitely many solutions and possesses a free variable. Let  $y = t$ , then rearranging the equation we have  $x = (h - bt)/a$  where  $t$  is any scalar. Denote the origin as  $O$  and any point on the line as  $P$ , then

$$\overrightarrow{OP} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{h}{a} - \frac{b}{a}t \\ t \end{bmatrix} = \begin{bmatrix} \frac{h}{a} \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{b}{a} \\ 1 \end{bmatrix}$$

---

<sup>1</sup>Let's say  $t = -0.25$ ,  $\overrightarrow{OP} = (-1, 0.5)^T + (-0.25)(2, 1)^T = (-1.5, 0.25)^T$ ,  $x - 2y = (-1.5) - 2(0.25) = -2$ .

This is one possible vector form (*parameterization*) of the line. Its idea can be borrowed from Example 3.2.3, with  $(\frac{h}{a}, 0)^T$  being the initial position/particular solution, and  $(-\frac{b}{a}, 1)^T$  as the direction of that line, multiplied by a free parameter (complementary solution) to complete the general solution. For example, if we have  $3x - 2y = 5$ , then by the same method, we get

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{5}{3} + \frac{2}{3}t \\ t \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ 0 \end{bmatrix} + t \begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix}$$

Bear in mind that the direction vector (representing complementary solution) can be scaled freely. In addition, any initial position vector (particular solution) can be chosen as long as it links to a point on the line and satisfies the equation. (You may refer to the discussion about particular/complementary solution in Section 3.2.1.) Hence there is no unique vector form for a line. For instance,

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix} + t_1 \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

is equivalent to

$$\begin{bmatrix} -1 \\ -1 \end{bmatrix} + t_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

for the line equation  $2x - y = -1$ .

**Short Exercise:** Check the equivalence of the two vector forms above by choosing a value for  $t_1$  and finding the corresponding  $t_2$  so that the vector points to the same position.<sup>2</sup>

---

<sup>2</sup>For example, if  $t_1 = 1$ , we have  $(1, 3)^T + (1)(2, 4)^T = (3, 7)^T$  as a point on the line, and for the another vector form  $(-1, -1)^T + t_2(1, 2)^T = (3, 7)^T$  to coincide we will have  $t_2 = 4$ . In this case, it can be shown that the general relation between the two forms is determined by  $t_2 = 2t_1 + 2$ , as

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix} + t_1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \left( \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right) + 2t_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} + (2t_1 + 2) \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Short Exercise: What is the vector form of the equation  $ax + by = h$  for the degenerate case  $a = 0$ ?<sup>3</sup>

### 5.1.2 Recovering Equation Form from Vector Form

On the other hand, inferring line equation from the vector form is not straightforward at first sight. Since the vector form of a line always contains an arbitrary parameter, which is absent in the equation form, the motivation is to remove the parameter through some manipulation.

Remember that from Properties 4.2.5 the dot product between orthogonal (perpendicular) vectors returns zero. This means that by carrying out dot product with the **normal vector** of the line which is orthogonal to the direction vector, on both sides of the vector form will eliminate the parameter and recover the line equation. For example, given that

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

We know that  $(4, -1)^T$  is a normal vector orthogonal to the direction vector (see the next short exercise). So, by taking dot product with  $(4, -1)^T$  on both sides, we have

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} &= \left( \begin{bmatrix} 1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ 4 \end{bmatrix} \right) \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} + t \begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} \\ 4x - y &= ((1)(4) + (3)(-1)) + ((1)(4) + (4)(-1))t = 1 + 0t = 1 \\ \Rightarrow 4x - y &= 1 \end{aligned}$$

Notice that the coefficients of the equation are the same as the components of the normal vector.

---

<sup>3</sup>The equation is reduced to  $y = \frac{h}{b}$  and we select  $x = t$  as the free variable instead.

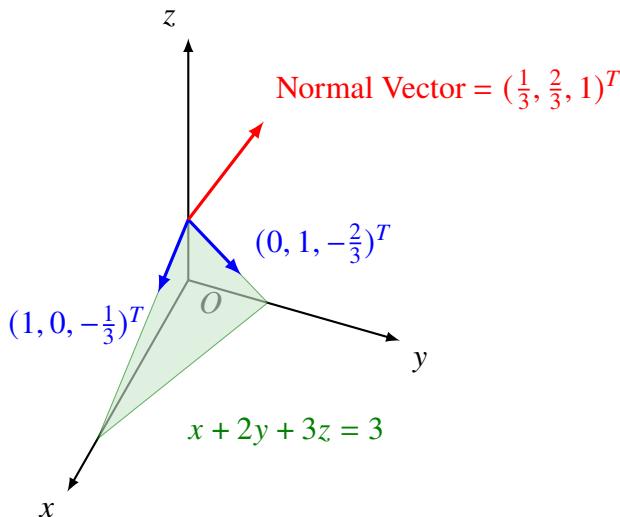
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t \\ \frac{h}{b} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{h}{b} \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Short Exercise: Verify that  $(a, b)$  is always orthogonal to  $(b, -a)$ , and vice versa.<sup>4</sup>

### 5.1.3 Generalizing to Higher Dimensions

Similar concepts can be applied on the equation and vector form for planes. General form of equation of a plane in three-dimensional space is  $ax+by+cz = h$ , which is a linear system of one equation with three unknowns. From the analysis in Section 3.2.1 we know there are two free variables and two (non-parallel) direction vectors for such a plane. By assigning any two (non-pivotal) unknowns as the free variables, we then obtain the vector form of the plane.

Recall from Section 4.2.2, the cross product of any two non-parallel vectors on the plane will give a third vector normal to the plane. Subsequently, we can take the dot product with this newly obtained normal vector to convert the vector form back to a plane equation just like what we have done for lines in the last subsection. Again, the coefficients of the plane equation match the components of the normal vector, differed at most by a multiplicative factor.




---

<sup>4</sup> $(a, b)^T \cdot (b, -a)^T = (a)(b) + (b)(-a) = 0$

The plane represented by the equation  $x + 2y + 3z = 3$ . Notice that the normal vector can be found via computing  $(1, 0, -\frac{1}{3})^T \times (0, 1, -\frac{2}{3})^T = (\frac{1}{3}, \frac{2}{3}, 1)^T$ . The normal vector is magnified for the purpose of illustration.

**Example 5.1.1.** Transform the plane equation  $2x + 3y + z = 4$  to vector form and convert the acquired vector form back to the starting equation to check consistency.

*Solution.* For the first part, we can let  $y = s$ ,  $z = t$ , then from the plane equation we have  $x = \frac{1}{2}(4 - 3s - t)$  and hence

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(4 - 3s - t) \\ s \\ t \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix}$$

where  $-\infty < s < \infty$ ,  $-\infty < t < \infty$  are the two free parameters. To recover the original equation, we can find the normal vector by doing cross product on the two direction vectors obtained above. By Properties 4.2.9, we can acquire a normal vector of

$$\begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ -\frac{3}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{vmatrix} = \hat{i} + \frac{3}{2}\hat{j} + \frac{1}{2}\hat{k}$$

The next step is to take the dot product on both sides of the vector equation with the normal vector just retrieved.

$$\begin{aligned} \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} \\ \begin{bmatrix} x \\ y \\ z \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} &= \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} + s \begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} \\ x + \frac{3}{2}y + \frac{1}{2}z &= 2 + s(0) + t(0) = 2 \\ \rightarrow 2x + 3y + z &= 4 \end{aligned}$$

□

The correspondence between the coefficients of a linear equation and components of its normal vector is not a coincidence. In fact, even for higher dimensional cases, where there is no intuitive geometric interpretation, it is still true.

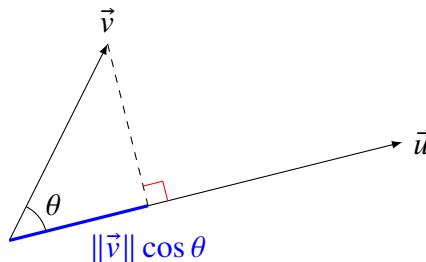
**Properties 5.1.1.** An equation of the form  $a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n = h$  in  $\mathbb{R}^n$  has a normal vector of  $(a_1, a_2, a_3, \dots, a_n)^T$ .

The procedures carried in the last example can be similarly applied to higher dimensional situations where the equation now represents a **hyperplane**<sup>5</sup>.

## 5.2 Further Geometric Applications of Dot Product

### 5.2.1 Projection

We have mentioned in Properties 4.2.4 that dot product between two vectors is related to the projection of one vector onto another. By rearranging the formula of Properties 4.2.4, we can derive the length of projection as follows.



**Properties 5.2.1.** For two real vectors  $\vec{u}$  and  $\vec{v}$  having the same dimension, the

---

<sup>5</sup>A hyperplane in the  $n$ -dimensional real space  $\mathbb{R}^n$  can be think of as an " $(n - 1)$ -dimensional flat surface".

(signed) scalar projection of  $\vec{v}$  onto  $\vec{u}$  is computed according to

$$\text{proj}_u v = \|\vec{v}\| \cos \theta = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|}$$

If we want to give directionality to the projection, then we can supply its unit vector  $\hat{u}$  to make it a *vector projection*:

$$\begin{aligned}\overrightarrow{\text{proj}}_u v &= (\text{proj}_u v) \hat{u} = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|} \hat{u} = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|^2} \vec{u} \\ &= (\text{proj}_u v) \frac{\vec{u}}{\|\vec{u}\|}\end{aligned}$$

where we have used Definition 4.1.5 to write out the unit vector  $\hat{u}$  as  $\frac{\vec{u}}{\|\vec{u}\|}$ .

**Example 5.2.1.** Find the projection of  $\vec{v} = -2\hat{i} + 3\hat{j} - \hat{k}$  onto  $\vec{u} = 4\hat{i} + \hat{j} - 3\hat{k}$ .

*Solution.* According to Properties 5.2.1, The signed scalar projection of  $\vec{v}$  into  $\vec{u}$  is

$$\begin{aligned}\text{proj}_u v &= \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|} \\ &= \frac{(-2)(4) + (3)(1) + (-1)(-3)}{\sqrt{(4)^2 + (1)^2 + (-3)^2}} \\ &= -\frac{2}{\sqrt{26}} = -\frac{\sqrt{26}}{13}\end{aligned}$$

and the vector projection is

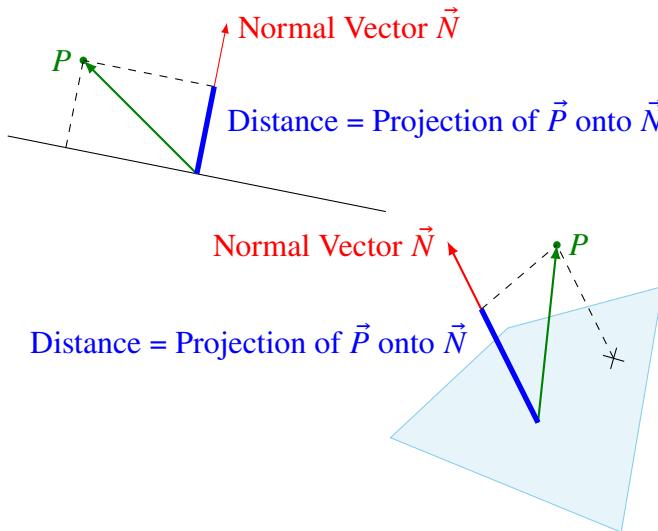
$$\begin{aligned}\overrightarrow{\text{proj}}_u v &= (\text{proj}_u v) \frac{\vec{u}}{\|\vec{u}\|} \\ &= \left(-\frac{\sqrt{26}}{13}\right) \left(\frac{4\hat{i} + \hat{j} - 3\hat{k}}{\sqrt{26}}\right)\end{aligned}$$

$$= -\frac{1}{13}(4\hat{i} + \hat{j} - 3\hat{k}) = \left(-\frac{4}{13}, -\frac{1}{13}, \frac{3}{13}\right)^T$$

□

## 5.2.2 Distance

Distance of a point to a line/plane (in  $\mathbb{R}^2/\mathbb{R}^3$  respectively) can be found by projecting any vector starting somewhere from the line/plane to the point, onto the normal vector of that line/plane, as illustrated in the figures below.



**Example 5.2.2.** Find the distance from the plane  $x - 2y + 3z = 6$  to the point  $(3, 3, 6)^T$ .

*Solution.* From the equation of the plane, and by Properties 5.1.1, it can be inferred that the normal vector of the plane is  $\hat{i} - 2\hat{j} + 3\hat{k}$ . We can select any point on the plane as we wish, let's say  $(4, 2, 2)^T$ , and the vector from such a point to the point  $(3, 3, 6)^T$  in question is simply their difference  $(3, 3, 6)^T - (4, 2, 2)^T =$

$-\hat{i} + \hat{j} + 4\hat{k}$ . Subsequently, the distance is found from the length of the projection of this vector  $-\hat{i} + \hat{j} + 4\hat{k}$  onto the normal vector of the plane  $\hat{i} - 2\hat{j} + 3\hat{k}$ . By Properties 5.2.1, it is

$$\frac{(-\hat{i} + \hat{j} + 4\hat{k}) \cdot (\hat{i} - 2\hat{j} + 3\hat{k})}{\|\hat{i} - 2\hat{j} + 3\hat{k}\|} = \frac{(-1)(1) + (1)(-2) + (4)(3)}{\sqrt{(1)^2 + (-2)^2 + (3)^2}} = \frac{9}{\sqrt{14}}$$

□

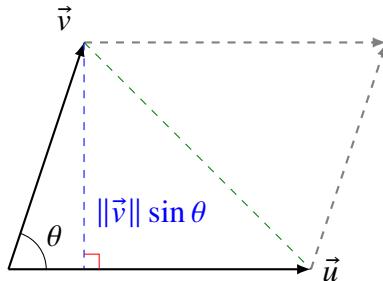
Sometimes the calculation may lead to a negative value for the projection and we may want to take the absolute value. The case of finding the distance of a point to a line of  $\mathbb{R}^3$  is considered in Exercise 5.3.

## 5.3 Further Geometric Applications of Cross Product

Unless specified, all vectors in this section is assumed to be of  $\mathbb{R}^3$ .

### 5.3.1 Area

The area of the parallelogram formed by two vectors  $\vec{u}, \vec{v}$  are simply the absolute value of their cross product.



**Properties 5.3.1.** Directly from Properties 4.2.10, the area of the parallelogram formed by two vectors  $\vec{u}$  and  $\vec{v}$  is

$$\|\vec{u} \times \vec{v}\| = \|\vec{u}\| \|\vec{v}\| \sin \theta$$

Similarly, the area of triangle made by  $\vec{u}$  and  $\vec{v}$  is half of the above:

$$\frac{1}{2} \|\vec{u} \times \vec{v}\| = \frac{1}{2} \|\vec{u}\| \|\vec{v}\| \sin \theta$$

**Example 5.3.1.** Find the area of the parallelogram formed by  $\vec{u} = (-1, -2, 4)^T$  and  $\vec{v} = (3, 0, 1)^T$ .

*Solution.* By Properties 4.2.9, the cross product between the two given vectors is

$$\begin{aligned}\vec{u} \times \vec{v} &= \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ -1 & -2 & 4 \\ 3 & 0 & 1 \end{vmatrix} \\ &= -2\hat{i} + 13\hat{j} + 6\hat{k}\end{aligned}$$

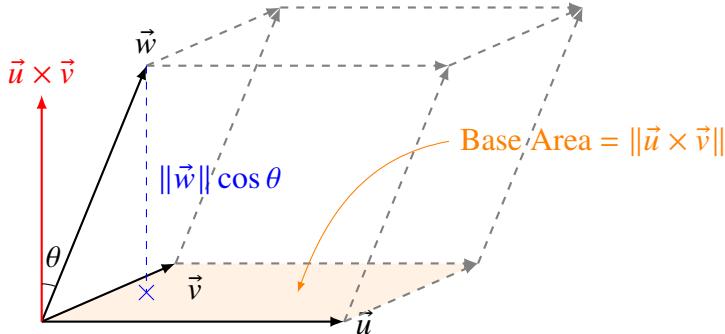
Therefore, as suggested by Properties 5.3.1, the required area is

$$\begin{aligned}\|\vec{u} \times \vec{v}\| &= \sqrt{(-2)^2 + (13)^2 + (6)^2} \\ &= \sqrt{209}\end{aligned}$$

□

## 5.3.2 Volume

Meanwhile, volume of parallelepiped (see the figure below) formed by three vectors  $\vec{u}, \vec{v}, \vec{w}$  is given by the absolute value of the so-called **scalar triple product** as follows.



**Properties 5.3.2 (Scalar Triple Product).** The volume of parallelepiped formed by three vectors  $\vec{u}$ ,  $\vec{v}$ , and  $\vec{w}$  is calculated as

$$\|\vec{u} \times \vec{v}\| \|\vec{w}\| \cos \theta = |(\vec{u} \times \vec{v}) \cdot \vec{w}| = \text{abs} \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

where

$$(\vec{u} \times \vec{v}) \cdot \vec{w} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

is the scalar triple product of  $\vec{u}$ ,  $\vec{v}$ , and  $\vec{w}$ . Also, by applying Properties 2.3.5, the determinant form of scalar triple product indicates that

$$\begin{aligned} (\vec{u} \times \vec{v}) \cdot \vec{w} &= (\vec{v} \times \vec{w}) \cdot \vec{u} = (\vec{w} \times \vec{u}) \cdot \vec{v} \\ &= -(\vec{v} \times \vec{u}) \cdot \vec{w} = -(\vec{w} \times \vec{v}) \cdot \vec{u} = -(\vec{u} \times \vec{w}) \cdot \vec{v} \end{aligned}$$

*Proof.* We will prove the determinant formula shown above for  $(\vec{u} \times \vec{v}) \cdot \vec{w}$  briefly. By Properties 4.2.9, we have

$$\vec{u} \times \vec{v} = (u_2 v_3 - u_3 v_2) \hat{i} + (u_3 v_1 - u_1 v_3) \hat{j} + (u_1 v_2 - u_2 v_1) \hat{k}$$

and then according to Definition 4.2.1

$$(\vec{u} \times \vec{v}) \cdot \vec{w} = (u_2 v_3 - u_3 v_2, u_3 v_1 - u_1 v_3, u_1 v_2 - u_2 v_1)^T \cdot (w_1, w_2, w_3)^T$$

### 5.3 Further Geometric Applications of Cross Product

---

$$= (u_2v_3 - u_3v_2)(w_1) + (u_3v_1 - u_1v_3)(w_2) + (u_1v_2 - u_2v_1)(w_3)$$

which is equal to

$$\begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = w_1(u_2v_3 - u_3v_2) - w_2(u_1v_3 - u_3v_1) + w_3(u_1v_2 - u_2v_1)$$

where we do cofactor expansion (Properties 2.3.3) along the third row of the determinant.  $\square$

If the volume of parallelepiped evaluated from the scalar triple product is zero, it implies that the three vectors involved are **co-planar**, i.e. lying on the same plane.

**Properties 5.3.3.** Given three vectors  $\vec{u}$ ,  $\vec{v}$ , and  $\vec{w}$ , if their scalar triple product  $(\vec{u} \times \vec{v}) \cdot \vec{w} = 0$  equals to zero, then  $\vec{u}$ ,  $\vec{v}$ , and  $\vec{w}$  are co-planar and lie on the same plane, and vice versa.

Note that if  $\vec{w} = \alpha\vec{u} + \beta\vec{v}$ , where  $\alpha$  and  $\beta$  are some scalars, then  $\vec{u}$ ,  $\vec{v}$ ,  $\vec{w}$  are co-planar, and  $(\vec{u} \times \vec{v}) \cdot \vec{w} = (\vec{u} \times \vec{v}) \cdot (\alpha\vec{u} + \beta\vec{v}) = \alpha((\vec{u} \times \vec{v}) \cdot \vec{u}) + \beta((\vec{u} \times \vec{v}) \cdot \vec{v}) = \alpha(0) + \beta(0) = 0$  as both  $\vec{u} \cdot (\vec{u} \times \vec{v})$  and  $\vec{v} \cdot (\vec{u} \times \vec{v})$  equal to zero.

**Example 5.3.2.** Find the volume of the parallelepiped formed by  $\vec{u} = (1, -2, 2)^T$ ,  $\vec{v} = (-1, -1, 1)^T$  and  $\vec{w} = (2, 1, 0)^T$ .

*Solution.* By Properties 5.3.2, the triple scalar product of the three given vectors is

$$(\vec{u} \times \vec{v}) \cdot \vec{w} = \begin{vmatrix} 1 & -2 & 2 \\ -1 & -1 & 1 \\ 2 & 1 & 0 \end{vmatrix} = -3$$

and the volume is  $|-3| = 3$ .  $\square$

## Generalization to other dimensions

Given that the volume of parallelepiped formed by three vectors is equal to the absolute value of the corresponding matrix determinant as derived above, it is natural to ask if similar results hold for other numbers of dimension. In fact, Properties 5.3.2 can be generalized to include length, area and the so-called *n-volume* (Volume equivalent of *n* vectors in the *n*-dimensional space).

**Properties 5.3.4.** For *n* vectors of  $\mathbb{R}^n$ , their *n*-volume is the absolute value of the determinant of matrix formed by these column (or row) vectors. When *n* = 1, 2, 3, the *n*-volume corresponds to the usual notions of length, area and volume.

We can check the legitimacy of the last sentence in Properties 5.3.4 by noticing it is consistent with Properties 5.3.1 about area of two vectors on the *x*-*y* plane. Given  $\vec{u} = (u_1, u_2)^T$  and  $\vec{v} = (v_1, v_2)^T$ , by Properties 5.3.4 the area of the parallelogram formed by them is

$$\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} = u_1 v_2 - v_1 u_2$$

Alternatively, we can treat  $\vec{u}$  and  $\vec{v}$  as two three-dimensional vectors  $(u_1, u_2, 0)^T$  and  $(v_1, v_2, 0)^T$  such that they have a zero *z*-component and remain lying on the *x*-*y* plane. Then according to the previous Properties 5.3.1, the area is computed by  $\|\vec{u} \times \vec{v}\|$ , where by Properties 4.2.9,

$$\begin{aligned} \vec{u} \times \vec{v} &= \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_1 & u_2 & 0 \\ v_1 & v_2 & 0 \end{vmatrix} \\ &= \hat{k} \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \text{ (Cofactor Expansion along the third row)} \\ &= (u_1 v_2 - v_1 u_2) \hat{k} = (0, 0, u_1 v_2 - v_1 u_2)^T \end{aligned}$$

Hence  $\|\vec{u} \times \vec{v}\| = \sqrt{(0)^2 + (0)^2 + (u_1 v_2 - v_1 u_2)^2} = u_1 v_2 - v_1 u_2$ , which coincides with the expression we just derived from Properties 5.3.4.

## Remarks

The solution of a linear system can be considered as a point/line/plane/hyperplane too, depending on the number of free variables and thus direction vectors in the complementary part (0/1/2 or more). We may also like to call it a *solution space*. However, while such shapes surely occupy space geometrically, we have been shying away from defining what really means by a *vector space* mathematically, which will be the main point of discussion in the next chapter.

## 5.4 Useful Vector Identities

In this section, we will prove some key vector identities that may be of utilities to some readers.

**Properties 5.4.1** (Vector Triple Product). The *vector triple product* of three vectors  $\vec{u}, \vec{v}, \vec{w}$  is defined as

$$\vec{u} \times (\vec{v} \times \vec{w}) = (\vec{u} \cdot \vec{w})\vec{v} - (\vec{u} \cdot \vec{v})\vec{w}$$

*Proof.* By Properties 4.2.9, the L.H.S. can be expanded into

$$\begin{aligned} & \vec{u} \times (\vec{v} \times \vec{w}) \\ &= (u_1 \hat{i} + u_2 \hat{j} + u_3 \hat{k}) \\ & \quad \times [(v_2 w_3 - v_3 w_2) \hat{i} + (v_3 w_1 - v_1 w_3) \hat{j} + (v_1 w_2 - v_2 w_1) \hat{k}] \\ &= \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_2 w_3 - v_3 w_2 & v_3 w_1 - v_1 w_3 & v_1 w_2 - v_2 w_1 \end{vmatrix} \end{aligned}$$

The  $\hat{i}$  component along the  $x$ -direction is

$$\begin{aligned} & u_2(v_1 w_2 - v_2 w_1) - u_3(v_3 w_1 - v_1 w_3) \\ &= u_2 w_2 v_1 + u_3 w_3 v_1 - u_2 v_2 w_1 - u_3 v_3 w_1 \end{aligned}$$

$$\begin{aligned}
 &= u_1 w_1 v_1 + u_2 w_2 v_1 + u_3 w_3 v_1 - u_1 v_1 w_1 - u_2 v_2 w_1 - u_3 v_3 w_1 \\
 &= (u_1 w_1 + u_2 w_2 + u_3 w_3) v_1 - (u_1 v_1 + u_2 v_2 + u_3 v_3) w_1 \\
 &= (\vec{u} \cdot \vec{w}) v_1 - (\vec{u} \cdot \vec{v}) w_1
 \end{aligned}$$

which is equal to the  $\hat{i}$  component on the R.H.S. and similar results can be shown for the  $\hat{j}$ ,  $\hat{k}$  components and the equality establishes.  $\square$

**Properties 5.4.2** (Jacobi Identity).

$$\vec{u} \times (\vec{v} \times \vec{w}) + \vec{v} \times (\vec{w} \times \vec{u}) + \vec{w} \times (\vec{u} \times \vec{v}) = \mathbf{0}$$

*Proof.* By Properties 5.4.1, we have

$$\begin{aligned}
 &\vec{u} \times (\vec{v} \times \vec{w}) + \vec{v} \times (\vec{w} \times \vec{u}) + \vec{w} \times (\vec{u} \times \vec{v}) \\
 &= [(\vec{u} \cdot \vec{w})\vec{v} - (\vec{u} \cdot \vec{v})\vec{w}] \\
 &\quad + [(\vec{v} \cdot \vec{u})\vec{w} - (\vec{v} \cdot \vec{w})\vec{u}] \\
 &\quad + [(\vec{w} \cdot \vec{v})\vec{u} - (\vec{w} \cdot \vec{u})\vec{v}] \\
 &= [(\vec{u} \cdot \vec{w})\vec{v} - (\vec{w} \cdot \vec{u})\vec{v}] \\
 &\quad + [(\vec{v} \cdot \vec{u})\vec{w} - (\vec{u} \cdot \vec{v})\vec{w}] \\
 &\quad + [(\vec{w} \cdot \vec{v})\vec{u} - (\vec{v} \cdot \vec{w})\vec{u}] \\
 &= 0\vec{v} + 0\vec{w} + 0\vec{u} = \mathbf{0}
 \end{aligned}$$

$\square$

**Properties 5.4.3** (Lagrange's Identity).

$$\|\vec{u} \times \vec{v}\|^2 = \|\vec{u}\|^2 \|\vec{v}\|^2 - (\vec{u} \cdot \vec{v})^2$$

*Proof.* Manipulating the geometric formulae of dot/cross product, we have

$$\begin{aligned}
 \|\vec{u} \times \vec{v}\|^2 &= \|\vec{u}\|^2 \|\vec{v}\|^2 \sin^2 \theta && \text{(Properties 4.2.10)} \\
 &= \|\vec{u}\|^2 \|\vec{v}\|^2 (1 - \cos^2 \theta)
 \end{aligned}$$

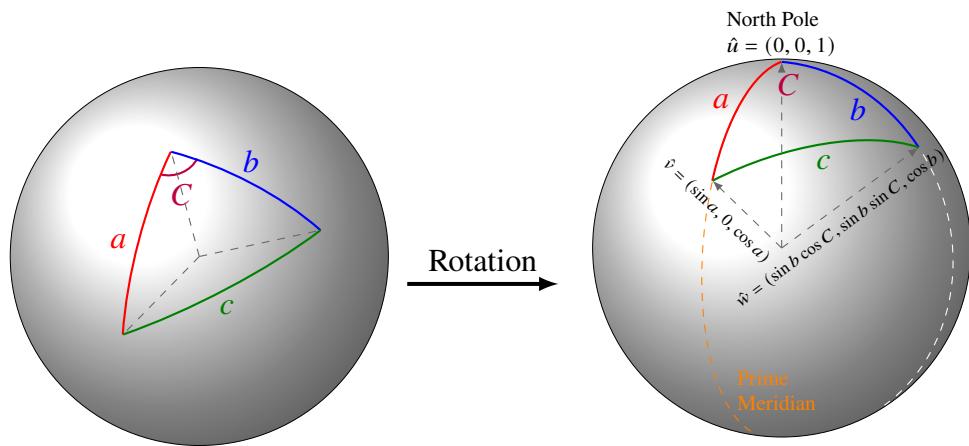


Figure 5.1: The spherical triangle on a unit sphere as described in Properties 5.4.4.

$$\begin{aligned}
 &= \|\vec{u}\|^2 \|\vec{v}\|^2 - \|\vec{u}\|^2 \|\vec{v}\|^2 \cos^2 \theta \\
 &= \|\vec{u}\|^2 \|\vec{v}\|^2 - (\vec{u} \cdot \vec{v})^2
 \end{aligned} \tag{Properties 4.2.4}$$

□

The last identity is the **Cosine Law for Spherical Trigonometry**.

**Properties 5.4.4** (Cosine Law for Spherical Trigonometry).

$$\cos c = \cos a \cos b + \sin a \sin b \cos C$$

where  $a, b, c$  are the (subtended angle of) three arcs (in radians) of a spherical triangle on a unit sphere and  $C$  is the angle between the two arcs  $a$  and  $b$ , as shown in Figure 5.1.

*Proof.* For the given spherical triangle, we can always rotate the coordinate system (see Figure 5.1) while keeping its shape intact, such that the corner  $C$  is positioned exactly at the north pole ( $\hat{u} = (0, 0, 1)^T$ ) and one of the two arcs starting from corner  $C$  (let's say  $a$ ) lies along the Prime Meridian (angle from the

$x$ -axis is  $0^\circ$ , i.e.  $y = 0$ ). The vector  $\hat{v}$  at the end of arc  $a$  will then have a direction of  $(\sin a, 0, \cos a)^T$ . The vector  $\hat{w}$  to the remaining corner at the intersection of arcs  $b$  and  $c$  will similarly have a  $z$ -component of  $\cos b$ , and its projection on  $x$ - $y$  plane will be  $\sin b$  and the  $x$ / $y$ -component will then be  $\sin b \cos C$  and  $\sin b \sin C$ , i.e.  $\hat{w} = (\sin b \cos C, \sin b \sin C, \cos b)^T$ . Now consider the dot product  $\hat{v} \cdot \hat{w}$ . The geometric meaning of dot product (Properties 4.2.4) implies that it is the angle between  $\hat{v}$  and  $\hat{w}$ , that is,  $\hat{v} \cdot \hat{w} = \cos c$ . On the other hand,

$$\begin{aligned}\hat{v} \cdot \hat{w} &= (\sin a, 0, \cos a)^T \cdot (\sin b \cos C, \sin b \sin C, \cos b)^T \\ &= (\sin a)(\sin b \cos C) + (0)(\sin b \sin C) + (\cos a)(\cos b) \\ &= \cos a \cos b + \sin a \sin b \cos C\end{aligned}$$

Therefore, equaling the two expressions of  $\hat{v} \cdot \hat{w}$  gives the desired formula of  $\cos c = \cos a \cos b + \sin a \sin b \cos C$ .  $\square$

## 5.5 Earth Science Applications

**Example 5.5.1.** Derive the *Haversine Formula* for finding the great-circle distance between any two points on a sphere with their latitudes/longitudes provided. Hence find the distance between New York ( $40.73^\circ\text{N}$ ,  $73.94^\circ\text{W}$ ) and Warsaw ( $52.24^\circ\text{N}$ ,  $21.02^\circ\text{E}$ ).

*Solution.* Denote the latitudes/longitudes of the two locations by  $\varphi_{1,2}$  and  $\lambda_{1,2}$ . Starting from the Cosine Law for Spherical Trigonometry (Properties 5.4.4) with corner  $C$  still fixed at north pole but arc  $a$  not necessarily along the Prime Meridian, we have  $C = \lambda_2 - \lambda_1$ ,  $a = \frac{\pi}{2} - \varphi_1$ ,  $b = \frac{\pi}{2} - \varphi_2$ , and

$$\cos c = \cos a \cos b + \sin a \sin b \cos C$$

$$\cos c = \cos\left(\frac{\pi}{2} - \varphi_1\right) \cos\left(\frac{\pi}{2} - \varphi_2\right) + \sin\left(\frac{\pi}{2} - \varphi_1\right) \sin\left(\frac{\pi}{2} - \varphi_2\right) \cos(\lambda_2 - \lambda_1)$$

$$\cos c = \sin \varphi_1 \sin \varphi_2 + \cos \varphi_1 \cos \varphi_2 \cos(\lambda_2 - \lambda_1)$$

The *haversine* of an angle  $\theta$  is  $\text{hav}(\theta) = \sin^2(\frac{\theta}{2}) = \frac{1}{2}(1 - \cos \theta)$  and therefore  $\cos \theta = 1 - 2 \text{hav}(\theta)$ . Subsequently,

$$\begin{aligned}\cos c &= \sin \varphi_1 \sin \varphi_2 + \cos \varphi_1 \cos \varphi_2 (1 - 2 \text{hav}(\lambda_2 - \lambda_1)) \\ \cos c &= \sin \varphi_1 \sin \varphi_2 + \cos \varphi_1 \cos \varphi_2 - 2 \cos \varphi_1 \cos \varphi_2 \text{hav}(\lambda_2 - \lambda_1) \\ \cos c &= \cos(\varphi_2 - \varphi_1) - 2 \cos \varphi_1 \cos \varphi_2 \text{hav}(\lambda_2 - \lambda_1) \\ (1 - 2 \text{hav}(c)) &= (1 - 2 \text{hav}(\varphi_2 - \varphi_1)) - 2 \cos \varphi_1 \cos \varphi_2 \text{hav}(\lambda_2 - \lambda_1) \\ \text{hav}(c) &= \text{hav}(\varphi_2 - \varphi_1) + \cos \varphi_1 \cos \varphi_2 \text{hav}(\lambda_2 - \lambda_1)\end{aligned}$$

where we have used the trigonometric identity  $\cos(\theta - \phi) = \cos \theta \cos \phi + \sin \theta \sin \phi$  in the middle. The Haversine Formula is now established and we can use it to calculate the angle  $c$  subtended by the arc between two locations and hence their distance by  $d = rc$  where  $r$  is the radius (of the Earth, 6370 km). For New York ( $40.73^\circ\text{N}$ ,  $73.94^\circ\text{W}$ ) and Warsaw ( $52.24^\circ\text{N}$ ,  $21.02^\circ\text{E}$ ),  $\lambda_1 = -73.94^\circ$ ,  $\lambda_2 = 21.02^\circ$ ,  $\varphi_1 = 40.73^\circ$ ,  $\varphi_2 = 52.24^\circ$ , and

$$\begin{aligned}\text{hav}(c) &= \text{hav}(52.24^\circ - 40.73^\circ) \\ &\quad + \cos(40.73^\circ) \cos(52.24^\circ) \text{hav}(21.02^\circ - (-73.94^\circ)) \\ &= \text{hav}(11.51^\circ) + \cos(40.73^\circ) \cos(52.24^\circ) \text{hav}(94.96^\circ) \\ &= \sin^2\left(\frac{11.51^\circ}{2}\right) + \cos(40.73^\circ) \cos(52.24^\circ) \sin^2\left(\frac{94.96^\circ}{2}\right) \\ \sin^2\left(\frac{c}{2}\right) &\approx 0.26214 \\ c &\approx 61.6^\circ = 1.075 \text{ rad}\end{aligned}$$

and therefore the required distance is  $d = rc = (6370 \text{ km})(1.075 \text{ rad}) \approx 6848 \text{ km}$ . The value computed by the Haversine Formula will be slightly off from the true value since the Earth is not a perfect sphere but rather an oblate one.  $\square$

**Example 5.5.2.** The Earth's magnetic field can be approximated by a magnetic dipole, so that the magnetic field lines on the Earth's surface are oriented from the geomagnetic North Pole to geomagnetic South Pole (like longitudinal lines but for the geomagnetic dipole). In 2020, the geomagnetic North Pole is at

80.7 °N, 72.7 °W. Find the magnetic declination (angle from the geographic North to geomagnetic North) at Tokyo (35.65 °N, 139.84 °E) according to this *geomagnetic dipole model*.

*Solution.* To find the magnetic declination we need to calculate the three arcs of the spherical triangle with its three corners at the geographic/geomagnetic North Pole and Tokyo. The arc distance between geographic/geomagnetic North Pole  $d$  is simply  $90^\circ - 80.7^\circ = 9.3^\circ$ . Similarly, the arc from the geographic North Pole to Tokyo is  $a = 90^\circ - 35.65^\circ = 54.35^\circ$ . We can use the Haversine Formula derived in the last example to obtain the arc from the geomagnetic North Pole to Tokyo, which yields

$$\begin{aligned}\text{hav}(t) &= \text{hav}(80.7^\circ - 35.65^\circ) \\ &\quad + \cos(35.65^\circ) \cos(80.7^\circ) \text{ hav}((-72.7^\circ) - 139.84^\circ) \\ &= \text{hav}(45.05^\circ) + \cos(35.65^\circ) \cos(80.7^\circ) \text{ hav}(-212.54^\circ) \\ &\approx 0.26777 \\ c &\approx 62.3^\circ\end{aligned}$$

Denote the declination angle by  $D$ . By Properties 5.4.4, we have

$$\begin{aligned}\cos d &= \cos a \cos t + \sin a \sin t \cos D \\ \cos(9.3^\circ) &= \cos(54.35^\circ) \cos(62.3^\circ) + \sin(54.35^\circ) \sin(62.3^\circ) \cos D \\ \cos D &\approx 0.9951 \\ D &\approx \pm 5.7^\circ\end{aligned}$$

To determine the sign, we note that concluded from the longitudes of Tokyo and geomagnetic North, the geomagnetic North is located to the east of Tokyo, and hence  $D = 5.7^\circ$ E. However, we note that the actual declination is 7.8 °W which has an opposite sign and is far from our answer (you can extract the value from <https://www.ngdc.noaa.gov/geomag/calculators/magcalc.shtml>). The reason is that the geomagnetic dipole is only a rough first-order approximation, while in reality the Earth's magnetic field has a much more complex structure.  $\square$

## 5.6 Python Programming

Projection as in Properties 5.2.1 can be calculated by numpy functions and let's wrap them up in our self-defined function as below.

```
def scalar_projection(u, v):
    """ Calculates the scalar projection of v onto u. """
    return np.dot(u,v) / np.linalg.norm(u)
```

This computes the scalar projection of  $\vec{v}$  onto  $\vec{u}$ . Testing with Example 5.2.1 shows

```
u = np.array([4., 1., -3.])
v = np.array([-2., 3., -1.])
print(scalar_projection(u, v))
```

a consistent output of  $-0.39223$ . Incorporating the unit vector function (`unit_vector()`) defined in the last chapter's programming section, we obtain the vector projection.

```
def vector_projection(u, v):
    """ Calculates the vector projection of v onto u. """
    return scalar_projection(u, v) * unit_vector(u)

print(vector_projection(u, v))
```

This results in  $[-0.3077 \ -0.0769 \ 0.2308]$  which matches the example's answer. Area of parallelogram formed by two vectors is the magnitude of their cross product and the corresponding function is typed below.

```
def area_parallelogram(u, v):
    """ Calculate the area of parallelogram formed by two
    vectors u and v. """
    return np.linalg.norm(np.cross(u,v))
```

`print(area_parallelogram(u, v))` then gives  $18.974$ . Meanwhile, the function to compute volume of parallelepiped made up of three vectors can be defined such that it uses the determinant formula in Properties 5.3.2.

```
def volume_paralleliped(u, v, w):
    """ Calculate the volume of parallelepiped formed by two
    vectors u, v, w. """
    return np.abs(np.linalg.det(np.c_[u,v,w]))
```

```
w = np.array([1., 2., -3.])
print(volume_paralleliped(u, v, w))
```

(`np.c_[]` is a short hand of combining arrays column by column) produces `14.00000...04` due to numerical round-off error (the true answer would be just 14). Finally, let's conclude this section by defining the Haversine Formula in Example 5.5.1.

```
def Haversine_dist(latlon1, latlon2):
    """ Haversine Formula for computing the great-circle
        distance between two places on the Earth.
        Input: (lat1, lon1), (lat2, lon2) in degrees.
        Output: Great-circle distance in km.
    """
    R_Earth = 6370 # Earth's Radius
    lat1, lon1 = latlon1[0], latlon1[1]
    lat2, lon2 = latlon2[0], latlon2[1]
    # Converting degree to radian
    lat1_rad, lon1_rad, lat2_rad, lon2_rad = np.deg2rad(lat1),
                                              np.deg2rad(lon1), np.deg2rad(lat2), np.deg2rad(lon2)
    # Haversine's Formula
    hav_c = np.sin((lat2_rad-lat1_rad)/2)**2 + np.cos(lat1_rad)
           *np.cos(lat2_rad)*np.sin((lon2_rad-lon1_rad)/2)**2
    arc_c = 2*np.arcsin(np.sqrt(hav_c)) # Inverting to get the
                                         # great-circle arc angle
    return(R_Earth*arc_c) # Arc angle to arc length
```

Using the latitudes and longitudes of New York and Warsaw in Example 5.5.1 for testing, `Haversine_dist((40.73, -73.94), (52.24, 21.02))` outputs `6847.76`.

## 5.7 Exercises

**Exercise 5.1** Parameterize the following equations into vector form.

(a)  $6x + 8y = 9$ ,

- (b)  $x + 9y + 9z = 7$ ,
- (c)  $y = 3$ ,  $-\infty < x < \infty$ , and
- (d)  $2x + z = 9$ ,  $-\infty < y < \infty$ .

**Exercise 5.2** Eliminate the parameters and find the direct equation.

(a)

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 9 \end{bmatrix} + t \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \\ 2 \end{bmatrix} + s \begin{bmatrix} 7 \\ 4 \\ 1 \end{bmatrix} + t \begin{bmatrix} 8 \\ 0 \\ 5 \end{bmatrix}$$

where  $-\infty < s, t < \infty$ .

**Exercise 5.3** Find the distance of the point  $(3, 2, 9)^T$  to the plane  $x + 2y + 5z = 10$ , as well as the distance of the point  $(3, 2, 9)^T$  to the line

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = t \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

where  $-\infty < t < \infty$ .

**Exercise 5.4** Prove that the shortest distance between two lines,  $\vec{u} = \vec{a} + s\hat{l}$  and  $\vec{v} = \vec{b} + t\hat{m}$ , where  $-\infty < s, t < \infty$ ,  $\vec{a}, \vec{b}$  are some arbitrary vectors and  $\hat{l}, \hat{m}$  are some fixed, non-parallel unit vectors representing direction of the two lines, is

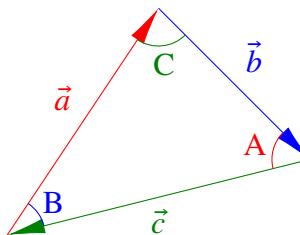
$$\text{Dist}(u, v) = \frac{(\hat{a} - \hat{b}) \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|}$$

Hints: Geometrically, the distance between these two lines is the projection of any vector from one line to another onto the vector normal to the plane made by  $\hat{l}$  and  $\hat{m}$ .

$$\frac{(\vec{v} - \vec{u}) \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|}$$

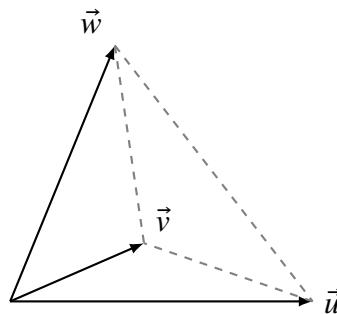
Draw a diagram to convince yourself it is true. What does it imply if  $\vec{a} \cdot (\hat{l} \times \hat{m}) = \vec{b} \cdot (\hat{l} \times \hat{m})$ ?

**Exercise 5.5** Prove Sine Law with vector notation by considering the triangle below



and equating three expressions of its area  $\frac{1}{2} \|\vec{a} \times \vec{b}\| = \frac{1}{2} \|\vec{b} \times \vec{c}\| = \frac{1}{2} \|\vec{c} \times \vec{a}\|$ . Properties 5.3.1 will be useful.

**Exercise 5.6** By extending Properties 5.3.2, derive a vector formula for the volume of a tetrahedron (pyramid).



**Exercise 5.7** For  $\vec{u} = (1, 2, 3)^T$ ,  $\vec{v} = (2, 1, 5)^T$ ,  $\vec{w} = (1, 4, 0)^T$ , find

- Area of the parallelogram formed by  $\vec{u}$  and  $\vec{v}$ ,
- Volume of the parallelepiped formed by  $\vec{u}$ ,  $\vec{v}$  and  $\vec{w}$ ,
- Redo the above for  $\vec{w} = (1, 5, 4)^T$ , what does the result tell you?

**Exercise 5.8** Find the geometric interpretation of solutions of the following systems of linear equations.

(a)

$$\begin{cases} x + 2y + 2z = 3 \\ 3x - y + 3z = 2 \\ x - 2y - z = -1 \end{cases}$$

(b)

$$\begin{cases} 2x - y - z = 3 \\ x + y + 2z = -1 \\ x + 4y + 7z = -6 \end{cases}$$



## Chapter 6

# Vector Spaces and Coordinate Bases

---

The previous chapters have provided a basic understanding of matrices and vectors separately. What bridge these two quantities together are the concepts of *vector (sub)spaces*, *linear combination*, *span*, *linear independence*. With all of these, we can revisit the process of Gaussian Elimination from the view of *column-row factorization*. Then, we will learn how to find *coordinate bases* for vector spaces so as to represent vectors in different coordinate systems. Finally, we are going to investigate about the so-called *four fundamental subspaces* induced by a matrix and see how they are interconnected via the *Rank-Nullity Theorem*.

## 6.1 Making of the Real $n$ -space $\mathbb{R}^n$

### 6.1.1 $\mathbb{R}^n$ as a Vector Space

We have briefly mentioned in Definition 4.1.2 that the real  $n$ -space  $\mathbb{R}^n$  is mathematically a vector space, but without stating the actual requirements. In fact, to be qualified as a *vector space*, a set has to satisfy the ten axioms below. We will limit ourselves to *real vector spaces* for now.

**Definition 6.1.1** (Axioms of a (Real) Vector Space). A *real* vector space is a non-empty set  $\mathcal{V}$  with the zero vector  $\mathbf{0}$ , such that for all elements (vectors)  $\vec{u}, \vec{v}, \vec{w} \in \mathcal{V}$  in the set, and *real* numbers (as the *scalars*)  $a, b \in \mathbb{R}$  (for a complex vector space replace  $\mathbb{R}$  by  $\mathbb{C}$  here), we have

1.  $\vec{u} + \vec{v} \in \mathcal{V}$  (Closure under Vector Addition: Addition between two vectors is defined and the resulting vector is still in the vector space.)
2.  $\vec{u} + \vec{v} = \vec{v} + \vec{u}$  (Commutative Property of Addition)
3.  $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$  (Associative Property of Addition)
4.  $\vec{u} + \mathbf{0} = \mathbf{0} + \vec{u} = \vec{u}$  (Zero Vector as the Additive Identity)
5. For any  $\vec{u}$ , there exists  $\vec{w}$  such that  $\vec{u} + \vec{w} = \mathbf{0}$ . This  $\vec{w}$  is denoted as  $-\vec{u}$ . (Existence of Additive Inverse)
6.  $a\vec{u} \in \mathcal{V}$  (Closure under Scalar Multiplication: Multiplying a vector by any scalar (a real/complex number for a real/complex vector space) is defined and the resulting vector is still in the vector space.)
7.  $a(\vec{u} + \vec{v}) = a\vec{u} + a\vec{v}$  (Distributive Property of Scalar Multiplication)
8.  $(a + b)\vec{u} = a\vec{u} + b\vec{u}$  (Distributive Property of Scalar Multiplication)
9.  $a(b\vec{u}) = (ab)\vec{u}$  (Associative Property of Scalar Multiplication)
10.  $1\vec{u} = \vec{u}$  (The real number 1 as the Multiplicative Identity)

The real  $n$ -space  $\mathbb{R}^n$  satisfies all the axioms above and is finite-dimensional, particularly  $n$ -dimensional (the notion of dimension here should be intuitive, but we will go through it more precisely later), with addition and scalar multiplication being the usual ones as defined in Section 4.1.2, and the zero vector is simply  $\mathbf{0} = (0, 0, 0, \dots, 0)^T$  with  $n$  zeros. We will not do it here but interested readers can try to justify all of them. To build the definition of a vector space from these axioms allows the generalization and application of its utilities to other sets that share the same abstract structure. **However, for most usages, we will focus on**

$\mathbb{R}^{n^1}$ , and the vector space axioms are provided above mainly for reference. We defer the treatment of complex vector spaces to Chapter 8.

### 6.1.2 Subspaces of $\mathbb{R}^n$

It will be very boring if we consider only the whole  $\mathbb{R}^n$  as a vector space. In last chapter, we show that geometrically there can be lower-dimensional shapes like lines/planes/hyperplanes residing in  $\mathbb{R}^n$ . This raises the question if we can similarly find **subspaces** embedded in  $\mathbb{R}^n$  that is a subset of  $\mathbb{R}^n$  which still fulfills the aforementioned vector space axioms such that it is a vector space in its own right. Nevertheless, to determine if a subset of vector space is a subspace, we don't need to check all the ten axioms but rather just two of them.

**Theorem 6.1.2** (Criteria for a Subspace). If  $\mathcal{W}$  is a non-empty subset of a (real) vector space  $\mathcal{V}$  (i.e.  $\mathcal{W} \subseteq \mathcal{V}$ ), then  $\mathcal{W}$  is called a (real) subspace of  $\mathcal{V}$  if the following criteria are satisfied:

1. For any  $\vec{u}, \vec{v} \in \mathcal{W}$ ,  $\vec{u} + \vec{v} \in \mathcal{W}$  (Closed under Addition)
2. For any scalar  $a$  ( $\in \mathbb{R}$ ) and  $\vec{u} \in \mathcal{W}$ ,  $a\vec{u} \in \mathcal{W}$  (Closed under Scalar Multiplication), particularly when  $a = 0$ ,  $0\vec{u} = \mathbf{0} \in \mathcal{W}$  so that a subspace always contains the zero vector of  $\mathcal{V}$ .

These are the same requirements of (1) and (6) in Definition 6.1.1. An equivalent condition is that, for any  $\vec{u}, \vec{v} \in \mathcal{W}$  and two scalars  $a$  and  $b$ ,  $a\vec{u} + b\vec{v} \in \mathcal{W}$ .

**Example 6.1.1.** Consider the following subsets of  $\mathbb{R}^2$  and decide if they are subspaces of  $\mathbb{R}^2$  by verifying the two criteria listed in Theorem 6.1.2.

- (a) The line  $x - 2y = 0$ ,
- (b) The  $y$ -axis,

<sup>1</sup>We actually have a very good reason to do so, as we will see in the next chapter: any  $n$ -dimensional real vector space is *isomorphic* to and can be treated like  $\mathbb{R}^n$ .

- (c) The positive  $y$ -axis,
- (d) The line  $2x + y = 1$ ,
- (e) The parabola  $y = x^2$ ,
- (f) The point  $(-1, 1)^T$ ,
- (g) The first quadrant  $x > 0, y > 0$ ,
- (h) The origin  $\mathbf{0} = (0, 0)^T$ ,
- (i)  $\mathbb{R}^2$  itself.

*Solution.* (a) The vector form of the line is  $\mathcal{W} = \{(x, y)^T = t(2, 1)^T \mid -\infty < t < \infty\}$ . To check the first condition, let's say  $\vec{u} = t_1(2, 1)^T \in \mathcal{W}$  and  $\vec{v} = t_2(2, 1)^T \in \mathcal{W}$  are vectors in  $\mathcal{W}$  for some  $t_1$  and  $t_2$ , then  $\vec{u} + \vec{v} = t_1(2, 1)^T + t_2(2, 1)^T = (t_1 + t_2)(2, 1)^T = s(2, 1)^T \in \mathcal{W}$  where  $s = t_1 + t_2$  also lies on the same straight line of  $x - 2y = 0$  and is another vector in  $\mathcal{W}$ , so  $\mathcal{W}$  is closed under addition. To check the second condition, this time we simply let  $\vec{u} = t(2, 1)^T \in \mathcal{W}$ . Subsequently,  $a\vec{u} = at(2, 1)^T = r(2, 1)^T \in \mathcal{W}$ , for any scalar  $a$  and  $r = at$ , so it is also closed under scalar multiplication. Hence the line  $x - 2y = 0$  is a subspace of  $\mathbb{R}^2$ .

- (b) Same arguments as above but with  $\mathcal{W} = \{(x, y)^T = t(0, 1)^T \mid -\infty < t < \infty\}$ , so the  $y$ -axis is also a subspace of  $\mathbb{R}^2$ .
- (c) For any point on the positive  $y$ -axis, multiplying it by a negative number places it on the negative  $y$ -axis instead, so it is not closed under scalar multiplication and thus not a subspace of  $\mathbb{R}^2$ .
- (d) Denote the collection of points on the line as  $\mathcal{W}$ . Pick  $\vec{u} = (1, -1)^T \in \mathcal{W}$  and  $\vec{v} = (0, 1)^T \in \mathcal{W}$ , then  $\vec{u} + \vec{v} = (1, 0)^T \notin \mathcal{W}$  as  $2(1) + (0) = 2 \neq 1$ , so it is not closed under addition and fails to be a subspace of  $\mathbb{R}^2$ .
- (e) Denote the collection of points on the parabola as  $\mathcal{W}$ . Pick  $\vec{u} = (1, 1)^T \in \mathcal{W}$  and  $\vec{v} = (2, 4)^T \in \mathcal{W}$ , then  $\vec{u} + \vec{v} = (3, 5)^T \notin \mathcal{W}$  is apparently not on

the parabola, so it is not closed under addition and can't be a subspace of  $\mathbb{R}^2$ .

- (f) It is easy to see that it fails to be closed under either addition or scalar multiplication (for example, take  $a(-1, 1)^T$  with  $a \neq 1$ ) and is not a subspace of  $\mathbb{R}^2$ .
- (g) Denote the collection of points on the first quadrant as  $\mathcal{W}$ . Pick  $\vec{u} = (1, 1)^T \in \mathcal{W}$  (or any other point), then multiplying it by  $-1$  will produce  $(-1)\vec{u} = -(1, 1)^T = (-1, -1)^T \notin \mathcal{W}$  which is outside the first quadrant. Therefore, it is not closed under scalar multiplication and hence not a subspace of  $\mathbb{R}^2$ .
- (h) It trivially satisfies the two criteria ( $\mathbf{0}$  is the only element in the set,  $\mathbf{0} + \mathbf{0} = \mathbf{0}$  and  $a\mathbf{0} = \mathbf{0}$  for any scalar  $a$ ) and is a subspace of  $\mathbb{R}^2$ .
- (i)  $\mathbb{R}^2$  is a vector space to begin with and technically a subset of itself (it trivially contains itself) so by definition it is a subspace of  $\mathbb{R}^2$ .

□

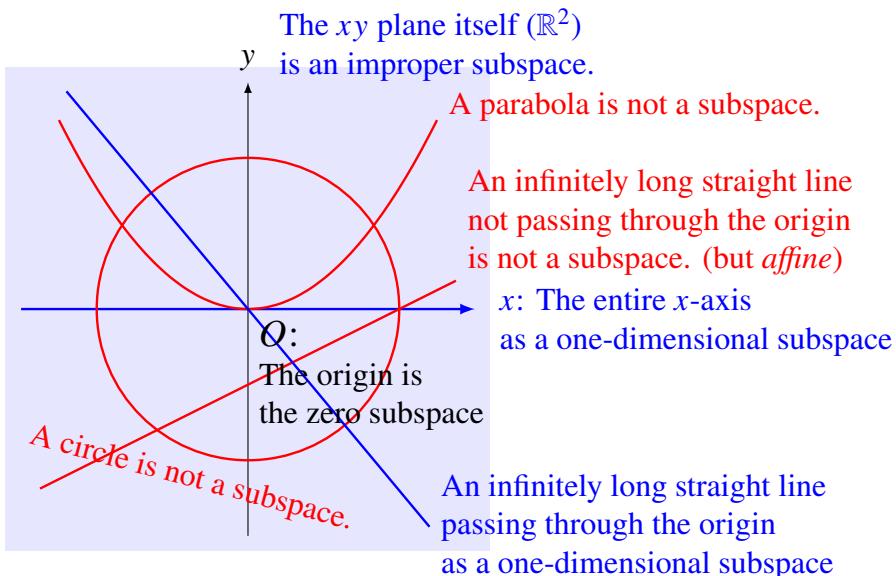
Generalizing the above discussion, we can easily infer that for  $\mathbb{R}^2$ , only the origin (the zero subspace), an infinitely long straight line that passes through the origin, or  $\mathbb{R}^2$  itself can be its subspaces (see the schematic in Figure 6.1). We often use the phrase ***proper subspaces*** to exclude the accommodating vector space itself ( $\mathbb{R}^2$  in this case). For any  $\mathbb{R}^n$ , the ***zero subspace***  $\{\mathbf{0}\}$  and ***improper subspace***  $\mathbb{R}^n$  are always two subspaces of it.

Short Exercise: Determine if the following subsets of  $\mathbb{R}^3$  is a subspace of  $\mathbb{R}^3$ .<sup>2</sup>

- (a) The origin  $\mathbf{0} = (0, 0, 0)^T$ ,
- (b) The point  $(1, 2, 3)^T$ ,
- (c) The line  $(x, y, z)^T = t(-1, 1, 2)^T$  for any scalar  $t$ ,
- (d) The line  $(x, y, z)^T = (1, -1, 3) + t(1, 2, -1)^T$  for any scalar  $t$ ,

---

<sup>2</sup>Yes, No, Yes, No, Yes, No, Yes, No, No. In fact, all possible subspaces of  $\mathbb{R}^3$  are  $\{\mathbf{0}\}$ , any infinitely long line/extending plane through the origin and  $\mathbb{R}^3$  itself.


 Figure 6.1: Some examples (blue) and non-examples (red) of subspaces in  $\mathbb{R}^2$ .

- (e) The plane  $x + 2y - 3z = 0$ ,
- (f) The plane  $x + y + 4z = 5$ ,
- (g)  $\mathbb{R}^3$  itself,
- (h) The sphere  $x^2 + y^2 + z^2 = 1$ ,
- (i) The cone  $x^2 + y^2 = z^2$ .

Further generalization motivated by the short exercise above leads to an intuitive result that, for  $\mathbb{R}^n$ , all its possible subspaces are geometrically "flat shapes" that pass through the origin and extend infinitely. On the other hand, any "curved shape" will not qualify as a subspace. From now on, we assume all vector (sub)spaces mentioned are finite-dimensional (again, we will clarify this notion later) unless otherwise specified.

### 6.1.3 Span by Linear Combinations of Vectors

The last section sees subspaces from a top-down perspective as some subsets of a larger vector space. Here, we are going to take another look at them with a bottom-up perspective, about how to generate a subspace of  $\mathbb{R}^n$  from some of its vectors. To do so, we need to first understand what is a *linear combination* of vectors.

**Definition 6.1.3** (Linear Combination of Vectors). A linear combination of vectors  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)} \in \mathcal{V}$  where  $\mathcal{V}$  is some vector space has the form of

$$\sum_{j=1}^q c_j \vec{v}^{(j)} = c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \dots + c_q \vec{v}^{(q)}$$

where the coefficients  $c_j$  are some scalars (real numbers for a real vector space) and the amount of vectors  $q$  has to be *finite*.

As a small example, if there are two vectors  $\vec{u} = (1, 2)^T$  and  $\vec{v} = (3, 4)^T \in \mathbb{R}^2$ , then  $\vec{h} = (5, 6)^T \in \mathbb{R}^2$  can be written as a linear combination of  $\vec{u}$  and  $\vec{v}$  because  $\vec{h} = (5, 6)^T = -(1, 2)^T + 2(3, 4)^T = -\vec{u} + 2\vec{v}$ .

Short Exercise: If  $\vec{h} = (1, 4)^T$  instead, express it as a linear combination of  $\vec{u}$  and  $\vec{v}$ .<sup>3</sup>

Attentive readers may realize that the short exercise above can be considered as a task to find out the solution (if any) for the system

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

Extending this, to decide whether a vector  $\vec{h} \in \mathbb{R}^n$  can be written as the linear combination of other vectors  $\vec{v}^{(j)} \in \mathbb{R}^n$ ,  $j = 1, 2, \dots, q$ , is equivalent to

---

<sup>3</sup>(1, 4)<sup>T</sup> = 4(1, 2)<sup>T</sup> - (3, 4)<sup>T</sup>.

determining whether the linear system  $A\vec{x} = \vec{h}$  has a solution, where  $A$  equals to (writing out  $\vec{v}^{(j)}$  in a matrix column by column)

$$A = \left[ \begin{array}{c|c|c|c} & & & \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \dots & \vec{v}^{(q)} \\ & & & \end{array} \right]$$

Here, the matrix product  $A\vec{x}$  is a compact way to represent a linear combination of the column vectors that have been condensed into  $A$ .

**Properties 6.1.4.** A linear combination  $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \dots + c_q\vec{v}^{(q)}$  made up of some vectors  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)} \in \mathbb{R}^n$  as in Definition 6.1.3, can be expressed by the matrix product  $A\vec{x}$ , where

$$A = \left[ \begin{array}{c|c|c|c} & & & \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \dots & \vec{v}^{(q)} \\ & & & \end{array} \right] \quad \vec{x} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_q \end{bmatrix}$$

From now on, we will just simply write  $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\dots|\vec{v}^{(q)}]$  and similarly for other matrices formed by column vectors when applicable to save space.

From this perspective, the first/second/last column of a matrix  $A$  can be extracted by

$$A \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad A \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad A \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

and it goes similarly for any other column. Below is a small example to demonstrate the equivalence between matrix-vector products and linear combi-

nations.

$$\begin{aligned}
 & \left[ \begin{array}{cccc|c} 5 & 1 & -1 & 2 & 0 \\ 2 & 3 & 0 & 7 & 1 \\ 4 & -2 & 3 & 1 & 0 \end{array} \right] \left[ \begin{array}{c} 1 \\ 3 \\ -2 \end{array} \right] = \left[ \begin{array}{c} 1 \\ 3 \\ -2 \end{array} \right] \\
 & \left[ \begin{array}{cccc|c} 5 & 1 & -1 & 2 & -1 \\ 2 & 3 & 0 & 7 & 2 \\ 4 & -2 & 3 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] = \left[ \begin{array}{cccc|c} 5 & 1 & -1 & 2 & -1 \\ 2 & 3 & 0 & 7 & 2 \\ 4 & -2 & 3 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \left( \begin{array}{c} -1 \\ 0 \\ 0 \\ 0 \end{array} \right) + \left( \begin{array}{c} 0 \\ 2 \\ 0 \\ 0 \end{array} \right) + \left( \begin{array}{c} 0 \\ 0 \\ 3 \\ 0 \end{array} \right) + \left( \begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \end{array} \right) \\
 & = (-1) \left[ \begin{array}{c} 5 \\ 2 \\ -4 \end{array} \right] + (2) \left[ \begin{array}{c} 1 \\ 3 \\ -2 \end{array} \right] + (3) \left[ \begin{array}{c} -1 \\ 0 \\ 3 \end{array} \right] + 0 \left[ \begin{array}{c} 2 \\ 7 \\ 1 \end{array} \right] \\
 & = \left[ \begin{array}{c} -6 \\ 4 \\ 1 \end{array} \right]
 \end{aligned}$$

**Example 6.1.2.** Show that  $\vec{h} = (2, 4, 3)^T$  cannot be written as a linear combination of  $\vec{v}^{(1)} = (-1, 0, 1)^T$  and  $\vec{v}^{(2)} = (1, 1, 0)^T$ .

*Solution.* Following the above discussion, the objective is equivalent to showing that the linear system

$$\left[ \begin{array}{cc|c} -1 & 1 & 2 \\ 0 & 1 & 4 \\ 1 & 0 & 3 \end{array} \right] \left[ \begin{array}{c} c_1 \\ c_2 \end{array} \right] = \left[ \begin{array}{c} 2 \\ 4 \\ 3 \end{array} \right]$$

has no solution. We can apply the method of Gaussian Elimination as demonstrated in Section 3.2.1, which leads to

$$\left[ \begin{array}{cc|c} -1 & 1 & 2 \\ 0 & 1 & 4 \\ 1 & 0 & 3 \end{array} \right] \rightarrow \left[ \begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & 4 \\ -1 & 1 & 2 \end{array} \right] \quad R_1 \leftrightarrow R_3$$

$$\rightarrow \left[ \begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{array} \right] \quad R_3 + R_1 - R_2 \rightarrow R_3$$

The last row is inconsistent and hence there is no solution to the linear system and  $\vec{h}$  cannot be expressed by a linear combination of  $\vec{v}^{(1)}$  and  $\vec{v}^{(2)}$ .  $\square$

With the idea of linear combination, we can define the *span* generated by a *finite* set of vectors.

**Definition 6.1.5** (Span). The span of  $q$  vectors in a set  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}\}$  where all of them are from the same vector space  $\mathcal{V}$ , i.e.  $\vec{v}^{(j)} \in \mathcal{V}$  for  $j = 1, 2, \dots, q$ , is another set that contains all their possible linear combinations as given in Definition 6.1.3, and is denoted by

$$\text{span}(\mathcal{B}) = \left\{ \sum_{j=1}^q c_j \vec{v}^{(j)} \mid \text{for all possible values of the scalars } c_j \text{ with } \vec{v}^{(j)} \in \mathcal{B} \right\}$$

Again we will limit ourselves to the cases where the coefficients  $c_j$  are real and  $q$  has to be finite. If the  $\vec{v}^{(j)}$  are from the real  $n$ -space, i.e.  $\vec{v}^{(j)} \in \mathbb{R}^n$ , then as suggested by Properties 6.1.4, the span can be thought in the form of

$$\text{span}(\mathcal{B}) = \{A\vec{x} \mid \text{for any } \vec{x} \in \mathbb{R}_q\}$$

with  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \dots | \vec{v}^{(q)}]$  is an  $n \times q$  matrix and  $\vec{x} = (c_1, c_2, \dots, c_q)^T$  being the coefficient vector.

For example, the span of  $\mathcal{B}_1 = \{(-1, 1)^T\}$  is simply  $t(-1, 1)^T$  where  $-\infty < t < \infty$ , or the line  $y = -x$ . The span of  $\mathcal{B}_2 = \{(1, 0, 2)^T, (0, 1, -1)^T\}$  (notice that the two vectors are not a constant multiple of each other and thus non-parallel) is  $s(1, 0, 2)^T + t(0, 1, -1)^T$  where  $-\infty < s, t < \infty$ , or represented by the plane  $2x - y - z = 0$  (see Section 5.1.3). Adding more vectors in the *spanning set* does not always imply the corresponding span will be larger. For example, the span of  $\mathcal{B}_3 = \{(1, 0)^T, (0, 1)^T\}$  and  $\mathcal{B}_4 = \{(1, 0)^T, (0, 1)^T, (1, 1)^T, (1, -1)^T\}$  are both apparently  $\mathbb{R}^2$ . This issue will be addressed in the next subsection.

**Example 6.1.3.** Show that any vector in  $\mathbb{R}^2$  can be written as infinitely many different linear combinations of the four vectors in the set  $\mathcal{B}_4$  mentioned above.

*Solution.* This is to decide if the linear system

$$\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

has infinitely many solutions for any pair of  $(x, y)$ . The augmented form

$$\left[ \begin{array}{cccc|c} 1 & 0 & 1 & 1 & x \\ 0 & 1 & 1 & -1 & y \end{array} \right]$$

is already in reduced row echelon form. There is a corresponding pivot for both  $x$  and  $y$  in the first two columns, and no zero row is present, which means that there would not be any inconsistency and we can always construct a family of solutions by setting the non-pivotal unknowns to be free variables, let's say  $c_3 = s$  and  $c_4 = t$ . Then we have  $c_1 = x - s - t$ ,  $c_2 = y - s + t$  from the rows. As a result, any linear combination in the form of

$$(x - s - t)(1, 0)^T + (y - s + t)(0, 1)^T + s(1, 1)^T + t(1, -1)^T$$

will produce the vector  $(x, y)^T$  with any value of  $s$  and  $t$  as desired, and there are infinitely many of them. This example shows that a vector (in this case any arbitrary vector of  $\mathbb{R}^2$ ) can possibly be written as more than one linear combinations of the constituent vectors in the spanning set (here  $\mathcal{B}_4$ ).  $\square$

An essential property of spans is that they are subspaces and vice versa. This fact integrates the top-down (it is a subset of a larger vector space) and bottom-up (it is formed by linear combinations of vectors) view of subspaces.

**Properties 6.1.6.** The span of a subset of some vectors in  $\mathcal{V}$  is a subspace of  $\mathcal{V}$ . A subspace of  $\mathcal{V}$  is always some span (not necessarily unique) of some vectors

in  $\mathcal{V}$ .

We leave the proof for showing the span → subspace direction in the footnote<sup>4</sup> and that for the subspace → span direction in the Appendix. Subsequently, we say  $\mathcal{W} = \text{span}(\mathcal{B})$  is a subspace of  $\mathcal{V}$  ( $\mathcal{W} \subseteq \mathcal{V}$ ) *generated* by the set  $\mathcal{B}$  and  $\mathcal{B}$  is known as a ***spanning/generating set*** for  $\mathcal{B}$ . This duality between subspace and span is consistent when we look at them from a geometric point of view: as mentioned at the end of Section 6.1.2 before, subspaces can be thought of as "flat shapes", or put differently, "linear objects" of infinite extent; Meanwhile a span is precisely consisted of all possible linear combinations of vectors. These spanning vectors represent straight directions that extend infinitely long and also produce a "linear shape". (also see Figure 6.2) Applying Properties 6.1.6 on Definition 6.1.5, we can say that the span generated by the column vectors  $\vec{v}^{(j)}$  in  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \cdots | \vec{v}^{(q)}]$  forms a subspace better known as the ***column space*** of  $A$ .

**Definition 6.1.7** (Column Space). The column space of an  $n \times q$  matrix  $A$  is the span generated by the  $q$  column vectors  $\in \mathbb{R}^n$  that make up  $A$  as suggested in Definition 6.1.5.

Finally, a result related to Properties 6.1.6 is noted below.

**Properties 6.1.8.** Any subspace of  $\mathcal{V}$  that contains a subset  $\mathcal{B}'$  of some vectors in  $\mathcal{V}$  also contains  $\text{span}(\mathcal{B}')$ .

---

<sup>4</sup>We check if the two criteria in Theorem 6.1.2 hold for a span. Let the span be the one defined in Definition 6.1.5, then any vector in the span can be written as  $\sum_{j=1}^q c_j \vec{v}^{(j)}$  for some constants  $c_j$ . Let  $\vec{u} = \sum_{j=1}^q \alpha_j \vec{v}^{(j)} \in \text{span}(\mathcal{B})$  and  $\vec{v} = \sum_{j=1}^q \beta_j \vec{v}^{(j)} \in \text{span}(\mathcal{B})$  are both in the span for some sets of constants  $\alpha_j$  and  $\beta_j$ , then their sum  $\vec{u} + \vec{v} = \sum_{j=1}^q \alpha_j \vec{v}^{(j)} + \sum_{j=1}^q \beta_j \vec{v}^{(j)} = \sum_{j=1}^q (\alpha_j + \beta_j) \vec{v}^{(j)} = \sum_{j=1}^q \gamma_j \vec{v}^{(j)} \in \text{span}(\mathcal{B})$  where  $\gamma_j = \alpha_j + \beta_j$  is also in the span and hence it closed under addition. Similarly, writing  $a\vec{w} = a(\sum_{j=1}^q \beta_j \vec{v}^{(j)}) = \sum_{j=1}^q (a\beta_j) \vec{v}^{(j)}$  shows that  $a\vec{w} \in \text{span}(\mathcal{B})$  and the span is closed under scalar multiplication and we are done.

### 6.1.4 Linear Independence, CR Factorization

Another key concept in this chapter is the problem of *linear independence*, which has profound implications in Linear Algebra. Given a set of vectors, if every one of them can not be expressed as a linear combination of other members, or speaking loosely, each of them is not "dependent" on other vectors, then such a set of vectors is said to be ***linearly independent***. Otherwise, if at least one of them can be expressed as some linear combination of other vectors, then the set is known as ***linearly dependent***.

To check linear independence of  $q$  vectors, one may indeed directly show that for every vector  $\vec{v}^{(j)}$  in the set,  $j = 1, 2, 3, \dots, q$ , it cannot be written as the linear combination of other vectors  $\vec{v}^{(k)}$  in the set,  $k \neq j$ . A slightly easier way is looking at the linear combination of just the first  $j - 1$  vectors (from  $\vec{v}^{(1)}$  up to  $\vec{v}^{(j-1)}$ ) for  $\vec{v}^{(j)}$ . However, it is very tedious if the amount of vectors is large. Fortunately, we have a theorem which significantly simplifies our work.

**Theorem 6.1.9.** For a set of vectors  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}\}$  where  $\vec{v}^{(j)} \in \mathcal{V}$ ,  $j = 1, 2, \dots, q$  that are from the same vector space, they are linearly independent if and only if, the equation

$$c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$$

has the trivial solution where all the coefficients are zeros ( $c_j = \mathbf{0}$ ) as its unique solution. Using the language in Properties 6.1.4 if  $\vec{v}^{(j)} \in \mathbb{R}^n$  come from a real  $n$ -space, it means that the homogeneous linear system  $A\vec{x} = \mathbf{0}$  where  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \cdots | \vec{v}^{(q)}]$  is an  $n \times q$  matrix only has the trivial solution  $\vec{x} = \mathbf{0}$ .

*Proof.* The "if" direction: We need to show that  $c_j = \mathbf{0}$  being the only solution to  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$  implies that  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  are linearly independent. We can prove the contrapositive where the opposite of the conclusion,  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  are linearly dependent, implies the opposite of the premise, i.e. there is non-trivial solution to the equation. This

requires that at least one of these vectors, without the loss of generality let's say  $\vec{v}^{(1)}$ , can be written as the linear combination of other vectors in the form of

$$\vec{v}^{(1)} = a_2 \vec{v}^{(2)} + a_3 \vec{v}^{(3)} + \cdots + a_q \vec{v}^{(q)}$$

Rearranging gives

$$\vec{v}^{(1)} - a_2 \vec{v}^{(2)} - a_3 \vec{v}^{(3)} - \cdots - a_q \vec{v}^{(q)} = \mathbf{0}$$

which shows that the coefficients  $c_1 = 1, c_2 = -a_2, c_3 = -a_3, \dots, c_q = -a_q$  is another solution other than  $c_j = \mathbf{0}$  to  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$  (concerning  $c_1$  particularly).

The "only if" direction: We want to show the converse that linear independence of  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  only permits  $c_j = \mathbf{0}$  as the unique solution to  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$ . To do so, we can again resort to its contrapositive, i.e. the existence of an alternative solution of  $c_j = a_j$  which are not all zeros to the equation in question, means that the vectors  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  are linearly dependent. Choose one of the  $a_j$  that is not zero and denote it by  $a_k$ , then

$$a_1 \vec{v}^{(1)} + \cdots + a_{k-1} \vec{v}^{(k-1)} + a_k \vec{v}^{(k)} + a_{k+1} \vec{v}^{(k+1)} + \cdots + a_q \vec{v}^{(q)} = \mathbf{0}$$

$$\vec{v}^{(k)} = -\frac{a_1}{a_k} \vec{v}^{(1)} - \cdots - \frac{a_{k-1}}{a_k} \vec{v}^{(k-1)} - \frac{a_{k+1}}{a_k} \vec{v}^{(k+1)} - \cdots - \frac{a_q}{a_k} \vec{v}^{(q)}$$

where we have divided the equation by the non-zero  $a_k$  to avoid dividing by zero and rearranged it to show that  $\vec{v}^{(k)}$  can be written in some linear combination of other vectors  $\vec{v}^{(j)}, j \neq k$  as shown above, and thus vectors in  $\mathcal{B}$  are linearly dependent.  $\square$

As a corollary, any set containing the zero vector  $\mathbf{0}$  must be linearly dependent. (Why?)<sup>5</sup>

---

<sup>5</sup>For any such a set  $\mathcal{B}_0 = \{\vec{u}_1, \vec{u}_2, \dots, \mathbf{0}\}$ , the linear system  $c_1 \vec{u}_1 + c_2 \vec{u}_2 + \cdots + c_0 \mathbf{0} = \mathbf{0}$  has a family of infinitely many solution with  $c_j = 0$  for  $j \neq 0$  and any value of  $c_0$ , which by Theorem 6.1.9 they are linearly dependent.

**Example 6.1.4.** Determine if  $\vec{u} = (1, 2, 1)^T$ ,  $\vec{v} = (3, 4, 2)^T$ ,  $\vec{w} = (6, 8, 1)^T$  are linearly independent.

By Theorem 6.1.9, this is equivalent to decide if  $A\vec{x} = \mathbf{0}$ , where  $A = [\vec{u}|\vec{v}|\vec{w}]$  has the trivial solution as the only solution. With the help of Theorem 3.1.2, we know that it is equivalent to check if  $\det(A)$  is zero or not. Since

$$|A| = \begin{vmatrix} 1 & 3 & 6 \\ 2 & 4 & 8 \\ 1 & 2 & 1 \end{vmatrix} = 6 \neq 0$$

We conclude that  $A\vec{x} = \mathbf{0}$  only has the trivial solution  $\vec{x} = \mathbf{0}$  and these three vectors are linearly independent.

Short Exercise: Redo the above example with  $\vec{u} = (1, 1, 3)^T$ ,  $\vec{v} = (1, 3, 2)^T$ ,  $\vec{w} = (2, 8, 3)^T$ .<sup>6</sup>

Including our earlier discussion in Section 3.2.1, Theorem 6.1.9 gives some interesting results.

1. If there are  $q$  vectors of  $\mathbb{R}^p$  in a set and  $p < q$ , i.e. the amount of vectors is more than their dimension, then  $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\dots|\vec{v}^{(q)}]$  is an  $p \times q$  matrix which has more columns ( $q$ ) than rows ( $p$ ). In this case  $A\vec{x} = \mathbf{0}$  must have at least one free variables and thus infinitely many solutions, hence the vectors must be linearly dependent.
2. Otherwise ( $p \geq q$ ), we can solve  $A\vec{x} = \mathbf{0}$  by Gaussian Elimination to see if it only has the trivial solution. If so (not), the vectors are linearly independent (dependent). Alternatively, if  $A$  is a square matrix, then we may check if its determinant is non-zero, just like what have been done in Example 6.1.4. Gaussian Elimination still works for any square matrix, and in case of linear independence (dependence),  $A$  will (not) be reduced to an identity matrix.

---

<sup>6</sup>The determinant of  $A = [\vec{u}|\vec{v}|\vec{w}]$  in the case is  $|A| = 0$ , and hence by the remark for Theorem 3.1.2 the linear system  $A\vec{x} = \mathbf{0}$  has infinitely many solutions, and these three vectors are linearly dependent by Theorem 6.1.9.

In many cases the number of vectors are indeed not equal to their dimension so the method of using determinant to check linear independence in the last example does not apply and we need to resort to Gaussian Elimination. In fact, Gaussian Elimination can disclose more information than just if a set of vectors is linearly (in)dependent as an entirety in both cases, but also how exactly these vectors are dependent on each other, soon to be explained. Before doing so, we note that the above observations lead to an extension of Theorem 3.2.1.

**Theorem 6.1.10** (Equivalence Statement, ver. 3). For an  $n \times n$  real square matrix  $A$ , the followings are equivalent:

- (a)  $A$  is invertible, i.e.  $A^{-1}$  exists,
- (b)  $\det(A) \neq 0$ ,
- (c) The reduced row echelon form of  $A$  is  $I$ ,
- (d) The linear system  $A\vec{x} = \vec{h}$  has a unique solution for any  $\vec{h}$ , particularly  $A\vec{x} = \mathbf{0}$  has only the trivial solution  $\vec{x} = \mathbf{0}$ ,
- (e) The  $n$  column vectors  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(n)}$  of  $\mathbb{R}^n$  as in  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \dots | \vec{v}^{(n)}]$  are linearly independent.

We now revisit the procedure of Gaussian Elimination and show that it actually explicitly reveals the so-called **dependence relations** between vectors (how a vector can be written as some linear combination of other vectors) as a by-product when determining linear (in)dependence. Let's illustrate this by an example: Given

$$\begin{aligned}\vec{v}^{(1)} &= (1, 2, 1)^T \\ \vec{v}^{(2)} &= (2, 4, 2)^T \\ \vec{v}^{(3)} &= (1, -1, -1)^T \\ \vec{v}^{(4)} &= (2, 1, 0)^T \\ \vec{v}^{(5)} &= (0, -3, -2)^T\end{aligned}$$

Note that the vectors are related by these dependence relations:  $\vec{v}^{(2)} = 2\vec{v}^{(1)}$ ,  $\vec{v}^{(4)} = \vec{v}^{(1)} + \vec{v}^{(3)}$  and  $\vec{v}^{(5)} = -\vec{v}^{(1)} + \vec{v}^{(3)}$ , while  $\vec{v}^{(1)}$  and  $\vec{v}^{(3)}$  are themselves

linearly independent of each other. Construct

$$A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \vec{v}^{(4)} | \vec{v}^{(5)}]$$

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

by concatenating the five vectors column by column. Now we carry out Gaussian Elimination as follows.

$$\begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 4 & -1 & 1 & -3 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & -3 & -3 & -3 \\ 0 & 0 & -2 & -2 & -2 \end{bmatrix} \quad R_2 - 2R_1 \rightarrow R_2$$

$$\qquad\qquad\qquad \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 & -2 \end{bmatrix} \quad R_3 - R_1 \rightarrow R_3$$

$$\qquad\qquad\qquad \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad -\frac{1}{3}R_2 \rightarrow R_2$$

$$\qquad\qquad\qquad \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad R_3 + 2R_2 \rightarrow R_3$$

$$\qquad\qquad\qquad \begin{bmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad R_1 - R_2 \rightarrow R_1$$

The new columns in the above rref matrix  $A_{\text{rref}} = [\vec{v}^{(1)'} | \vec{v}^{(2)'} | \vec{v}^{(3)'} | \vec{v}^{(4)'} | \vec{v}^{(5)'}]$  follow the exact same dependence relation:  $\vec{v}^{(2)'} = 2\vec{v}^{(1)'}$ ,  $\vec{v}^{(4)'} = \vec{v}^{(1)'} + \vec{v}^{(3)'}$  and  $\vec{v}^{(5)'} = -\vec{v}^{(1)'} + \vec{v}^{(3)'}$ .  $\vec{v}^{(1)'}$  and  $\vec{v}^{(3)'}$  are clearly still linearly independent of each other too. This demonstrates that dependence relations (and by extension linear independent vectors) are preserved under elementary row operations during Gaussian Elimination. (The detailed argument is put in the following footnote.<sup>7</sup>)

We now introduce a helper theorem so that we may proceed.

---

<sup>7</sup>We will show this for the case of row addition/subtraction only because the other two types of elementary row operations are easy to check. Without loss of generality, take a dependence relation in the form of  $\vec{v}^{(r+1)} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + \cdots + c_r\vec{v}^{(r)}$  where  $A =$

**Theorem 6.1.11** (Plus/Minus Theorem). Let  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}\}$  be a set of vectors in the vector space  $\mathcal{V}$ , i.e.  $\vec{v}^{(j)} \in \mathcal{V}, 1 \leq j \leq q$ , we have the following two results:

$[\vec{v}^{(1)} | \vec{v}^{(2)} | \dots | \vec{v}^{(r)} | \vec{v}^{(r+1)}]$ , i.e.

$$\begin{bmatrix} \vdots \\ \vec{v}_p^{(r+1)} \\ \vdots \\ \vec{v}_q^{(r+1)} \\ \vdots \end{bmatrix} = c_1 \begin{bmatrix} \vdots \\ \vec{v}_p^{(1)} \\ \vdots \\ \vec{v}_q^{(1)} \\ \vdots \end{bmatrix} + c_2 \begin{bmatrix} \vdots \\ \vec{v}_p^{(2)} \\ \vdots \\ \vec{v}_q^{(2)} \\ \vdots \end{bmatrix} + \dots + c_r \begin{bmatrix} \vdots \\ \vec{v}_p^{(r)} \\ \vdots \\ \vec{v}_q^{(r)} \\ \vdots \end{bmatrix}$$

The elementary row operation of adding  $c_q$  times row  $R_q$  to row  $R_p$  then produces a new matrix  $A'$  with column vectors

$$\vec{v}^{(j)'} = \begin{bmatrix} \vdots \\ \vec{v}_p^{(j)} + c_q \vec{v}_q^{(j)} \\ \vdots \\ \vec{v}_q^{(j)} \\ \vdots \end{bmatrix}$$

for all  $j$ . Therefore,

$$\begin{aligned} \vec{v}^{(r+1)'} &= \begin{bmatrix} \vdots \\ \vec{v}_p^{(r+1)} + c_q \vec{v}_q^{(r+1)} \\ \vdots \\ \vec{v}_q^{(r+1)} \\ \vdots \end{bmatrix} \\ &= \begin{bmatrix} \vdots \\ (c_1 \vec{v}_p^{(1)} + c_2 \vec{v}_p^{(2)} + \dots + c_r \vec{v}_p^{(r)}) + c_q (c_1 \vec{v}_q^{(1)} + c_2 \vec{v}_q^{(2)} + \dots + c_r \vec{v}_q^{(r)}) \\ \vdots \\ (c_1 \vec{v}_q^{(1)} + c_2 \vec{v}_q^{(2)} + \dots + c_r \vec{v}_q^{(r)}) \\ \vdots \end{bmatrix} \end{aligned}$$

- (a) If  $\mathcal{B}$  is a linearly independent set and  $\vec{v}$  is not in  $\text{span}(\mathcal{B})$ , then  $\mathcal{B} \cup \{\vec{v}\}$  formed after inserting  $\vec{v}$  into the set is still linearly independent,
- (b) If  $\vec{w}$  is a vector in some other set (also denoted by  $\mathcal{B}$ ) that can be expressed as a linear combination of other vectors in the now linearly dependent set, then the new set  $\mathcal{B} - \{\vec{w}\}$  remained after removing  $\vec{w}$  from  $\mathcal{B}$  has the same span, i.e.

$$\text{span}(\mathcal{B}) = \text{span}(\mathcal{B} - \{\vec{w}\})$$

*Proof.* We include the proof for (a) as a footnote since (a) is less of a concern.<sup>8</sup>

---

by the original dependence relation, which is then equal to

$$\begin{aligned} & \left[ \begin{array}{c} c_1(\vec{v}_p^{(1)} + c_q \vec{v}_q^{(1)}) \\ \vdots \\ c_1 \vec{v}_q^{(1)} \\ \vdots \end{array} \right] + \left[ \begin{array}{c} c_2(\vec{v}_p^{(2)} + c_q \vec{v}_q^{(2)}) \\ \vdots \\ c_2 \vec{v}_q^{(2)} \\ \vdots \end{array} \right] + \cdots + \left[ \begin{array}{c} c_r(\vec{v}_p^{(r)} + c_q \vec{v}_q^{(r)}) \\ \vdots \\ c_r \vec{v}_q^{(r)} \\ \vdots \end{array} \right] \\ &= c_1 \vec{v}^{(1)'} + c_2 \vec{v}^{(2)'} + \cdots + c_r \vec{v}^{(r)'} \end{aligned}$$

This shows that the same dependence relation holds between the new column vectors  $\vec{v}^{(j)'}, j = 1, 2, \dots, r+1$ .

<sup>8</sup>We will prove the contrapositive that given  $\mathcal{B}$  is a linearly independent set then,  $\mathcal{B} \cup \{\vec{v}\}$  is linearly dependent only if  $\vec{v} := \vec{v}^{(q+1)}$  is in  $\text{span}(\mathcal{B})$ : if  $\mathcal{B} \cup \{\vec{v}\}$  is linearly dependent, then there is non-trivial solution  $c_j = d_j$  where  $d_j$  are not all zeros to the equation  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + \cdots + c_q \vec{v}^{(q)} + c_{q+1} \vec{v}^{(q+1)} = \mathbf{0}$  by Theorem 6.1.9. Since  $\mathcal{B}$  is required to be linearly independent,  $d_{q+1} \neq 0$ , for otherwise  $d_{q+1} = 0$  and then at least one of the  $c_j = d_j, j \neq q+1$ , will be non-zero due to the linear dependence of the union set  $\mathcal{B} \cup \{\vec{v}\}$  and lead to a non-trivial solution to  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$  instead, which contradicts the assumed linear independence of  $\mathcal{B}$  alone, so we have  $d_1 \vec{v}^{(1)} + d_2 \vec{v}^{(2)} + \cdots + d_q \vec{v}^{(q)} + d_{q+1} \vec{v}^{(q+1)} = \mathbf{0}$  and because  $d_{q+1} \neq 0$  we can obtain

$$\vec{v}^{(q+1)} = -\frac{1}{d_{q+1}}(d_1 \vec{v}^{(1)} + d_2 \vec{v}^{(2)} + \cdots + d_q \vec{v}^{(q)})$$

showing that  $\vec{v}^{(q+1)}$  is a linear combination of  $\vec{v}^{(j)} \in \mathcal{B}, 1 \leq j \leq q$ .

For (b), assign the vector  $\vec{v}^{(k)}$  that is being removed where  $1 \leq k \leq q$  as  $\vec{w}$ . We can write  $\vec{w} = a_1\vec{v}^{(1)} + a_2\vec{v}^{(2)} + \cdots + a_{k-1}\vec{v}^{(k-1)} + a_{k+1}\vec{v}^{(k+1)} + \cdots + a_q\vec{v}^{(q)}$  using other vectors in  $\mathcal{B}$  where  $a_j, j \neq k$  are some constants. For any vector  $\vec{v} = b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_k\vec{v}^{(k)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)}$  in  $\text{span}(\mathcal{B})$  with  $b_j$  being the coefficients, it can be rewritten as a linear combination of the remaining vectors:

$$\begin{aligned}
 \vec{v} &= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_k\vec{v}^{(k)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} \\
 &= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} + b_k\vec{v}^{(k)} \\
 &= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} + b_k\vec{v}^{(k)} \\
 &\quad + b_k(a_1\vec{v}^{(1)} + a_2\vec{v}^{(2)} + \cdots + a_{k-1}\vec{v}^{(k-1)} + a_{k+1}\vec{v}^{(k+1)} + \cdots + a_q\vec{v}^{(q)}) \\
 &= (b_1 + b_k a_1)\vec{v}^{(1)} + (b_2 + b_k a_2)\vec{v}^{(2)} + (b_{k-1} + b_k a_{k-1})\vec{v}^{(k-1)} \\
 &\quad + (b_{k+1} + b_k a_{k+1})\vec{v}^{(k+1)} + \cdots + (b_q + b_k a_q)\vec{v}^{(q)} \\
 &\in \text{span}(\mathcal{B} - \{\vec{v}^{(k)}\}) = \text{span}(\mathcal{B} - \{\vec{w}\})
 \end{aligned}$$

Therefore for all  $\vec{v} \in \text{span}(\mathcal{B})$ ,  $\vec{v} \in \text{span}(\mathcal{B} - \{\vec{w}\})$  and hence  $\text{span}(\mathcal{B}) \subseteq \text{span}(\mathcal{B} - \{\vec{w}\})$ . It is trivial to show  $\text{span}(\mathcal{B} - \{\vec{w}\}) \subseteq \text{span}(\mathcal{B})$ , and thus  $\text{span}(\mathcal{B}) = \text{span}(\mathcal{B} - \{\vec{w}\})$ . This part of the theorem is very relevant to the span of sets  $\mathcal{B}_3$  and  $\mathcal{B}_4$  in the previous Example 6.1.3.  $\square$

With these results, Gaussian Elimination enables us to carry out the ***Column-row (CR) Factorization*** over a matrix. First, note that by part (b) of Theorem 6.1.11 above, the column space (Definition 6.1.7) of a matrix  $A$  can be expressed as the span of a ***minimal generating set*** by removing linearly dependent column vectors in  $A$  (which does not change the span and still generates the same subspace) and only keeping the linearly independent ones. Meanwhile, the manners of linear (in)dependence over the column vectors of  $A$  can be inferred by Gaussian Elimination as just demonstrated in the last example. After obtaining the minimal generating set, we can express any vector in the column space as a unique linear combination of these linearly independent vectors inside the set due to the following properties.

**Properties 6.1.12.** For a set of vectors  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}\}$ ,  $\vec{v}^{(j)} \in \mathcal{V}$  for  $j = 1, 2, \dots, q$  which are linearly independent, any vector  $\vec{v} \in \text{span}(\mathcal{B})$  in their span can be written as a unique linear combination of these generating vectors in  $\mathcal{B}$ . Otherwise, if the vectors in  $\mathcal{B}$  are linearly dependent, there will be infinitely many such linear combinations to assemble  $\vec{v}$ .

Again, we will simply provide the proof in a footnote for reference.<sup>9</sup> Return to

---

<sup>9</sup>We will show the first part only. Since  $\vec{v}$  already belongs to  $\text{span}(\mathcal{B})$ , it must be possible to express  $\vec{v}$  as some linear combination(s) of vectors  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  in  $\mathcal{B}$  by Definition 6.1.5. Now it suffices to show that it is unique. Assume the contrary that there are two distinct linear combinations of vectors in  $\mathcal{B}$  that represent  $\vec{v}$ , and hence we can express it by

$$\begin{aligned}\vec{v} &= d_1 \vec{v}^{(1)} + d_2 \vec{v}^{(2)} + d_3 \vec{v}^{(3)} + \cdots + d_q \vec{v}^{(q)} \\ &= g_1 \vec{v}^{(1)} + g_2 \vec{v}^{(2)} + g_3 \vec{v}^{(3)} + \cdots + g_q \vec{v}^{(q)}\end{aligned}$$

where  $d_j, g_j$  are two sets of coefficients are not exactly the same. Subtracting one expression by another leads to

$$\begin{aligned}(d_1 \vec{v}^{(1)} + d_2 \vec{v}^{(2)} + d_3 \vec{v}^{(3)} + \cdots + d_q \vec{v}^{(q)}) - (g_1 \vec{v}^{(1)} + g_2 \vec{v}^{(2)} + g_3 \vec{v}^{(3)} + \cdots + g_q \vec{v}^{(q)}) &= \vec{v} - \vec{v} \\ (d_1 - g_1) \vec{v}^{(1)} + (d_2 - g_2) \vec{v}^{(2)} + (d_3 - g_3) \vec{v}^{(3)} + \cdots + (d_q - g_q) \vec{v}^{(q)} &= \mathbf{0}\end{aligned}$$

Since  $d_j, g_j$  are assumed to be not completely identical, it is a non-trivial solution to the equation  $c_1 \vec{v}^{(1)} + c_2 \vec{v}^{(2)} + c_3 \vec{v}^{(3)} + \cdots + c_q \vec{v}^{(q)} = \mathbf{0}$ , where  $c_j = d_j - g_j$  are not all zeros.

our example, where

$$A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \vec{v}^{(4)} | \vec{v}^{(5)}]$$

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

We have found that the corresponding rref is

$$A_{\text{rref}} = \begin{bmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and thus concluded that the first/third column vectors are linearly independent, and the second/fourth/fifth column vectors are linearly dependent on the first/third ones and can be expressed as a linear combination of them. Now let

$$C = [\vec{v}^{(1)} | \vec{v}^{(3)}] = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix}$$

using the two linear independent vectors. By Properties 6.1.12 and 6.1.4, each of the  $\vec{v}^{(j)}$  can be expressed as a unique linear combination in the form of a matrix product between  $C$  and a column vector that contains the coefficients in front of the chosen linear independent vectors that make up the  $\vec{v}^{(j)}$ . The required column vectors to produce them are hence exactly the corresponding columns in the rref which retain the dependence relations, with row(s) of all zeros removed. For instance,

$$\vec{v}^{(5)} = -\vec{v}^{(1)'} + \vec{v}^{(3)'}$$

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

---

This contradicts our assumed linear independence of  $\mathcal{B}$  and hence the linear combination of  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(q)}$  to generate  $\vec{v}$  must be unique.

$$\begin{aligned}
 &= \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} && \text{(Properties 6.1.4)} \\
 &= [\vec{v}^{(1)} | \vec{v}^{(3)}] \begin{bmatrix} -1 \\ 1 \end{bmatrix} = C \begin{bmatrix} -1 \\ 1 \end{bmatrix}
 \end{aligned}$$

Denote

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

which is consisted of the non-zero rows of  $A_{\text{ref}}$ , then similarly

$$\begin{aligned}
 \vec{v}^{(1)} &= \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = CR_1 \\
 \vec{v}^{(2)} &= \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = CR_2 \\
 \vec{v}^{(3)} &= \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = CR_3 \\
 \vec{v}^{(4)} &= \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = CR_4
 \end{aligned}$$

where  $R_j$  is the  $j$ -th column of  $R$ . Therefore,

$$\begin{aligned}
 A &= [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \vec{v}^{(4)} | \vec{v}^{(5)}] = \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 4 & -1 & 1 & -3 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix} \\
 &= [CR_1 | CR_2 | CR_3 | CR_4 | CR_5] \\
 &= C([R_1 | R_2 | R_3 | R_4 | R_5]) = CR = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}
 \end{aligned}$$

from the second line to the third line we use the fact that the same matrix multiplied to the left in every column of another matrix can be factored out (why?)<sup>10</sup> and this is the desired CR Factorization of  $A$ . In general, for any matrix, its CR Factorization is derived as follows.

**Properties 6.1.13** (CR Factorization). The Column-Row Factorization of any matrix  $A$  that has a rref of  $A_{\text{rref}}$ , is given by  $A = CR$ , where  $C$  contains the  $r$  linearly independent columns of  $A$  at which the  $r$  leading 1s of  $A_{\text{rref}}$  are located, and  $R$  is simply the first  $r$  rows of  $A_{\text{rref}}$  that hold these leading 1s, with all the full-zero rows below removed.

The  $k$ -th row of  $R$  contains the coefficients in front of the  $k$ -th column vector in  $C$  required to generate each column in the original  $A$  matrix.

---

<sup>10</sup> For an  $m \times r$  matrix  $A$ , and an  $r \times n$  matrix  $B$ , we have, by Definition 1.1.1 (with a slightly different notation),

$$\begin{aligned} A[B_1|B_2|\cdots|B_n] &= \left[ \begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & & a_{2r} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mr} \end{array} \right] \left[ \begin{array}{c|c|c|c} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rn} \end{array} \right] \\ &= \left[ \begin{array}{cccc} \sum_{k=1}^r a_{1k}b_{k1} & \sum_{k=1}^r a_{1k}b_{k2} & \cdots & \sum_{k=1}^r a_{1k}b_{kn} \\ \sum_{k=1}^r a_{2k}b_{k1} & \sum_{k=1}^r a_{2k}b_{k2} & & \sum_{k=1}^r a_{2k}b_{kn} \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^r a_{mk}b_{k1} & \sum_{k=1}^r a_{mk}b_{k2} & \cdots & \sum_{k=1}^r a_{mk}b_{kn} \end{array} \right] \\ &= [AB_1|AB_2|\cdots|AB_n] \end{aligned}$$

where  $B_j$  now denotes the  $j$ -th column of  $B$ :

$$B_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{rj} \end{bmatrix} \quad \text{and hence } AB_j = \left[ \begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & & a_{2r} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mr} \end{array} \right] \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{rj} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^r a_{1k}b_{kj} \\ \sum_{k=1}^r a_{2k}b_{kj} \\ \vdots \\ \sum_{k=1}^r a_{mk}b_{kj} \end{bmatrix}$$

**Example 6.1.5.** Show that  $\vec{u} = (2, 1, -1, 1)^T$ ,  $\vec{v} = (1, 2, 1, -1)^T$ ,  $\vec{w} = (0, 1, 1, 2)^T$  are linearly independent and find the CR Factorization of  $A = [\vec{u}|\vec{v}|\vec{w}]$ . What if  $\vec{w} = (1, -1, -2, 2)^T$  instead?

*Solution.* From Theorem 6.1.9, we need to show that the system  $A\vec{x} = \mathbf{0}$  has only the trivial solution  $\vec{x} = \mathbf{0}$ , where

$$A = [\vec{u}|\vec{v}|\vec{w}] = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 2 \end{bmatrix}$$

To do so we can apply Gaussian Elimination as below.

$$\begin{array}{c} \left[ \begin{array}{ccc|c} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 1 & -1 & 2 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 \end{array} \right] \quad R_1 \leftrightarrow R_4 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 3 & -1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 3 & -4 & 0 \end{array} \right] \quad R_2 - R_1 \rightarrow R_2 \\ \quad \quad \quad R_3 + R_1 \rightarrow R_3 \\ \quad \quad \quad R_4 - 2R_1 \rightarrow R_4 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 3 & -4 & 0 \end{array} \right] \quad \frac{1}{3}R_2 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & -3 & 0 \end{array} \right] \quad R_4 - 3R_2 \rightarrow R_4 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -3 & 0 \end{array} \right] \quad \frac{1}{3}R_3 \rightarrow R_3 \end{array}$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_4 + 3R_1 \rightarrow R_4$$

(the zero column to the right can be omitted) The forward phase leads to a redundant row and the presence of pivots in every column indicates that the trivial solution of  $\vec{x} = 0$  is the only solution, hence the three vectors  $\vec{u}, \vec{v}, \vec{w}$  are linearly independent (refer to Section 3.2.1). The backward phase concerning the matrix  $A$  itself is instantaneous, yielding its rref:

$$\left[ \begin{array}{ccc} 1 & -1 & 2 \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right]$$

By Properties 6.1.13, the CR Factorization of  $A$  is then trivially

$$A = \left[ \begin{array}{ccc} 2 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 2 \end{array} \right] = \left[ \begin{array}{ccc} 2 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 2 \end{array} \right] \left[ \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] = CR$$

with  $C = A$  and  $R = I_3$  as all columns in  $A$  are linearly independent. In general, if the  $n$  column vectors in an  $m \times n$  matrix  $A$  are linearly independent<sup>11</sup>, then its rref will be in the form of

$$A_{\text{rref}} = \left[ \begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \end{array} \right]$$

---

<sup>11</sup>Bear in mind that it is necessary to have  $m \geq n$ .

where the top is an  $n \times n$  identity matrix  $I_n$ , followed by  $m - n$  rows of full zeros at the bottom. The CR Factorization of  $A$  will then be simply comprised of  $C = A$  and  $R = I_n$ . For the second case where  $\vec{w} = (1, -1, -2, 2)^T$ , we can repeat the same analysis by deriving the rref of the modified  $A$  matrix.

$$\begin{aligned}
 \left[ \begin{array}{ccc|c} 2 & 1 & 1 & \\ 1 & 2 & -1 & \\ -1 & 1 & -2 & \\ 1 & -1 & 2 & \end{array} \right] &\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & \\ 1 & 2 & -1 & \\ -1 & 1 & -2 & \\ 2 & 1 & 1 & \end{array} \right] && R_1 \leftrightarrow R_4 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & \\ 0 & 3 & -3 & \\ 0 & 0 & 0 & \\ 0 & 3 & -3 & \end{array} \right] && R_2 - R_1 \rightarrow R_2 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & \\ 0 & 1 & -1 & \\ 0 & 0 & 0 & \\ 0 & 3 & -3 & \end{array} \right] && R_3 + R_1 \rightarrow R_3 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & \\ 0 & 1 & -1 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \end{array} \right] && \frac{1}{3}R_2 \rightarrow R_2 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 2 & \\ 0 & 1 & -1 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \end{array} \right] && R_4 - 3R_2 \rightarrow R_4 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & \\ 0 & 1 & -1 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \end{array} \right] && R_1 + R_2 \rightarrow R_1
 \end{aligned}$$

The final rref reveals that  $\vec{u}$  and  $\vec{v}$  are two linearly independent vectors in the column space of  $A$ , in addition to the dependence relation of  $\vec{w} = \vec{u} - \vec{v}$ . Hence its new CR Factorization, by Properties 6.1.13, is

$$\left[ \begin{array}{ccc} 2 & 1 & 1 \\ 1 & 2 & -1 \\ -1 & 1 & -2 \\ 1 & -1 & 2 \end{array} \right] = \left[ \begin{array}{cc} 2 & 1 \\ 1 & 2 \\ -1 & 1 \\ 1 & -1 \end{array} \right] \left[ \begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & -1 \end{array} \right]$$

□

## 6.2 Coordinate Bases for $\mathbb{R}^n$ and its Subspaces

### 6.2.1 Coordinate Bases for $\mathbb{R}^n$

Back in Definition 4.1.3, we have introduced the  $n$  standard unit vectors  $\hat{e}^{(1)}, \hat{e}^{(2)}, \dots, \hat{e}^{(n)}$  for the real  $n$ -space  $\mathbb{R}^n$ . Obviously the standard unit vectors are linearly independent and their span is exactly  $\mathbb{R}^n$ . We often refer to the coefficients  $x_j$  in front of  $\hat{e}^{(j)}$  of a vector  $\vec{x} = (x_1, x_2, \dots, x_n)^T = x_1\hat{e}^{(1)} + x_2\hat{e}^{(2)} + \dots + x_n\hat{e}^{(n)}$  in  $\mathbb{R}^n$  as the *Cartesian coordinates* of  $\vec{x}$ . The coordinates  $x_j$  are unique, guaranteed by Properties 6.1.12. However, sometimes we may want to express an  $\mathbb{R}^n$  vector in another **coordinate basis (system)** with axes different from the standard unit vectors, that is, other than the **standard basis**  $S = \{\hat{e}^{(1)}, \hat{e}^{(2)}, \dots, \hat{e}^{(n)}\}$ . Motivated by the properties of the Cartesian coordinate system above, in which the standard unit vectors are linearly independent and span  $\mathbb{R}^n$  such that every vector in  $\mathbb{R}^n$  can be expressed as a unique linear combination of them, we require all other coordinate bases for  $\mathbb{R}^n$  to carry the same properties. The coefficients of this linear combination will then become the coordinates of that vector with respect to this basis.

**Definition 6.2.1** (Coordinate Basis for  $\mathbb{R}^n$ ). A coordinate basis  $\mathcal{B}$  for  $\mathbb{R}^n$  should consists of  $n$  vectors  $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)}\}$  where  $\vec{v}^{(j)} \in \mathbb{R}^n$ , which

- (a) are linearly independent, and
- (b) span (generate)  $\mathbb{R}^n$ .

Some may wonder why the definition above has explicitly stated that the number of vectors in a coordinate basis for  $\mathbb{R}^n$  is exactly  $n$ , although many people would probably think it is reasonable and accept this without a doubt. For the sake of completeness, we will explain that this is a result coming naturally from the conditions of linear independence and spanning  $\mathbb{R}^n$ . We have previously shown that Theorem 6.1.9 implies that in  $\mathbb{R}^n$  if there are more vectors  $q$  than the dimension  $n$  then they will be linearly dependent. So linear independence

requires  $q \leq n$ . To span  $\mathbb{R}^n$ , it is apparent that  $q \geq n$ .<sup>12</sup> Hence the number of vectors  $q$  must be equal to  $n$ .

The following theorem shows that we actually only need to check either one of the conditions in Definition 6.2.1.

**Theorem 6.2.2.** A set of  $n$  vectors of  $\mathbb{R}^n$  is linearly independent if and only if they span  $\mathbb{R}^n$ .

*Proof.* Linear Independence  $\rightarrow$  Spanning  $\mathbb{R}^n$ : Assume  $\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)}$  are linear independent with  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \dots | \vec{v}^{(n)}]$  being a square matrix. The application of part (e)  $\rightarrow$  (d) of Theorem 6.1.10 immediately shows that there is always a (unique) solution to  $A\vec{x} = \vec{h}$  for any  $\vec{h}$  of  $\mathbb{R}^n$ . Recall that  $A\vec{x}$  represents the span of  $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)}\}$  (Definition 6.1.5) so it implies that these vectors generates the entire  $\mathbb{R}^n$ .

Spanning  $\mathbb{R}^n \rightarrow$  Linear Independence: Assume the opposite of the implication that  $\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)}$  are linear dependent, then by (c) and (e) of Theorem 6.1.10 the reduced row echelon form of  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \dots | \vec{v}^{(n)}]$  is not the identity matrix and contains at least one row of full zeros. Following a logic similar to Footnote 12, these vectors cannot span  $\mathbb{R}^n$  and the contrapositive is proved.  $\square$

---

<sup>12</sup> To formally show this, express the span of  $q$   $\mathbb{R}^n$  vectors  $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \dots + c_q\vec{v}^{(q)}$  by  $A\vec{x}$  where  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \vec{v}^{(3)} | \dots | \vec{v}^{(q)}]$  is an  $n \times q$  matrix and  $\vec{x} = (c_1, c_2, c_3, \dots, c_q)^T$  consists of  $q$  coefficients as unknowns (Properties 6.1.4). If  $q < n$ , then  $A\vec{x} = \vec{h}$  is an overdetermined system so that we can always find some row of full zeros in the rref of  $A$  to the left of the augmented matrix as we solve the system by Gaussian Elimination. Along the column vector to the right of the augmented matrix that undergoes the reduction process together, we can always set the number on such a row to some non-zero number (let's say, 1) if not already, to make sure it is inconsistent. Invert the entire process of Gaussian Elimination over the augmented matrix to recover  $A$  from its reduced form. To the right of the augmented matrix will then appear  $\vec{h}_{\text{inconst}}$ . This system  $A\vec{x} = \vec{h}_{\text{inconst}}$  is inconsistent by the design above (just do the same steps of Gaussian Elimination again and the inconsistent 1 to the right will reappear), which shows that the span does not include  $\vec{h}_{\text{inconst}}$  and cannot cover the entire  $\mathbb{R}^n$ .

**Example 6.2.1.** Show that  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\} = \{(1, 2, 1)_S^T, (-1, 1, 0)_S^T, (1, -1, 2)_S^T\}$  forms a basis for  $\mathbb{R}^3$  and express  $\vec{v}$  in  $\mathcal{B}$  (a.k.a.  $[\vec{v}]_B$ ) where  $[\vec{v}]_S = (2, 1, 2)_S^T$ , the subscript  $S$  ( $B$ ) emphasizes that the coordinates are relative to the standard basis  $\mathcal{S}$  (the new basis  $\mathcal{B}$ ).

*Solution.* By Definition 6.2.1 and Theorem 6.2.2, the first part is equivalent to checking if the three  $\mathbb{R}^3$  vectors in  $\mathcal{B}$  are linearly independent. By (b) to (e) of Theorem 6.1.10, we can simply check if  $\det(A)$  is non-zero where

$$A = [[\vec{v}^{(1)}]_S | [\vec{v}^{(2)}]_S | [\vec{v}^{(3)}]_S] = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix}$$

A simple calculation reveals that  $\det(A) = 6 \neq 0$  so  $\mathcal{B}$  is indeed a valid basis for  $\mathbb{R}^3$ . To express  $(2, 1, 2)_S^T$  in  $\mathcal{B}$  is to find  $[\vec{v}]_B = ([v_1]_B, [v_2]_B, [v_3]_B)_B^T$  where  $[v_j]_B$  is the  $j$ -th component (coefficient) of  $\vec{v}$  in the  $\mathcal{B}$  coordinate system such that the corresponding linear combination of  $\vec{v}^{(j)}$  below produces the desired vector that is consistent in the  $\mathcal{S}$  basis as well:

$$\begin{aligned} [v_1]_B(\vec{v}^{(1)}) + [v_2]_B(\vec{v}^{(2)}) + [v_3]_B(\vec{v}^{(3)}) &= \vec{v} \\ [v_1]_B(1, 2, 1)_S^T + [v_2]_B(-1, 1, 0)_S^T + [v_3]_B(1, -1, 2)_S^T &= (2, 1, 2)_S^T \end{aligned}$$

or put in matrix form,

$$\begin{aligned} [v_1]_B[\vec{v}^{(1)}]_S + [v_2]_B[\vec{v}^{(2)}]_S + [v_3]_B[\vec{v}^{(3)}]_S &= [\vec{v}]_S \\ A[\vec{v}]_B = [[\vec{v}^{(1)}]_S | [\vec{v}^{(2)}]_S | [\vec{v}^{(3)}]_S] \begin{bmatrix} [v_1]_B \\ [v_2]_B \\ [v_3]_B \end{bmatrix} &= \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} [v_1]_B \\ [v_2]_B \\ [v_3]_B \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \end{aligned}$$

We can either use matrix inverse or Gaussian Elimination to solve for the  $[v_j]_B$ , yielding  $[v_1]_B = 1$ ,  $[v_2]_B = -\frac{1}{2}$ ,  $[v_3]_B = \frac{1}{2}$ , and hence  $[\vec{v}]_B = (1, -\frac{1}{2}, \frac{1}{2})_B^T$ . Notice, the matrix equation

$$A[\vec{v}]_B = [[\vec{v}^{(1)}]_S | [\vec{v}^{(2)}]_S | [\vec{v}^{(3)}]_S] \begin{bmatrix} [v_1]_B \\ [v_2]_B \\ [v_3]_B \end{bmatrix} = [\vec{v}]_S$$

shows that the matrix  $A$  transforms the coordinate system from the  $\mathcal{B}$  to  $\mathcal{S}$  basis in which a given vector  $\vec{v}$  is expressed, and we will write  $A = P_B^S$  (thus  $P_B^S[\vec{v}]_B = [\vec{v}]_S$ ) for clarity in the future. Also, the vector  $\vec{v}$  is still the same one despite having different coordinate representations since only the reference frame is changed. For this we will henceforth use the equivalence symbol to write along the lines of  $\vec{v} \equiv [\vec{v}]_B \equiv [\vec{v}]_S$ . Notice that  $P_B^S[\vec{v}^{(j)}]_B = P_B^S(e^{(j)})_B$  returns the  $j$ -th basis vector in  $\mathcal{B}$  ( $j$ -th column of  $P_B^S$ ) expressed in the original standard basis  $\mathcal{S}$ , namely  $[\vec{v}^{(j)}]_S$ , where  $(e^{(j)})_B = [\vec{v}^{(j)}]_B$  is the numeric coordinate representation (emphasized by the absence of hat symbol over  $e$ ) of the  $j$ -th basis vector in the  $\mathcal{B}$  coordinate system with the  $j$ -th component being 1 and other being 0. For instance<sup>13</sup>,

$$P_B^S[\vec{v}^{(2)}]_B = P_B^S(e^{(2)})_B = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = [\vec{v}^{(2)}]_S$$

From now on, we simply omit the subscript  $S$  and write  $\vec{v}$  in place of  $[\vec{v}]_S$  if not specified, to denote vectors in the standard basis as implicitly assumed before.  $\square$

### 6.2.2 Coordinate Bases for Subspaces of $\mathbb{R}^n$

Now that we are able to construct a coordinate basis for  $\mathbb{R}^n$ , it is natural to ask if we can also extend this and come up with some coordinate basis for any subspace of  $\mathbb{R}^n$  (since a subspace is itself a vector space too), in the sense that any vector in the subspace can be uniquely expressed by the basis vectors (*linear independence*) and the basis *spans* the subspace exactly, just like any basis for

---

<sup>13</sup>In contrast, the standard unit vector  $\hat{e}^{(2)}$  with a hat is deemed as a "true" vector with  $[\hat{e}^{(2)}]_S = (0, 1, 0)_S^T$ .  $P_B^S[\hat{e}^{(2)}]_B = [\hat{e}^{(2)}]_S$  and thus

$$[\hat{e}^{(2)}]_B = (P_B^S)^{-1}[\hat{e}^{(2)}]_S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{5}{6} & \frac{1}{6} & \frac{1}{2} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \left(\frac{1}{3}, \frac{1}{6}, -\frac{1}{6}\right)_B^T$$

$\mathbb{R}^n$ . Actually, we have already done this for the column space of a matrix  $A$  back in the derivation of CR Factorization in Section 6.1.4 where we created a minimal generating set from the column vectors that compose  $A$ . For other subspaces of  $\mathbb{R}^n$  the procedure is similar. If we are given a subspace as a span of some vectors, then to find a basis for it, we carry out CR Factorization as if these vectors are the columns of a matrix and retain the linearly independent vectors. These linearly independent vectors still span the subspace by part (b) of Theorem 6.1.11, and any vector in the subspace can be written as a unique linear combination of them by Properties 6.1.12.

**Properties 6.2.3.** For the subspace generated by a spanning set  $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(q)}\}$ , its basis can be found by applying CR Factorization (Properties 6.1.13) over  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \dots | \vec{v}^{(q)}]$ . Then a possible basis is  $\{\vec{v}^{(j)}\}$  for all  $j$  such that the  $j$ -th column of rref of  $A$  contains a leading 1.

**Example 6.2.2.** Find a basis  $\mathcal{B}$  for the subspace generated by  $\mathcal{G} = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\}$ . Hence express  $(3, -1, 4, 0)^T$  in this basis.

*Solution.* By Properties 6.2.3, we need to find the CR Factorization for

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

We proceed with Gaussian Elimination.

$$\begin{array}{c} \left[ \begin{array}{cccc} 1 & 1 & 0 & -2 \\ 0 & -1 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & -2 & 0 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{cccc} 1 & 1 & 0 & -2 \\ 0 & -1 & 0 & 1 \\ 0 & -1 & 1 & 4 \\ 0 & -3 & 0 & 3 \end{array} \right] \quad R_3 - 2R_1 \rightarrow R_3 \\ \qquad \qquad \qquad R_4 - R_1 \rightarrow R_4 \\ \rightarrow \left[ \begin{array}{cccc} 1 & 1 & 0 & -2 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 1 & 4 \\ 0 & -3 & 0 & 3 \end{array} \right] \quad -R_2 \rightarrow R_2 \end{array}$$

$$\begin{aligned} &\rightarrow \left[ \begin{array}{cccc} 1 & 1 & 0 & -2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cccc} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_4 + 3R_2 \rightarrow R_4 \\ &\quad \quad \quad R_1 - R_2 \rightarrow R_1 \end{aligned}$$

So a possible basis for the subspace:  $\text{span}(\mathcal{G})$ , contains the first three generating vectors in  $\mathcal{G}$ , so that  $\mathcal{B} = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T\}$ . To express  $\vec{v} = (3, -1, 4, 0)^T$  in this basis, we need to find  $[\vec{v}]_{\mathcal{B}}$  just like in Example 6.2.1, which is derived by

$$\left[ \begin{array}{ccc} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 2 & 1 & 1 \\ 1 & -2 & 0 \end{array} \right] \left[ \begin{array}{c} [v_1]_{\mathcal{B}} \\ [v_2]_{\mathcal{B}} \\ [v_3]_{\mathcal{B}} \end{array} \right] = \left[ \begin{array}{c} 3 \\ -1 \\ 4 \\ 0 \end{array} \right]$$

We repeat the same steps of Gaussian Elimination over the augmented matrix where now the left portion consisted of the three linearly independent basis vectors only.

$$\begin{aligned} \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -1 \\ 2 & 1 & 1 & 4 \\ 1 & -2 & 0 & 0 \end{array} \right] &\rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -1 \\ 0 & -1 & 1 & -2 \\ 0 & -3 & 0 & -3 \end{array} \right] \quad R_3 - 2R_1 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 1 & -2 \\ 0 & -3 & 0 & -3 \end{array} \right] \quad R_4 - R_1 \rightarrow R_4 \\ &\rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad -R_2 \rightarrow R_2 \\ &\rightarrow \left[ \begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3 \\ &\quad \quad \quad R_4 + 3R_2 \rightarrow R_4 \end{aligned}$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_1 - R_2 \rightarrow R_1$$

The last full-zero row is consistent and  $[\vec{v}]_B = ([v_1]_B, [v_2]_B, [v_3]_B)_B^T = (2, 1, -1)_B^T$ . As a final note, since the fourth vector  $(-2, 1, 0, 1)^T$  in the generating set can be written as a non-zero linear combination of the first three vectors (with the coefficients of  $-1, -1, 3$ ), we can replace any one of the three vectors in the basis by  $(-2, 1, 0, 1)^T$ .  $\square$

In general, any vector  $\vec{v}^{(j)}$  in a basis can be replaced by another vector that is the linear combination of the basis vectors where the coefficient corresponding to  $\vec{v}^{(j)}$  is particularly not zero.<sup>14</sup> Moreover, we can now properly define the "dimension" of any subspace of  $\mathbb{R}^n$ . It is simply the number of (linearly independent) vectors in its basis. Some may wonder if it is possible for two bases of the same vector space to have different number of vectors so that the notion of its dimension will be problematic. In fact, all bases of a *finite-dimensional* vector (sub)space must possess the same amount of vectors, and we simply note this below.<sup>15</sup>

**Properties 6.2.4.** If  $\mathcal{V}$  is a vector space with a finite basis, then all bases of  $\mathcal{V}$  are finite and have the same number of vectors.

From the statement above, we see that if we can find any basis with exactly  $n$  vectors for a vector space  $\mathcal{V}$  where  $n$  is finite, then  $n$  will be the unique integer such that every basis of  $\mathcal{V}$  is consisted of this number of vectors.  $n$  is then referred to as the **dimension** of  $\mathcal{V}$ , and we define  $\dim(\mathcal{V}) = n$ .  $\mathcal{V}$  is then known as a **finite-dimensional** vector space. If a vector space is not finite-dimensional, i.e. a finite basis cannot be found, then it is called **infinite-dimensional**. Moreover,

<sup>14</sup>This preserves the span and linear independence. Refer to Properties 6.3.4 later for the span part (where the vectors are to be viewed in as rows and the replacement is effectively multiplications and additions of rows), plus (a) of Properties 6.2.7 for linear independence.

<sup>15</sup>It, along with other results below, actually comes from a more general theorem called the *Steinitz Replacement Theorem* which is proved in the Appendix.

**Properties 6.2.5.** For any subspace  $\mathcal{W}$  of a vector space  $\mathcal{V}$ ,  $\dim(\mathcal{W}) \leq \dim(\mathcal{V})$ . If  $\dim(\mathcal{W}) = \dim(\mathcal{V})$ ,  $\mathcal{W} = \mathcal{V}$ .

and

**Theorem 6.2.6.** If a vector space  $\mathcal{V}$  is generated by a spanning set  $\mathcal{G}$  with a finite amount of vectors, then some subset of  $\mathcal{G}$  is a basis for  $\mathcal{V}$ , and  $\mathcal{V}$  has finite bases.

which is a broader restatement of Properties 6.2.3.

Finally, we expand Theorem 6.2.2 (Equivalent requirements of a basis) to any finite-dimensional vector (sub)space. The results are simply stated below.

**Properties 6.2.7.** If  $\mathcal{V}$  is a vector space with  $\dim(\mathcal{V}) = n$ , then

- (a) Any generating set for  $\mathcal{V}$  contains at least  $n$  vectors. If, furthermore, it is made of exactly  $n$  vectors, then it will be a basis for  $\mathcal{V}$ ;
- (b) Any linearly independent subset of  $\mathcal{V}$  that has exactly  $n$  vectors is a basis for  $\mathcal{V}$ ;
- (c) Every linearly independent subset  $\mathcal{G}_1$  of  $\mathcal{V}$  with  $m \leq n$  vectors can be extended to a basis for  $\mathcal{V}$ , i.e. there exists another subset  $\mathcal{G}_2$  of  $\mathcal{V}$  with  $n - m$  (linearly independent) vectors such that  $\mathcal{B} = \mathcal{G}_1 \cup \mathcal{G}_2$  is a basis for  $\mathcal{V}$ .

A point worth mentioning is that part (c) of the properties above allows the possibility of completing a basis from its fragment, which will be used in many arguments from time to time.

### 6.2.3 Direct Sum Representation

Since we can create subspaces from multiple individual vectors, we may like to know if we can go one step further and make a larger vector space from smaller subspaces by composing them together. This then leads to the *direct sum* representation. Let's begin with the definition of *sum of subspaces* first.

**Definition 6.2.8** (Subspace Sum). Given two subspaces  $\mathcal{W}_1, \mathcal{W}_2$ , of a vector space  $\mathcal{V}$ , their subspace sum is

$$\mathcal{W}_1 + \mathcal{W}_2 = \{\vec{w}_1 + \vec{w}_2 \mid \text{all } \vec{w}_1 \in \mathcal{W}_1, \vec{w}_2 \in \mathcal{W}_2\}$$

consisted of all possible vectors resulted from addition between any pair of vectors from  $\mathcal{W}_1, \mathcal{W}_2 \subseteq \mathcal{V}$  respectively. Note that  $(\mathcal{W}_1 + \mathcal{W}_2) \subseteq \mathcal{V}$  is a subspace of  $\mathcal{V}$ .

For example, if  $\mathcal{W}_1 = \text{span}(\{(1, 0, 1)^T\})$  and  $\mathcal{W}_2 = \text{span}(\{(1, 1, 0)^T, (0, 1, 1)^T\})$ , then according to the definition of spans in Definition 6.1.5 and that of a subspace sum above,  $\mathcal{W}_1 + \mathcal{W}_2 = \text{span}(\{(1, 0, 1)^T, (1, 1, 0)^T, (0, 1, 1)^T\})$ , which is just the span of the union of generating vectors from the two smaller spans, and can be shown to be equal to  $\mathbb{R}^3$  following the same idea used when doing Example 6.2.1. Extending this, we have

$$\mathcal{W}_1 + \mathcal{W}_2 + \cdots + \mathcal{W}_n = \{\vec{w}_1 + \vec{w}_2 + \cdots + \vec{w}_n \mid \vec{w}_j \in \mathcal{W}_j, 1 \leq j \leq n\}$$

In the small example above,  $\dim(\mathcal{W}_1) + \dim(\mathcal{W}_2) = 1 + 2 = 3 = \dim(\mathcal{W}_1 + \mathcal{W}_2)$ , as the spanning vectors collected from the two subspaces are linearly independent of each other, i.e. the basis vector in  $\mathcal{W}_1$  cannot be expressed as the linear combination of those in  $\mathcal{W}_2$  and vice versa. In this case, the dimensions of the two subspaces can be *directly* added together, and hence it constitutes a **direct sum**, whose requirement is given below.

**Definition 6.2.9** (Direct Sum). A direct sum between two subspaces  $\mathcal{W}_1, \mathcal{W}_2$  is their subspace sum  $\mathcal{W}_1 + \mathcal{W}_2$  as defined in Definition 6.2.8 which additionally satisfies  $\mathcal{W}_1 \cap \mathcal{W}_2 = \{\mathbf{0}\}$ , and is denoted as  $\mathcal{W}_1 \oplus \mathcal{W}_2$ , and we have  $\dim(\mathcal{W}_1 \oplus \mathcal{W}_2) = \dim(\mathcal{W}_1) + \dim(\mathcal{W}_2)$ .

Here we show that the condition of  $\mathcal{W}_1 \cap \mathcal{W}_2 = \{\mathbf{0}\}$  is equivalent to the above condition that the basis vectors from  $\mathcal{W}_1$  and  $\mathcal{W}_2$  combined are linearly independent (sometimes we also simply say that the subspaces  $\mathcal{W}_1$  and  $\mathcal{W}_2$  are linearly independent). Let  $\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(p)}$  and  $\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(q)}$  be the basis vectors for  $\mathcal{W}_1$  and  $\mathcal{W}_2$  respectively. If these basis vectors are linearly

independent, then by Theorem 6.1.9, the equation

$$c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \cdots + c_p\vec{u}^{(p)} + c_{p+1}\vec{v}^{(1)} + \cdots + c_{p+q}\vec{v}^{(q)} = \mathbf{0}$$

only has  $c_j = 0$  as the trivial solution,  $1 \leq j \leq p+q$ . Rearranging, we have

$$\begin{aligned} c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \cdots + c_p\vec{u}^{(p)} &\in \mathcal{W}_1 \\ = -c_{p+1}\vec{v}^{(1)} - \cdots - c_{p+q}\vec{v}^{(q)} &\in \mathcal{W}_2 \end{aligned}$$

But since  $c_j = 0$  is the only solution to this, it shows that there is only the zero vector in both  $\mathcal{W}_1$  and  $\mathcal{W}_2$  at the same time. The converse essentially follows the same argument in reverse. We say that  $\mathcal{W}_1$  and  $\mathcal{W}_2$  are a (*subspace complement*<sup>16</sup>) to each other in  $\mathcal{W}_1 \oplus \mathcal{W}_2$ , as any non-zero vector from  $\mathcal{W}_1 \oplus \mathcal{W}_2$  that has purely zero components in one of the two smaller subspaces will always be found in another.

**Properties 6.2.10** (Subspace Complement). If a vector space can be written as a direct sum of two smaller subspaces, i.e.  $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$ , then  $\mathcal{W}_1 = \mathcal{W}_2^C$  and  $\mathcal{W}_2 = \mathcal{W}_1^C$  are said to be the (subspace) complement (denoted by the superscript  $C$ ) to each other in  $\mathcal{V}$ .

As a counter-example, consider Example 6.2.2, suppose  $\mathcal{W}_1 = \text{span}(\mathcal{B}_1) = \text{span}(\{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T\})$  and  $\mathcal{W}_2 = \text{span}(\mathcal{B}_2) = \text{span}(\{(0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\})$  be the subspaces spanned the first/last two vectors in  $\mathcal{G}$  respectively. It is not hard to see that  $\mathcal{B}_1$  and  $\mathcal{B}_2$  are themselves linearly independent (and hence they are bases for  $\mathcal{W}_1$  and  $\mathcal{W}_2$  individually), and  $\dim(\mathcal{W}_1) = \dim(\mathcal{W}_2) = 2$ . Nevertheless, in that example, we already know that the four vectors, when put together, are not linearly independent:  $(-2, 1, 0, 1)^T$  is equal to  $(1, 0, 2, 1)^T - (1, -1, 1, -2)^T + 3(0, 0, 1, 0)^T$ , and they only generates a three-dimensional subspace. Hence  $\dim(\mathcal{W}_1 + \mathcal{W}_2) = \dim(\text{span}(\mathcal{B}_1) + \text{span}(\mathcal{B}_2)) = \dim(\text{span}(\mathcal{G})) = 3 \neq 4 = 2 + 2 = \dim(\mathcal{W}_1) + \dim(\mathcal{W}_2)$ , and therefore they cannot form a direct sum. Geometrically, these two subspaces are like two planes intersecting along a straight line.

---

<sup>16</sup>A subspace complement is not the same as the set-theoretic complement.

The direct sum of multiple subspaces are then recursively defined as

$$\begin{aligned}\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3 \oplus \cdots \oplus \mathcal{W}_{n-1} \oplus \mathcal{W}_n \\ = (\cdots ((\mathcal{W}_1 \oplus \mathcal{W}_2) \oplus \mathcal{W}_3) \oplus \cdots \oplus \mathcal{W}_{n-1}) \oplus \mathcal{W}_n\end{aligned}$$

where we add up the subspaces one by one. Below shows an example of this.

**Example 6.2.3.** Given  $\mathcal{W}_1 = \text{span}\{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T\}$ ,  $\mathcal{W}_2 = \text{span}\{(0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\}$ ,  $\mathcal{W}_3 = \text{span}\{(1, 1, -3, 0, -1)^T\}$ , show that  $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3$  is a valid direct sum which equals to  $\mathbb{R}^5$ .

*Solution.* First, let's derive  $\mathcal{W}_1 \oplus \mathcal{W}_2$ . It is obvious that the two generating vectors from each of  $\mathcal{W}_1$  and  $\mathcal{W}_2$  are linearly independent themselves as they are not constant multiples of another. Now following similar ideas in Example 6.2.2, we are going to show that every column in the matrix formed by combining basis vectors of both  $\mathcal{W}_1$  and  $\mathcal{W}_2$

$$\left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 2 & 0 & 1 & -1 \\ 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 1 \end{array} \right]$$

is pivotal after Gaussian Elimination, as follows.

$$\begin{array}{c} \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 2 & 0 & 1 & -1 \\ 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & -4 & 1 & -1 \\ 0 & -2 & 0 & -2 \\ 0 & -1 & 0 & 1 \end{array} \right] \\ R_3 - 2R_1 \rightarrow R_3 \\ R_4 - R_1 \rightarrow R_4 \\ \\ \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 13 & -1 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 3 & 1 \end{array} \right] \\ R_3 + 4R_2 \rightarrow R_3 \\ R_4 + 2R_2 \rightarrow R_4 \\ R_5 + R_2 \rightarrow R_5 \end{array}$$

$$\begin{aligned}
 & \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 13 & -1 \end{array} \right] & R_3 \leftrightarrow R_5 \\
 & \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 13 & -1 \end{array} \right] & \frac{1}{3}R_3 \rightarrow R_3 \\
 & \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & -4 \\ 0 & 0 & 0 & -\frac{16}{3} \end{array} \right] & R_4 - 6R_3 \rightarrow R_4 \\
 & \quad R_5 - 13R_3 \rightarrow R_5 \\
 & \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{16}{3} \end{array} \right] & -\frac{1}{4}R_4 \rightarrow R_4 \\
 & \rightarrow \left[ \begin{array}{cccc} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right] & R_5 + \frac{16}{3}R_4 \rightarrow R_5
 \end{aligned}$$

and we are done (the backward phase is optional for this). Therefore, the four column vectors are linearly independent when considered as a whole and the direct sum  $\mathcal{W}_1 \oplus \mathcal{W}_2 = \text{span}(\{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T, (0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\})$  makes sense, with  $\dim(\mathcal{W}_1 \oplus \mathcal{W}_2) = \dim(\mathcal{W}_1) + \dim(\mathcal{W}_2) = 2 + 2 = 4$ ,  $\mathcal{W}_1 \oplus \mathcal{W}_2 \subset \mathbb{R}^5$ . Now, we attempt to compose  $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3 = (\mathcal{W}_1 \oplus \mathcal{W}_2) \oplus \mathcal{W}_3$ , which requires showing that the only generating vector  $(1, 1, -3, 0, -1)^T$  in  $\mathcal{W}_3$  is linearly independent from (the basis vectors of)  $\mathcal{W}_1 \oplus \mathcal{W}_2$ . One way to do this is to show that the augmented system formed by

appending  $(1, 1, -3, 0, -1)^T$  to the matrix at the start

$$\left[ \begin{array}{ccccc|c} 1 & 2 & 0 & 0 & 1 \\ 0 & 1 & 3 & 0 & 1 \\ 2 & 0 & 1 & -1 & -3 \\ 1 & 0 & 0 & -2 & 0 \\ 0 & -1 & 0 & 1 & -1 \end{array} \right]$$

has no solution and thus  $(1, 1, -3, 0, -1)^T$  cannot be written as their linear combination (see part (a) of Theorem 6.1.11). We can simply repeat the exactly same reduction steps performed above, which would lead to

$$\left[ \begin{array}{ccccc|c} 1 & 2 & 0 & 0 & 1 \\ 0 & 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 & -\frac{1}{4} \\ 0 & 0 & 0 & 0 & -\frac{1}{3} \end{array} \right]$$

where the last row is inconsistent. Therefore  $(1, 1, -3, 0, -1)^T$  is linearly independent from the preceding four vectors and  $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3$  is a valid direct sum, and  $\dim(\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3) = \dim(\mathcal{W}_1 \oplus \mathcal{W}_2) + \dim(\mathcal{W}_3) = 4 + 1 = 5$ . By Properties 6.2.5,  $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3 = \mathbb{R}^5$ .  $\square$

The importance of direct sum is that the coordinates of two vectors in respective bases from the two subspaces can be simply concatenated when we add up both the vectors and bases, and *this representation will be unique*. Going in the opposite direction, we can also "split" the coordinates of a direct sum back into the respective subspaces. Let's illustrate this with  $\mathcal{W}_1$  and  $\mathcal{W}_2$  in the above example. Using the given sets of generating vectors  $\mathcal{B}_1 = \{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T\}$  and  $\mathcal{B}_2 = \{(0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\}$  as bases for  $\mathcal{W}_1$  and  $\mathcal{W}_2$ , the coordinates  $(1, 2)_{\mathcal{B}_1}^T$  and  $(1, -1)_{\mathcal{B}_2}^T$  in the  $\mathcal{B}_1$  and  $\mathcal{B}_2$  system, represent the vectors

$$(1, 0, 2, 1, 0)^T + 2(2, 1, 0, 0, -1)^T = (5, 2, 2, 1, -2)^T$$

and  $(0, 3, 1, 0, 0)^T - (0, 0, -1, -2, 1)^T = (0, 3, 2, 2, -1)^T$

in  $\mathbb{R}^5$  respectively. When they are summed, it yields  $(5, 2, 2, 1, -2)^T + (0, 3, 2, 2, -1)^T = (5, 5, 4, 3, -3)^T$ . The basis formed by combining  $\mathcal{B}_1$  and  $\mathcal{B}_2$  will be

$$\mathcal{B}_1 \cup \mathcal{B}_2 = \{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T, (0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\}$$

and the merged coordinates  $(1, 2, 1, -1)_{B_1+B_2}^T$  then correspond exactly to

$$\begin{aligned} & (1, 0, 2, 1, 0)^T + 2(2, 1, 0, 0, -1)^T + (0, 3, 1, 0, 0)^T - (0, 0, -1, -2, 1)^T \\ &= (5, 5, 4, 3, -3)^T \in \mathcal{W}_1 \oplus \mathcal{W}_2 \subset \mathbb{R}^5 \end{aligned}$$

The new coordinate representation  $(1, 2, 1, -1)_{B_1+B_2}^T$  is unique as  $\mathcal{B}_1 \cup \mathcal{B}_2$  has been shown to be linearly independent in Example 6.2.3 and Properties 6.1.12 applies over the direct sum  $\mathcal{W}_1 \oplus \mathcal{W}_2$ , and it can be partitioned cleanly as  $(1, 2, 1, -1)_{B_1+B_2}^T = (1, 2)_{B_1}^T + (1, -1)_{B_2}^T$ .

On the other hand, the uniqueness property will not hold if the subspace sum is not a direct sum. Let's use Example 6.2.2 again as a demonstration, where  $\mathcal{B}_1 = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T\}$  and  $\mathcal{B}_2 = \{(0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\}$  and we have already shown that they are not linearly independent when combined. Take

$$\begin{aligned} (1, 2)_{B_1}^T &= (1, 0, 2, 1)^T + 2(1, -1, 1, -2)^T = (3, -2, 4, -3)^T \\ \text{and } (-1, 1)_{B_2}^T &= -(0, 0, 1, 0)^T + (-2, 1, 0, 1)^T = (-2, 1, -1, 1)^T \end{aligned}$$

Their concatenated sum will be

$$\begin{aligned} & (1, 2, -1, 1)_{B_1+B_2}^T \\ &= (1, 0, 2, 1)^T + 2(1, -1, 1, -2)^T - (0, 0, 1, 0)^T + (-2, 1, 0, 1)^T \\ &= (3, -2, 4, -3)^T + (-2, 1, -1, 1)^T \\ &= (1, -1, 3, -2)^T \end{aligned}$$

but

$$(0, 1, 2, 0)_{B_1+B_2}^T = (1, -1, 1, -2)^T + 2(0, 0, 1, 0)^T$$

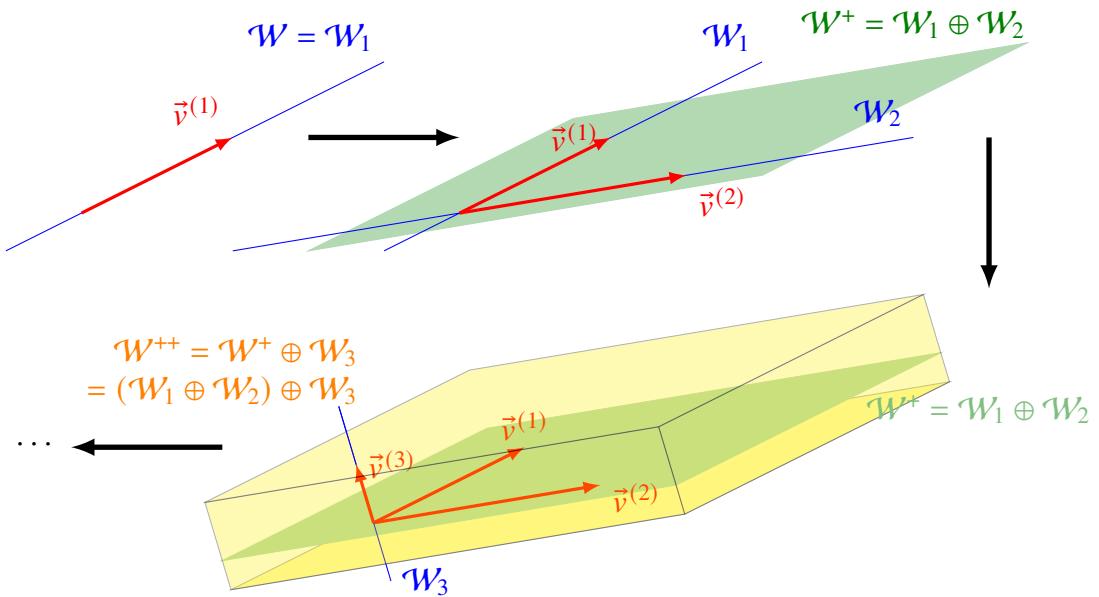


Figure 6.2: Iteratively adding one-dimensional subspaces to the direct sum. Note that the lines, plane and "cuboid" all extend infinitely. We can only visualize up to a three-dimensional direct sum but it goes on even for higher dimensions.

$$= (1, -1, 3, -2)^T = (1, 2, -1, 1)^T_{B_1+B_2}$$

is aptly an alternative representation.

Another aspect about direct sum is that all finite-dimensional, particularly  $n$ -dimensional vector spaces with some basis consisted of  $n$  vectors can be regarded to be a direct sum of the  $n$  one-dimensional subspaces generated by each of these basis vectors individually. We have provided a schematic (Figure 6.2) to better illustrate this.

## 6.3 The Four Fundamental Subspaces Induced by Matrices

### 6.3.1 Row Space, Column Space

In Definition 6.1.7, we have developed the notion of column space. For a  $m \times n$  matrix  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \cdots | \vec{v}^{(n)}]$ , its column space is the subspace generated by the  $n$   $\mathbb{R}^m$  vectors  $\vec{v}^{(j)}$ ,  $j = 1, 2, \dots, n$ . Similarly, we can also define the **row space** of a matrix. We formally define both of them below.

**Definition 6.3.1** (Column/Row Space). For an  $m \times n$  real matrix  $A$ , its column space  $C(A)$  is the subspace spanned by its  $n$  column vectors,  $\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)} \in \mathbb{R}^m$  as in  $A = [\vec{v}^{(1)} | \vec{v}^{(2)} | \cdots | \vec{v}^{(n)}]$ ; Meanwhile its row space is the subspace spanned by its  $m$  row vectors  $\vec{w}^{(1)T}, \vec{w}^{(2)T}, \dots, \vec{w}^{(m)T} \in \mathbb{R}^n$  as in

$$A = \begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \hline \vdots \\ \hline \vec{w}^{(m)T} \end{bmatrix}$$

Notice that the row (column) space of a matrix is just the column (row) space of its transpose, hence we denote the row space of  $A$  as  $C(A^T)$ .

For instance, in Example 6.2.2, the matrix

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

actually has a column space of  $C(A) = \text{span}(\{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\}) = \text{span}(\{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T\})$  of dimension 3 despite the vectors are in  $\mathbb{R}^4$ . In middle of deriving this result

we have produced the reduced row echelon form of  $A$ , which is

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

from which we can see the number of pivots, or *rank*, is also 3. In fact, just like the case above, the *rank* of a matrix always indicates the dimension of its column space. This is due to Properties 6.2.3 and 6.2.4, leading to the following equivalent definition.

**Definition 6.3.2** (Rank). The rank of a matrix  $A$  is the number of leading 1s in its reduced row echelon form, which is also the amount of linearly independent vectors in any basis of its column space, i.e. the dimension of the column space.

We can approach this from another angle, which involves restating previous results related to Gaussian Elimination. In Section 6.1.4, we have shown that elementary row operations preserve (the amount of) linear independent vectors, hence

**Properties 6.3.3.** Elementary row operations does not change the number of dimensions in the column space of a matrix.

The matrix  $A$  will then have the same number of dimensions in its column space throughout the Gaussian Elimination procedure, which coincides with the number of linearly independent vectors and thus pivots in the final reduced row echelon form, establishing the equivalence in Definition 6.3.2. However, notice that elementary row operations do change the actual column space. On the other hand, for row space, we have an even stronger result.

**Properties 6.3.4.** Elementary row operations does not change the row space of a matrix, and thus its dimension.

which is not hard to accept. Swapping rows, and multiplying a row by some constant obviously does not affect the span of rows in the matrix. Adding

to/subtracting from a row  $R_p$  (also as a row vector  $\vec{w}^{(p)T}$ ) by the constant multiple of another row  $R_q$  ( $\vec{w}^{(q)T}$ ) also will not alter it. To see this, observe that the newly resulted row vector is just a linear combination of the two input rows, i.e. the new  $R_p$  becomes  $\vec{w}^{(r)T} = \vec{w}^{(p)T} + c\vec{w}^{(q)T}$  (and hence  $\vec{w}^{(p)T} = \vec{w}^{(r)T} - c\vec{w}^{(q)T}$ ). Using part (b) of Theorem 6.1.11 twice, we have

$$\begin{aligned} &= \text{span}(\{\dots, \vec{w}^{(p)}, \dots, \vec{w}^{(q)}, \dots, \vec{w}^{(r)}\}) \\ C(A^T) &= \text{span}(\{\dots, \vec{w}^{(p)}, \dots, \vec{w}^{(q)}, \dots\}) \\ &= \text{span}(\{\dots, \vec{w}^{(r)}, \dots, \vec{w}^{(q)}, \dots\}) = C(A'^T) \end{aligned}$$

where  $A'$  denotes the matrix after the addition/subtraction elementary row operation. Our next key result relies on the observation that the dimensions of row and column space of a matrix in its reduced row echelon form are the same, or in other words,

**Properties 6.3.5.** A matrix in reduced row echelon form has the same amount of (linearly independent) vectors in the basis of its row and column space.

We will not read off the detailed arguments in the proof, but instead note that it is essentially an analysis of positions of the leading 1s and zeros in any reduced row echelon form. However, we will give an example to elucidate how it holds. Take a reduced row echelon form of

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

It is obvious that its column space is spanned by the basis  $\{(1, 0, 0, 0)^T, (0, 1, 0, 0)^T, (0, 0, 1, 0)^T\}$ , while a basis of its row space can be simply formed by the first three non-zero row vectors  $\{(1, 1, 0, 0, 1), (0, 0, 1, 0, 0), (0, 0, 0, 1, 1)\}$ . In this case, the dimension of row/column space of the reduced row echelon form is both 3. With these observations, we can derive the desired result, sometimes referred to as "*Column rank equals to row rank*".

**Properties 6.3.6.** For any matrix, the dimension of its column space is equal to that of its row space, i.e.

$$\dim(C(A)) = \dim(C(A^T))$$

*Proof.* Any matrix has a unique reduced row echelon form due to Theorem 2.2.6, whose row/column space has the same number of dimensions by Properties 6.3.5. According to Properties 6.3.3 and 6.3.4, the elementary row operations done to convert the matrix to its reduced row echelon form leave both the dimensions of row and column space conserved, and thus the column rank and row rank in the starting matrix are equal.  $\square$

**Example 6.3.1.** Given a matrix

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

find a basis for its column/row space  $C(A)$  and  $C(A^T)$  and check if Properties 6.3.6 holds.

*Solution.* We first apply Gaussian Elimination to  $A$ , which leads to

$$\begin{aligned} \left[ \begin{array}{cccc} 1 & 1 & -2 & 1 \\ 1 & 2 & 1 & -1 \\ 1 & 0 & -5 & 3 \end{array} \right] &\rightarrow \left[ \begin{array}{cccc} 1 & 1 & -2 & 1 \\ 0 & 1 & 3 & -2 \\ 0 & -1 & -3 & 2 \end{array} \right] && R_2 - R_1 \rightarrow R_2 \\ &\rightarrow \left[ \begin{array}{cccc} 1 & 1 & -2 & 1 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right] && R_3 - R_1 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cccc} 1 & 0 & -5 & 3 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right] && R_3 - R_2 \rightarrow R_3 \\ &\rightarrow \left[ \begin{array}{cccc} 1 & 0 & -5 & 3 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right] && R_1 - R_2 \rightarrow R_1 \end{aligned}$$

### 6.3 The Four Fundamental Subspaces Induced by Matrices

---

The number of pivotal columns is 2, and from the reduced row echelon form we obtain the dependence relations where the first two column vectors  $(1, 1, 1)^T$  and  $(1, 2, 0)^T$  are linearly independent while the last two column vectors  $(-2, 1, -5)^T = -5(1, 1, 1)^T + 3(1, 2, 0)^T$  and  $(1, -1, 3)^T = 3(1, 1, 1)^T - 2(1, 2, 0)^T$  are linear combinations of the previous two. Hence  $C(A)$  has a basis of  $\{(1, 1, 1)^T, (1, 2, 0)^T\}$  and  $\dim(C(A)) = 2$ . On the other hand, to find the row space we consider  $A^T$  and repeat the elimination process again as follows. However, notice that according to the dependence relations for the column vectors in  $A$  above, we can immediately do the corresponding addition/subtraction operations for the rows in  $A^T$ , to reduce the third/fourth rows, obtaining

$$\left[ \begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & 0 \\ -2 & 1 & -5 \\ 1 & -1 & 3 \end{array} \right] \rightarrow \left[ \begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \quad R_3 + 5R_1 - 3R_2 \rightarrow R_3, \\ R_4 - 3R_1 + 2R_2 \rightarrow R_4$$

and the next step is straight-forward:

$$\left[ \begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \quad R_2 - R_1 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{ccc} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \quad R_1 - R_2 \rightarrow R_1$$

which reveals that the first two columns (representing the first two row vectors in  $A$ ) are linearly independent and the third column (the last row vector in  $A$ ) is redundant ( $(1, 0, -5, 3)^T = 2(1, 1, -2, 1)^T - (1, 2, 1, -1)^T$ ). Therefore  $C(A^T)$  has a basis of  $\{(1, 1, -2, 1)^T, (1, 2, 1, -1)^T\}$ , and  $\dim(C(A^T)) = 2 = \dim(C(A))$ , and Properties 6.3.6 is true in this case.  $\square$

Finally, in view of Definitions 6.1.5 and 6.3.1, the analysis about solving linear systems in Section 3.2 can be summarized as

**Properties 6.3.7.** A linear system  $A\vec{x} = \vec{h}$  is consistent if and only if  $\vec{h}$  is in the column space of  $A$ .

### 6.3.2 Null Space, Rank-Nullity Theorem

As we have briefly mentioned in the end of last chapter, the solution of a linear system  $A\vec{x} = \vec{h}$ , where  $A$  is an  $m \times n$  matrix and  $\vec{x} \in \mathbb{R}^n$ , can be viewed as some sort of a solution space. In Section 3.2.1 we know that it is made up of the particular and complementary solution, where the latter corresponding to the family of  $\vec{x} = \vec{x}_0$  ( $= \vec{x}_c$  using the notation in that section) that satisfies the homogeneous part  $A\vec{x} = \mathbf{0}$ . The set  $\vec{x}_0 \in \mathbb{R}^n$  can be shown to form a subspace of  $\mathbb{R}^n$ <sup>17</sup>, and this subspace is then called the **null space** of  $A$ .

**Definition 6.3.8** (Null Space). For an  $m \times n$  real matrix  $A$ , its null space  $N(A)$  is the subspace consisted of all solution vectors  $\vec{x} = \vec{x}_0 \in \mathbb{R}^n$  to the matrix equation  $A\vec{x} = \mathbf{0}$ . The dimension of null space is called **nullity**.

This definition of nullity as the dimension of null space is consistent with that in Section 3.2.1 where nullity is initially given by the number of columns in the matrix minus the amount of leading 1s (rank) in its rref, or equivalently the number of non-pivotal columns. To see this, observe that any solution  $\vec{x}_0$  to  $A\vec{x} = \mathbf{0}$  is also the solution to  $A_{\text{rref}}\vec{x} = \mathbf{0}$  and vice versa, via elementary matrices. Hence the null space and nullity of  $A$  will be the same as that of  $A_{\text{rref}}$ . Previously we have assigned free variables to the non-pivotal columns (let's say there is  $k$  of them) of  $A_{\text{rref}}$  and derive  $\vec{x}_0$  where they are generated by  $k$  pairs of free variables and column vectors  $(\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \dots, \vec{x}_0^{(k)})$ . It is clear that such a procedure will ensure these  $k$  vectors are linearly independent as each of them has a component of 1 in the position corresponding to that particular free variable indicated by the rref and 0s in other positions corresponding to other

---

<sup>17</sup>To show this we check the two conditions in Theorem 6.1.2. Let  $\vec{x}_0^{(1)}$  and  $\vec{x}_0^{(2)}$  be two vectors in the null space  $\vec{x}_0$ . Then we have: 1.  $A(\vec{x}_0^{(1)} + \vec{x}_0^{(2)}) = A\vec{x}_0^{(1)} + A\vec{x}_0^{(2)} = \mathbf{0} + \mathbf{0} = \mathbf{0}$ , so  $\vec{x}_0^{(1)} + \vec{x}_0^{(2)} \in \vec{x}_0$ , and 2.  $A(a\vec{x}_0^{(1)}) = a(A\vec{x}_0^{(1)}) = a\mathbf{0} = \mathbf{0}$ , hence  $a\vec{x}_0^{(1)} \in \vec{x}_0$ .

### 6.3 The Four Fundamental Subspaces Induced by Matrices

free variables (see Example 3.2.6 for an instance). We claim that they also span the entire null space of  $A_{\text{ref}}$ .<sup>18</sup> Hence by the definition given in Section 6.2.2 they form a basis for the null space of  $A_{\text{ref}}$  as well as  $A$  and by Properties 6.2.4 the dimension of null space of  $A$  is also  $k$ .

Using Definitions 6.3.1, 6.3.2, and 6.3.8 to rephrase, the preceding discussion means that the rank of a matrix plus its nullity equals to its number of columns, which leads to the so-called **Rank-nullity Theorem**.

**Theorem 6.3.9** (Rank-nullity Theorem). For a real  $m \times n$  matrix  $A$ , we have

$$\begin{aligned}\dim(C(A)) + \dim(\mathcal{N}(A)) &= \text{rank}(A) + \text{nullity}(A) = n \\ &= \dim(C(A^T)) + \dim(\mathcal{N}(A))\end{aligned}$$

where  $\dim(C(A)) = \dim(C(A^T)) = \text{rank}(A)$  by Properties 6.3.6.

An invertible square matrix has a reduced row echelon form of an identity matrix according to Theorem 6.1.10, and since an identity matrix has *full rank*<sup>19</sup> (Definition 6.3.2), by the Rank-nullity Theorem 6.3.9 above, we have

**Properties 6.3.10.** A  $n \times n$  square matrix is invertible if and only if its rank and nullity are  $n$  and 0.

A notable relationship between row space and null space is that any pair of two vectors coming from the respective subspaces will be orthogonal to each other.

<sup>18</sup>Assume the contrary so that the span of  $\{\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \dots, \vec{x}_0^{(k)}\}$  does not cover the whole null space, then the dimension of null space has to be greater than  $k$  by Properties 6.2.5. Without loss of generality, let the "correct" dimension of null space to be  $k+1$ . Then by (c) of Properties 6.2.7, there exists  $\vec{x}_0^{(k+1)}$  such that  $\{\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \dots, \vec{x}_0^{(k)}, \vec{x}_0^{(k+1)}\}$  are basis of the null space. This  $\vec{x}_0^{(k+1)}$  can be made to take the value of 0 at all  $k$  positions where the free variables reside by subtracting it by appropriate multiples of  $\vec{x}_0^{(j)}$ ,  $j \neq k+1$ , without altering the null space (c.f. Footnote 14), and by doing so non-zero components of  $\vec{x}_0^{(k+1)}$  only appear in positions corresponding to leading 1s in  $A_{\text{ref}}$ . This causes a contradiction since  $A_{\text{ref}}\vec{x}_0^{(k+1)} = \mathbf{0}$  then implies that there exists a non-trivial dependence relation between the pivotal column vectors themselves.

<sup>19</sup>An  $m \times n$  matrix  $A$  is said to have full rank if  $\text{rank}(A) = \min(m, n)$

**Properties 6.3.11.** Given a real matrix  $A$ , any vector in its row space  $C(A^T)$  is orthogonal to all vectors in its null space  $N(A)$  and vice versa.

*Proof.* Let the shape of  $A$  be  $m \times n$ , we can express  $A$  in the form of its row vectors as

$$A = \begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \vdots \\ \hline \vec{w}^{(m)T} \end{bmatrix}$$

and the corresponding homogeneous system  $A\vec{x} = \mathbf{0}$  then can be written as

$$A\vec{x} = \begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \vdots \\ \hline \vec{w}^{(m)T} \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{w}^{(1)T} \cdot \vec{x} \\ \hline \vec{w}^{(2)T} \cdot \vec{x} \\ \vdots \\ \hline \vec{w}^{(m)T} \cdot \vec{x} \end{bmatrix} = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where for a solution  $\vec{x} = \vec{x}_0$  in the null space of  $A$ , each of the dot products  $\vec{w}^{(i)T} \cdot \vec{x}_0 = 0$ ,  $i = 1, 2, \dots, m$ , has to be equal to zero. Any vector in the row space of  $A$  can be expressed as  $\vec{w} = c_1\vec{w}^{(1)} + c_2\vec{w}^{(2)} + \dots + c_m\vec{w}^{(m)}$  by Definitions 6.3.1 and 6.1.5, and subsequently, its dot product with  $\vec{x}_0$

$$\begin{aligned} \vec{w}^T \cdot \vec{x}_0 &= (c_1\vec{w}^{(1)} + c_2\vec{w}^{(2)} + \dots + c_m\vec{w}^{(m)})^T \cdot \vec{x}_0 \\ &= c_1(\vec{w}^{(1)T} \cdot \vec{x}_0) + c_2(\vec{w}^{(2)T} \cdot \vec{x}_0) + \dots + c_m(\vec{w}^{(m)T} \cdot \vec{x}_0) \\ &= c_1(0) + c_2(0) + \dots + c_m(0) = 0 \end{aligned}$$

is also zero, therefore they are orthogonal by Properties 4.2.5, which implies that any vector in  $C(A^T)$  is orthogonal to any another vector in  $N(A)$ .  $\square$

As a corollary, this is equivalent to all vectors in the generating set or basis for the row space for a matrix being orthogonal to all vectors in those for its null space. The following additional observation will be useful later.

**Properties 6.3.12.** Non-zero orthogonal vectors are linearly independent.

*Proof.* We will only prove the case with two vectors in  $\mathbb{R}^n$  but those with multiple vectors can be derived in the same essence. Consider  $c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} = \mathbf{0}$  where  $\vec{u}^{(1)}$  and  $\vec{u}^{(2)}$  are orthogonal, i.e.  $\vec{u}^{(1)} \cdot \vec{u}^{(2)} = 0$ . Taking dot product with  $\vec{u}_1$  on both sides gives

$$\begin{aligned}\vec{u}^{(1)} \cdot (c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)}) &= c_1(\vec{u}^{(1)} \cdot \vec{u}^{(1)}) + c_2(\vec{u}^{(1)} \cdot \vec{u}^{(2)}) = \vec{u}^{(1)} \cdot \mathbf{0} \\ c_1\|\vec{u}^{(1)}\|^2 + c_2(0) &= c_1\|\vec{u}^{(1)}\|^2 = 0\end{aligned}$$

Since  $\vec{u}_1$  is non-zero,  $\|\vec{u}_1\|^2 > 0$ , and  $c_1$  must be zero. In a similar vein, we can show that  $c_2$  is zero as well. Therefore the only solution to the equation  $c_1\vec{u}_1 + c_2\vec{u}_2 = \mathbf{0}$  is the trivial solution  $c_1 = c_2 = 0$ . By Theorem 6.1.9, the two vectors are linearly independent.  $\square$

**Example 6.3.2.** For the matrix in Example 6.3.1, find its null space and check if Properties 6.3.11 and Theorem 6.3.9 hold.

*Solution.* The homogeneous system corresponding to the matrix is

$$\left[ \begin{array}{cccc|c} 1 & 1 & -2 & 1 & 0 \\ 1 & 2 & 1 & -1 & 0 \\ 1 & 0 & -5 & 3 & 0 \end{array} \right]$$

which can be reduced, following the same steps in Example 6.3.1, to

$$\left[ \begin{array}{cccc|c} 1 & 0 & -5 & 3 & 0 \\ 0 & 1 & 3 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

where there are two non-pivotal columns and hence two free parameters can be assigned to them. Let  $x_3 = s$  and  $x_4 = t$ , then  $x_1 = 5s - 3t$  and  $x_2 = -3s + 2t$ .

So the solution to the system is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 5s - 3t \\ -3s + 2t \\ s \\ t \end{bmatrix} = s \begin{bmatrix} 5 \\ -3 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -3 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

and thus a basis for the null space is  $\{(5, -3, 1, 0)^T, (-3, 2, 0, 1)^T\}$  where these two vectors are clearly linearly independent (by observing the 0 and 1 of the last two components). As found in Example 6.3.1, the basis for its row space is  $\{(1, 1, -2, 1)^T, (1, 2, 1, -1)^T\}$ . Subsequently, checking orthogonality between the two bases is straight-forward, and we will only do this for the first vector in the row space basis against the null space basis.

$$(5, -3, 1, 0)^T \cdot (1, 1, -2, 1)^T = (5)(1) + (-3)(1) + (1)(-2) + (0)(1) = 0$$

$$(-3, 2, 0, 1)^T \cdot (1, 1, -2, 1)^T = (-3)(1) + (2)(1) + (0)(-2) + (1)(1) = 0$$

Furthermore, the dimension of null space, or the nullity, is  $\dim \mathcal{N}(A) = 2$ . Previously we have also found that  $\dim C(A) = \text{rank}(A) = 2$ . So  $\text{rank}(A) + \text{nullity}(A) = 2 + 2 = 4$ , and Theorem 6.3.9 is true.  $\square$

Short Exercise: Show that <sup>20</sup>

$$\dim(C(A^T)) + \dim(\mathcal{N}(A^T)) = \dim(C(A)) + \dim(\mathcal{N}(A^T)) = m$$

$\mathcal{N}(A^T)$  is also known as the **left null space** of  $A$ .

By Properties 6.3.11 and 6.3.12, vectors in the row space and null space of an  $m \times n$  matrix  $A$  are linearly independent of each other and can form a direct sum  $C(A^T) \oplus \mathcal{N}(A)$  according to Definition 6.2.9. Note that they are the complement to each other with respect to this direct sum according to Properties 6.2.10. Since  $C(A^T) \subseteq \mathbb{R}^n$ ,  $\mathcal{N}(A) \subseteq \mathbb{R}^n$  and hence  $C(A^T) \oplus \mathcal{N}(A) \subseteq \mathbb{R}^n$ , from Theorem 6.3.9 and Properties 6.2.5, we conclude that  $C(A^T) \oplus \mathcal{N}(A)$  which has a dimension of  $n$ , is just  $\mathbb{R}^n$ . In other words, the row space and null space of a matrix can reconstruct the real  $n$ -space by forming their direct sum. The similar can be

---

<sup>20</sup>Replace  $A$  by  $A^T$  in Theorem 6.3.9 to get  $\dim(C(A^T)) + \dim(\mathcal{N}(A^T)) = m$ .

said for its column and left null space. Furthermore, since all vectors in the row space (column space) are orthogonal to those in the (left) null space (and vice versa) via Properties 6.3.11, we say that they are actually an *orthogonal complement* to each other.

**Properties 6.3.13.** For a real  $m \times n$  matrix  $A$ , we have

$$C(A^T) \oplus N(A) = \mathbb{R}^n \quad C(A) \oplus N(A^T) = \mathbb{R}^m$$

where  $C(A^T)^\perp = N(A)$ ,  $N(A)^\perp = C(A^T)$ , and  $C(A)^\perp = N(A^T)$ ,  $N(A^T)^\perp = C(A)$ , with  $^\perp$  denoting an orthogonal complement.

We conclude the relationships between the column, row, null, and left null space, a.k.a the *Four fundamental subspaces* induced by a matrix, with a diagram (Figure 6.3).

Last but not least, we end this chapter with a useful result.

**Properties 6.3.14.** For two (real)  $m \times r$  and  $r \times n$  matrices  $A$  and  $B$ , the rank of  $AB$

$$\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$$

is capped by the smaller of ranks of  $A$  and  $B$ .

*Proof.* The column space of  $AB$  is a subset of that of  $A$ , i.e.  $C(AB) \subseteq C(A)$ , because the columns of  $AB$  can be viewed as

$$\begin{aligned} AB &= [A_1 | A_2 | \cdots | A_r] \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rn} \end{bmatrix} \\ &= [b_{11}A_1 + b_{21}A_2 + \cdots + b_{r1}A_r | \cdots | b_{1n}A_1 + b_{2n}A_2 + \cdots + b_{rn}A_r] \end{aligned}$$

(similar to the explanation for the CR Factorization, where  $A_j$  is the  $j$ -th column of  $A$ ) which shows that the columns of  $AB$  are linear combinations of columns

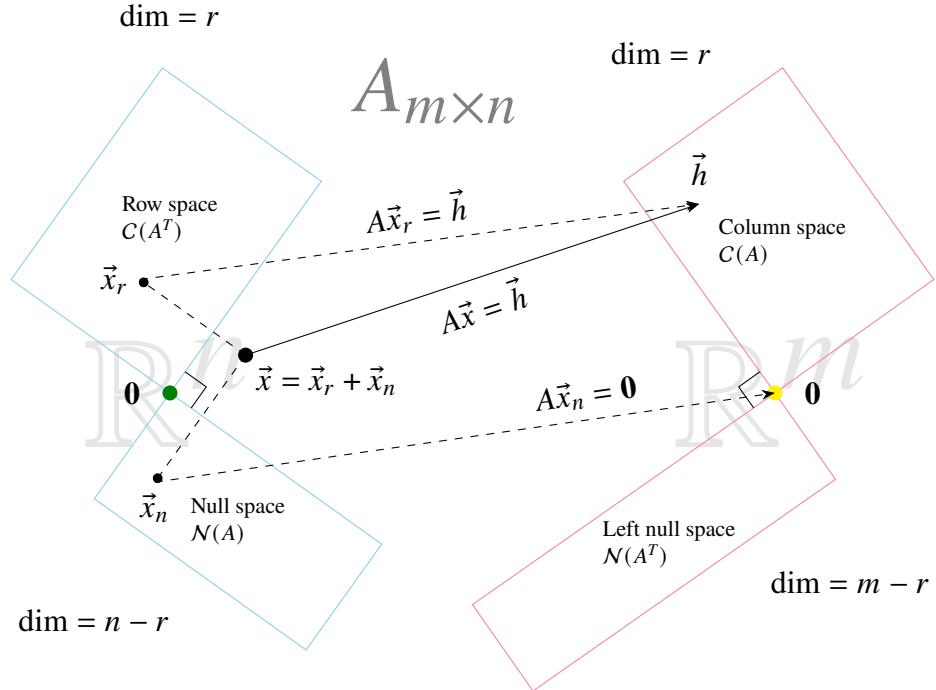


Figure 6.3: The relationships between the four fundamental subspaces for an  $m \times n$  real matrix  $A$  of rank  $r$ : the row space  $C(A^T)$ , null space  $N(A)$ , column space  $C(A)$ , left null space  $N(A^T)$ . Any vector  $\vec{x} \in \mathbb{R}^n$  can be partitioned into  $\vec{x} = \vec{x}_r + \vec{x}_n$  uniquely, where  $\vec{x}_r \in C(A^T) \subseteq \mathbb{R}^n$  and  $\vec{x}_n \in N(A) \subseteq \mathbb{R}^n$  are in the row/null space of  $A$  respectively. The matrix  $A$  maps  $\vec{x}_n$  to the zero vector in  $\mathbb{R}^m$  and  $\vec{x}_r$  to some vector  $\vec{h} \in C(A) \subseteq \mathbb{R}^m$  in the column space of  $A$ . The total effect on  $\vec{x}$  multiplied by  $A$ , is the sum of the two responses:  $A\vec{x} = A(\vec{x}_r + \vec{x}_n) = A\vec{x}_r + A\vec{x}_n = \vec{h} + \mathbf{0} = \vec{h}$ .

in  $A$  and hence are in the column space of  $A$ . By Properties 6.1.8, the column space of  $A$  then contains the column space of  $AB$ . Applying Properties 6.2.5 we have  $\text{rank}(AB) = \dim(C(AB)) \leq \dim(C(A)) = \text{rank}(A)$ . The same argument on the row space of  $AB$  and  $B$  similarly shows that  $\text{rank}(AB) \leq \text{rank}(B)$  and the desired inequality follows.  $\square$

Short Exercise: Show that  $\text{rank}(AB) = \text{rank}(A)$  if  $B$  is an invertible square matrix.<sup>21</sup>

## 6.4 Python Programming

To check linear independence and find a basis for columns in a matrix (or in general, any basis from a spanning set), we can use the `columnspace` method in `sympy`. Let's test it with the matrix in Example 6.3.1.

```
import sympy

myMatrix = sympy.Matrix([[1., 1., -2., 1.],
                        [1., 2., 1., -1.],
                        [1., 0., -5., 3.]])
print(myMatrix.columnspace())
```

which gives

```
[Matrix([
[1.0],
[1.0],
[1.0]]),
Matrix([
[1.0],
[2.0],
[ 0]])]
```

as expected. The rank can be found in two ways.

```
print(myMatrix.rank()) # or len(myMatrix.columnspace())
```

---

<sup>21</sup> $\text{rank}(A) = \text{rank}(ABB^{-1}) \leq \text{rank}(AB) \leq \text{rank}(A)$  by applying Properties 6.3.14 twice. So  $\text{rank}(AB)$  is "sandwiched" by and must be equal to  $\text{rank}(A)$ .

This returns 2 correctly. We can make a basis for the row space similarly by the `rowspace` method. In the same manner, the null space is computed by the `nullspace` method:

```
print(myMatrix.nullspace())
```

producing an output of

```
[Matrix([
[ 5.0],
[-3.0],
[ 1],
[ 0]]),
Matrix([
[-3.0],
[ 2.0],
[ 0],
[ 1]])]
```

The nullity is then simply calculated by `len(myMatrix.nullspace())`, which gives a right answer of 2. Finally, CR Factorization is computed via the `rank_decomposition` method, where

```
C, R = myMatrix.rank_decomposition()
print(C, R)
```

gives

```
Matrix([[1.00, 1.00],
       [1.00, 2.00],
       [1.00, 0]])
Matrix([[1, 0, -5.00, 3.00],
       [0, 1, 3.00, -2.00]])
```

where the matrix  $C$  is essentially the same as that comes from `columnspace`.

## 6.5 Exercises

**Exercise 6.1** For  $\vec{v}^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ ,  $\vec{v}^{(2)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ ,  $\vec{v}^{(3)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$ , find the constants  $a, b, c$  such that their linear combination  $a\vec{v}^{(1)} + b\vec{v}^{(2)} + c\vec{v}^{(3)}$  equals to

- (a)  $(3, 2, 9)^T$ ,
- (b)  $(9, 1, 5)^T$ .

**Exercise 6.2** Determine if the following sets of vectors are linearly independent.

- (a)  $\vec{u} = (2, -1)^T$ ,  $\vec{v} = (-4, 2)^T$ ,
- (b)  $\vec{u} = (1, 2, 3)^T$ ,  $\vec{v} = (6, 7, 9)^T$ ,  $\vec{w} = (4, 8, 5)^T$ , and
- (c)  $\vec{u} = (1, 3, 3)^T$ ,  $\vec{v} = (3, 2, 9)^T$ ,  $\vec{w} = (1, -4, 3)^T$ .

**Exercise 6.3** Given a spanning set  $\mathcal{G} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \vec{v}^{(4)}\}$  in which  $\vec{v}^{(1)} = (1, 3, 0, 1)^T$ ,  $\vec{v}^{(2)} = (1, -1, 2, -1)^T$ ,  $\vec{v}^{(3)} = (-1, 2, 1, 2)^T$ ,  $\vec{v}^{(4)} = (3, 0, 1, -2)^T$ , determine if

- (a)  $(1, 4, 3, 2)^T$ , and
- (b)  $(1, 2, -3, 1)^T$ .

are in the subspace generated by  $\mathcal{G}$ .

**Exercise 6.4** For the basis  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$ , where  $\vec{v}^{(1)} = (6, 1, 2)^T$ ,  $\vec{v}^{(2)} = (1, 0, 1)^T$ ,  $\vec{v}^{(3)} = (2, 3, 3)^T$  (relative to the standard basis  $\mathcal{S}$ ), do the following coordinate conversion.

- (a) Find the coordinates of  $(5, 2, 3)^T$  in the  $\mathcal{B}$  frame,
- (b) Transform  $(1, -1, 1)_B^T$  from the  $\mathcal{B}$  system back to the the standard basis  $\mathcal{S}$ .

**Exercise 6.5** Show that  $\mathcal{B} = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \vec{w}^{(3)}\}$ , where  $\vec{w}^{(1)} = (1, -1, 0, 1)^T$ ,  $\vec{w}^{(2)} = (2, 1, 1, 0)^T$ ,  $\vec{w}^{(3)} = (1, 2, -1, 1)^T$  forms a basis for the subspace generated by itself, and find the coordinates of  $\vec{v} = (1, -1, -2, 3)^T$  with respect to this basis.

**Exercise 6.6** Prove that for any two subspaces  $\mathcal{W}_1, \mathcal{W}_2 \subseteq \mathcal{V}$ . Their intersection  $\mathcal{W}_1 \cap \mathcal{W}_2$  is also a subspace of  $\mathcal{V}$ . How about their union?

**Exercise 6.7** Show that  $\mathcal{W}_1 = \text{span}(\{(1, 0, 0, 1)^T, (0, 1, -1, 1)^T\})$  and  $\mathcal{W}_2 = \text{span}(\{(1, 0, 1, -1)^T\})$  can be composed to produce a direct sum  $\mathcal{W}_1 \oplus \mathcal{W}_2$ . Find bases for  $\mathcal{W}_1, \mathcal{W}_2$  and hence this direct sum. Express  $(2, 0, 1, 0)^T$  in the direct sum basis as the combined coordinates of the two smaller subspaces.

**Exercise 6.8** Find (bases for) the column, row, null and left null space of

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

and check if Theorem 6.3.9 and Properties 6.3.13 hold in this case.

## Chapter 7

# More on Coordinate Bases, Linear Transformations

---

In this chapter we will go deeper about what actually a matrix represents in the big picture. Matrices by nature is a rule of *linear transformations* (or *linear mappings*) between two vector spaces. We are going to study several special types of linear transformations, which ultimately reveals the relationship between any  $n$ -dimensional real vector space and the real  $n$ -space  $\mathbb{R}^n$ , as an *isomorphism*. We then move to discuss how a change of coordinates works for vectors and matrices, as well as the *Gram-Schmidt* process to make an *orthonormal* basis.

## 7.1 Ideas of Linear Transformations

### 7.1.1 Linear Maps between Vector Spaces

Consider two vector spaces, we may want to know if vectors in one of the vector spaces, let's say  $\mathcal{U}$ , can be associated to or *transformed* into those in another vector space  $\mathcal{V}$ , according to some rules. This is known as a *transformation/mapping* from the vector space  $\mathcal{U}$  to  $\mathcal{V}$ . Of the most concern is the class of *linear transformations/mappings* which obeys the two properties listed below.

**Definition 7.1.1** (Linear Transformation/Map). A linear transformation (or linear map) from a vector space  $\mathcal{U}$  to another vector space  $\mathcal{V}$  is a mapping:  $T : \mathcal{U} \rightarrow \mathcal{V}$ , such that for all vectors  $\vec{u}^{(1)}, \vec{u}^{(2)} \in \mathcal{U}$ , and any scalar  $a$ , it satisfies:

1.  $T(\vec{u}^{(1)} + \vec{u}^{(2)}) = T(\vec{u}^{(1)}) + T(\vec{u}^{(2)})$  (Additivity), and
2.  $T(a\vec{u}^{(j)}) = aT(\vec{u}^{(j)})$  (Homogeneity).

These two properties combined are known as *linearity*. An equivalent condition is  $T(a\vec{u}^{(1)} + b\vec{u}^{(2)}) = aT(\vec{u}^{(1)}) + bT(\vec{u}^{(2)})$ , where  $b$  is any scalar as well.

Notice that if  $\mathcal{U}/\mathcal{V}$  coincides with the real  $n/m$ -space  $\mathbb{R}^n/\mathbb{R}^m$ , and we express any vector  $\vec{u} \equiv [\vec{u}]_B$ <sup>1</sup> in  $\mathcal{U}$  by  $n$  coordinates using some basis  $\mathcal{B}$  (similarly for  $\vec{v} \equiv [\vec{v}]_H \in \mathcal{V}$  that has  $m$  coordinates in some basis  $\mathcal{H}$ ). Then  $[T]_B^H = A$  where  $A$  is any  $m \times n$  matrix satisfies the requirements of and is a linear transformation from  $\mathcal{U}$  to  $\mathcal{V}$  according to the rule  $T(\vec{u}) \equiv [T]_B^H [\vec{u}]_B = A[\vec{u}]_B$ . (Short Exercise: show this satisfies the conditions outlined in Definition 7.1.1 above!<sup>2</sup>) This implies that all matrices can be considered as some sort of linear mappings (for now, between  $\mathbb{R}^n$  and  $\mathbb{R}^m$ ). In fact, the converse, which states that any linear transformation (between finite-dimensional vector spaces) can be represented by a matrix, is also true as well, and will be discussed in the following parts of this section.

Provided that  $\mathcal{U}/\mathcal{V}$  is  $n/m$ (finite)-dimensional, let's now explicitly fix a basis  $\mathcal{B} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(n)}\}$  for  $\mathcal{U}$  (again, similarly we have  $\mathcal{H} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(m)}\}$  for  $\mathcal{V}$ ). For each  $\vec{u}^{(j)}$ ,  $j = 1, 2, \dots, n$ , denote  $\vec{w}^{(j)} = T(\vec{u}^{(j)})$  as the resulting vectors in  $\mathcal{V}$  after applying the transformation  $T$  over the basis vectors for  $\mathcal{U}$ . Notice that  $[\vec{u}^{(j)}]_B = (e^{(j)})_B$  where the  $j$ -th basis vector of  $\mathcal{B}$  is explicitly represented in a coordinate form with the  $j$ -th entry being 1 and others being 0 (where the usual hat symbol on  $e$  is not present) relative to the  $\mathcal{B}$  system. Due to Definition 6.2.1 and Properties 6.1.12,  $T(\vec{u}^{(j)}) = \vec{w}^{(j)} \in \mathcal{V}$  can be expressed as a *unique* linear combination as  $\vec{w}^{(j)} = a_1^{(j)}\vec{v}^{(1)} + a_2^{(j)}\vec{v}^{(2)} + \dots + a_m^{(j)}\vec{v}^{(m)} = \sum_{i=1}^m a_i^{(j)}\vec{v}^{(i)}$

<sup>1</sup>We use " $\equiv$ " to associate any vector to its representation in some fixed coordinate system.

<sup>2</sup> $T(\vec{u}^{(1)} + \vec{u}^{(2)}) \equiv A([\vec{u}^{(1)}]_B + [\vec{u}^{(2)}]_B) = A[\vec{u}^{(1)}]_B + A[\vec{u}^{(2)}]_B \equiv T(\vec{u}^{(1)}) + T(\vec{u}^{(2)})$  and  $T(a\vec{u}^{(1)}) \equiv A(a[\vec{u}^{(1)}]_B) = a(A[\vec{u}^{(1)}]_B) \equiv aT(\vec{u}^{(1)})$

of the basis vectors  $\vec{v}^{(i)}$  from  $\mathcal{H}$ , i.e.

$$T(\vec{u}^{(j)}) = \sum_{i=1}^m a_i^{(j)} \vec{v}^{(i)}$$

The matrix formed by the above coefficients  $A = a_i^{(j)}$  is then the desired, *unique matrix representation* of our linear transformation  $T$ . To see this, insert  $\vec{u}^{(j)}$  and compare both sides of  $T(\vec{u}) \equiv A[\vec{u}]_B$ . Subsequently,

$$\begin{aligned} A[\vec{u}^{(j)}]_B &= a_i^{(j)}(e^{(j)})_B \\ &= \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(j)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(j)} & & a_2^{(n)} \\ \vdots & & & \vdots & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(j)} & \cdots & a_m^{(n)} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \text{ (the } j\text{-th entry)} \\ \vdots \\ 0 \text{ (the last index is } n) \end{bmatrix} \\ &= \begin{bmatrix} a_1^{(j)} \\ a_2^{(j)} \\ \vdots \\ a_m^{(j)} \end{bmatrix} \end{aligned}$$

Due to the structure of  $(e^{(j)})_B$ , this matrix product yields exactly the  $j$ -th column of  $A = a_i^{(j)}$  as shown above (see the remark under Properties 6.1.4). Moreover, the coordinates of  $\vec{w}^{(j)} = T(\vec{u}^{(j)})$  in the  $\mathcal{H}$  system

$$\begin{aligned} T(\vec{u}^{(j)}) &\equiv [T(\vec{u}^{(j)})]_H = [\vec{w}^{(j)}]_H \\ &= \left[ \sum_{i=1}^m a_i^{(j)} \vec{v}^{(i)} \right]_H \\ &= \sum_{i=1}^m a_i^{(j)} [\vec{v}^{(i)}]_H \\ &= \sum_{i=1}^m a_i^{(j)} (e^{(i)})_H \end{aligned}$$

$$\begin{aligned}
 &= a_1^{(j)} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2^{(j)} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + a_m^{(j)} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}_{\text{(the last index is } m\text{)}} \\
 &= \begin{bmatrix} a_1^{(j)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_2^{(j)} \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ a_m^{(j)} \end{bmatrix} = \begin{bmatrix} a_1^{(j)} \\ a_2^{(j)} \\ \vdots \\ a_m^{(j)} \end{bmatrix}
 \end{aligned}$$

also gives the same  $j$ -th column of  $A_{ij} = a_i^{(j)}$ . By the same argument, this holds for any  $j$ . Hence, the association of the matrix  $[A]_B^H = a_i^{(j)}$  to the linear transformation  $T$  is consistent, where we have now added the subscript  $B$  and superscript  $H$  to emphasize the transformation are carried out in reference to these two specific coordinate bases. This same line of reasoning also shows that, to construct the matrix representation of a linear transformation, we compute  $T(\vec{u}_j) = \vec{w}^{(j)}$  for each of the  $\vec{u}_j$  in the  $\mathcal{B}$  basis and find its coordinates in the  $\mathcal{H}$  frame, namely  $[\vec{w}^{(j)}]_H$ , which readily become the  $j$ -th column of the matrix to be found. To be more clear, we have

$$\begin{aligned}
 [T]_B^H &= [[T(\vec{u}^{(1)})]_H | [T(\vec{u}^{(2)})]_H | \cdots | [T(\vec{u}^{(n)})]_H] \\
 &= [[\vec{w}^{(1)}]_H | [\vec{w}^{(2)}]_H | \cdots | [\vec{w}^{(n)}]_H] \\
 &= \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(n)} \\ \vdots & & \ddots & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(n)} \end{bmatrix} = a_i^{(j)} = [A]_B^H
 \end{aligned}$$

Notice that here the  $i/j$  subscript/superscript has been exchanged when compared to like Properties 1.2.3.

**Definition 7.1.2** (Matrix Representation of a Linear Transformation). A linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  as defined in Definition 7.1.1 between two finite-dimensional vector spaces, with respect to the bases  $\mathcal{B} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(n)}\}$

for  $\mathcal{U}$  and  $\mathcal{H} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(m)}\}$  for  $\mathcal{V}$ , has a unique matrix representation of

$$[T]_B^H = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(n)} \\ \vdots & & \ddots & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(n)} \end{bmatrix}$$

where the entries  $a_i^{(j)}$  are those found according to the relations  $T(\vec{u}^{(j)}) = \sum_{i=1}^m a_i^{(j)} \vec{v}^{(i)}$ , or in matrix notation,  $[T]_B^H [\vec{u}]_B = [\vec{v}]_H$ .

Let's illustrate how it works out using an easy example using the familiar  $\mathbb{R}^n$  and  $\mathbb{R}^m$ .

**Example 7.1.1.** Let  $\mathcal{U} = \mathbb{R}^3$  and  $\mathcal{V} = \mathbb{R}^2$ , it can be easily verified that  $\mathcal{B} = \{(1, 2, 1)^T, (0, 1, -1)^T, (2, -1, 1)^T\}$  is a basis for  $\mathcal{U}$ , and the same goes for  $\mathcal{V}$  with a basis  $\mathcal{H} = \{(1, 2)^T, (2, -1)^T\}$ . If a linear transformation  $T : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  obeys the rule  $T((x, y, z)^T) = (x + 2y, x - y + z)^T$  (by the way, you should verify if this is really a linear transformation), find its matrix representation  $[T]_B^H$  with respect to the  $\mathcal{B}$  and  $\mathcal{H}$  system. Then, use the results to compute  $T((-1, 4, -1)^T)$ .

*Solution.* Following Definition 7.1.2, we set out to find how the linear transformation will apply on the basis vectors in  $\mathcal{B}$ . For the first one, we have

$$T((1, 2, 1)^T) = ((1) + 2(2), (1) - (2) + (1))^T = (5, 0)^T$$

which can be subsequently written as a linear combination of the two basis vectors in  $\mathcal{H}$ :

$$(5, 0)^T = 1(1, 2)^T + 2(2, -1)^T$$

Hence  $a_1^{(1)} = 1$ ,  $a_2^{(1)} = 2$ , and this gives us the first column of  $[T]_B^H$  as

$$\begin{bmatrix} 1 & * & * \\ 2 & * & * \end{bmatrix}$$

We repeat the same procedure on the other two basis vectors  $(0, 1, -1)^T$  and  $(2, -1, 0)^T$  of  $\mathcal{B}$ , where it can be shown that

$$\begin{aligned} T((0, 1, -1)^T) &= ((0) + 2(1), (0) - (1) + (-1))^T = (2, -2)^T \\ &= -\frac{2}{5}(1, 2)^T + \frac{6}{5}(2, -1)^T \\ T((2, -1, 1)^T) &= ((2) + 2(-1), (2) - (-1) + (1))^T = (0, 4)^T \\ &= \frac{8}{5}(1, 2)^T - \frac{4}{5}(2, -1)^T \end{aligned}$$

Therefore, the required matrix representation is

$$[T]_B^H = \begin{bmatrix} 1 & -\frac{2}{5} & \frac{8}{5} \\ 2 & \frac{6}{5} & -\frac{4}{5} \end{bmatrix}$$

For the second part, we start by expressing  $(-1, 4, 1)^T$  in the basis  $\mathcal{B}$ . As  $(-1, 4, -1)^T = 1(1, 2, 1)^T + 1(0, 1, -1)^T - 1(2, -1, 1)^T$ , we have  $(-1, 4, 1)^T = (1, 1, -1)_B^T$ , and then

$$\begin{aligned} [T((1, 1, -1)_B^T)]_H &= [T]_B^H (1, 1, -1)_B^T \\ &= \begin{bmatrix} 1 & -\frac{2}{5} & \frac{8}{5} \\ 2 & \frac{6}{5} & -\frac{4}{5} \end{bmatrix}_B^H \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}_B \\ &= \begin{bmatrix} -1 \\ 4 \end{bmatrix}_H = (-1, 4)_H^T \end{aligned}$$

implying that  $T((-1, 4, -1)^T) = (-1, 4)_H^T = -1(1, 2)^T + 4(2, -1)^T = (7, -6)^T$  in the usual standard basis  $\mathcal{S}$ . This can be cross-checked by directly invoking the given definition of  $T$ , where  $T((-1, 4, -1)^T) = ((-1) + 2(4), (-1) - (4) + (-1))^T = (7, -6)^T$  as well.  $\square$

Up until now, we have been playing around with the simple real  $n$ -space only, but the real (no pun intended) power of the notion of a general vector space lies in its abstraction: Any mathematical object that satisfies the criteria in Definition 6.1.1 is a (real) vector space, and the results that we have already

established in the previous parts for the real  $n$ -space are readily transferable to them.<sup>3</sup> Two prime examples of abstract vector spaces are the set of (real) polynomials  $\mathcal{P}^n$  with a degree up to  $n$ <sup>4</sup> and the family of continuous ( $k$ -times continuously differentiable) functions  $C^0$  ( $C^k$ ) over a fixed interval. Now we will see how the concept of linear transformation is laid out when these abstract vector spaces are involved.

**Example 7.1.2.** Consider  $\mathcal{U} = \mathcal{P}^2$ , and  $\mathcal{V} = \mathcal{P}^1$ , and let the bases for  $\mathcal{U}$  and  $\mathcal{V}$  be  $\mathcal{B} = \{1, x, x^2\}$  and  $\mathcal{H} = \{1, x\}$ . (They are known as the standard bases for  $\mathcal{P}^2$  and  $\mathcal{P}^1$  respectively. In general the standard basis for  $\mathcal{P}^n$  is  $\{1, x, x^2, \dots, x^{n-1}, x^n\}$  and thus  $(n+1)$ -dimensional. Readers are advised to justify why they constitute a basis for the polynomial spaces.) Let  $T : \mathcal{U} \rightarrow \mathcal{V}$  be  $T[p(x)] = p'(x)$  the differentiation operator and find its matrix representation with respect to  $\mathcal{B}$  and  $\mathcal{H}$ .

*Solution.* We essentially do the same thing as in Example 7.1.1 but applied over polynomials now. From elementary calculus, we have

$$\begin{aligned} T(1) &= \frac{d}{dx}(1) = 0 \\ T(x) &= \frac{d}{dx}(x) = 1 \\ T(x^2) &= \frac{d}{dx}(x^2) = 2x \end{aligned}$$

and by Definition 7.1.2, the desired matrix representation is

$$[T]_B^H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

---

<sup>3</sup>Again, this appeals to the isomorphic nature of same-dimensional real vector spaces, which will be discussed soon.

<sup>4</sup>We shall argue for some criteria in Definition 6.1.1 for  $\mathcal{P}^n$  here. For instances, condition (1) is obvious as adding up two polynomials with a degree up to  $n$  can only result in another polynomial with a maximum degree of  $n$ . In condition (4), the zero vector for  $\mathcal{P}^n$  is simply the constant zero function 0, which is considered to have a degree of  $-1$  by convention.

Notice that we can express, quite trivially

$$1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}_H$$

$$x = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}_B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}_H$$

$$x^2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}_B$$

using vector notation in the given two standard bases. We can verify the form of  $[T]_B^H$  by a test polynomial  $c_0 + c_1x + c_2x^2$ , whose vector representation in  $\mathcal{B}$  is clearly  $(c_0, c_1, c_2)_B^T$ . Then, multiplying  $[T]_B^H$  to its left gives

$$[T((c_0, c_1, c_2)_B^T)]_H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}_B^H \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}_B = \begin{bmatrix} c_1 \\ 2c_2 \end{bmatrix}_H$$

which corresponds to the polynomial  $c_1 + 2c_2x$ . This coincides with the usual result of differentiation, that is,  $\frac{d}{dx}(c_0 + c_1x + c_2x^2) = c_1 + 2c_2x$ .  $\square$

In each of the previous examples, we consider a linear transformation between two vector spaces that are of the same type (the usual real vectors/polynomials). Below shows what happen when they are mixed together. Actually, due to the abstraction provided by the nature of vector space, the outcome follows easily.

**Example 7.1.3.** Let  $\mathcal{U} = \mathbb{R}^3$  and  $\mathcal{V} = \mathcal{P}^2$ , while  $\mathcal{B} = \{(1, 0, 0)^T, (0, 1, 0)^T, (0, 0, 1)^T\}$  and  $\mathcal{H} = \{1, x, x^2\}$  be the standard bases for  $\mathcal{U}$  and  $\mathcal{V}$  respectively. Show that, the rather trivial linear transformation  $T((c_0, c_1, c_2)^T) = c_0 + c_1x + c_2x^2$  has a matrix representation of an identity with respect to  $\mathcal{B}$  and  $\mathcal{H}$ .

*Solution.* Again, we repeat what we have done in the previous two examples. It is apparent that

$$\begin{aligned} T((1, 0, 0)_B^T) &= (1) + (0)x + (0)x^2 = 1 = (1, 0, 0)_H^T \\ T((0, 1, 0)_B^T) &= (0) + (1)x + (0)x^2 = x = (0, 1, 0)_H^T \\ T((0, 0, 1)_B^T) &= (0) + (0)x + (1)x^2 = x^2 = (0, 0, 1)_H^T \end{aligned}$$

So by Definition 7.1.2, the desired matrix representation is simply

$$[T]_B^H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which is the  $3 \times 3$  identity matrix. This is expected as the linear transformation is essentially  $T[(c_0, c_1, c_2)_B^T] = (c_0, c_1, c_2)_H^T$  where  $(c_0, c_1, c_2)_H^T = c_0 + c_1x + c_2x^2$ , which means that the numeric coordinate representation of vectors in the two spaces is preserved under such a linear transformation between them and the only visible change is the subscript.  $\square$

Most of the readers should find it boring in the above example as we are just stating the obvious. It is a straight-forward, "one-to-one" association between the standard bases of the real  $n$ -space and space of polynomials with degree  $n - 1$ . However, the important message is that given such an association we can always identify any vector of some space as a vector in another space of a completely different class, which is very powerful as many operations become transferable between these two spaces. In this sense, this kind of "one-to-one" mapping is not limited to mappings that have an identity representation, or affected by the bases used for the two vector spaces as we will see in the following subsection.

### 7.1.2 One-to-one and Onto, Kernel and Range

Continuing our discussion above, to identify a vector (one and only one) from one vector space as another vector in another vector space through a linear mapping, we require it to be **one-to-one (injective)**. On the other hand, another important

property of a linear transformation is that whether it is *onto (surjective)*, which means that every vector (*image*) in the latter vector space (*codomain*) is being mapped onto by (or speaking loosely, "comes from") some vector(s) (*preimage*) in the former vector space (*domain*). The formal definitions of these two properties are given as below.

**Properties 7.1.3** (Injective Transformation). A transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  is called *one-to-one* if for any pair of two (may or may not be distinct) vectors  $\vec{u}^{(1)}, \vec{u}^{(2)} \in \mathcal{U}, T(\vec{u}^{(1)}) = T(\vec{u}^{(2)})$  implies  $\vec{u}^{(1)} = \vec{u}^{(2)}$ , i.e. an image has one and only one corresponding preimage. Furthermore, if  $T$  is linear, then an equivalent condition is that  $T(\vec{u}) = \mathbf{0}$  implies  $\vec{u} = \mathbf{0}$  as the only possibility.

To show the equivalence of the two conditions above, notice that  $T(\mathbf{0}) = \mathbf{0}$  if  $T$  is linear. (why?)<sup>5</sup> For any  $\vec{u}$  such that  $T(\vec{u}) = \mathbf{0}$ , we have

$$T(\vec{u}) = \mathbf{0} = T(\mathbf{0})$$

and hence  $\vec{u}$  must be  $\mathbf{0}$  if  $T(\vec{u}^{(1)}) = T(\vec{u}^{(2)})$  implies  $\vec{u}^{(1)} = \vec{u}^{(2)}$ . The proof of the converse is left as an exercise.

**Properties 7.1.4** (Surjective Transformation). A transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  is called *onto* if for any vector  $\vec{v} \in \mathcal{V}$  (image), there exists at least one vector(s)  $\vec{u} \in \mathcal{U}$  (preimage) such that  $T(\vec{u}) = \vec{v}$ .

As an illustration, in Example 7.1.2, the differentiation operator  $T(p(x)) = p'(x)$  from  $\mathcal{P}^2$  to  $\mathcal{P}^1$  is onto but not one-to-one. To see these, note that given any image  $\vec{v} = d_0 + d_1x \in \mathcal{P}^1$ , all preimages in the form of  $\vec{u} = K + d_0x + \frac{d_1}{2}x^2 \in \mathcal{P}^2$  where  $K$  can be any number satisfies  $T(\vec{u}) = \vec{v}$  by elementary calculus, and the surjectivity is obvious. To explicitly disprove injectivity, fix an image  $\vec{v} = d_0 + d_1x$  with specific  $d_0$  and  $d_1$ , and note that both  $\vec{u}_1 = K_1 + d_0x + \frac{d_1}{2}x^2$  and  $\vec{u}_2 = K_2 + d_0x + \frac{d_1}{2}x^2$  where  $K_1, K_2$  are distinct satisfy  $T(\vec{u}_1) = T(\vec{u}_2) = \vec{v}$ , but  $\vec{u}_1 \neq \vec{u}_2$ .

---

<sup>5</sup> $T(\mathbf{0}) = T(0\vec{u}) = 0T(\vec{u}) = \mathbf{0}$  for arbitrary  $\vec{v}$  due to the homogeneity property as required in Definition 7.1.1.

However, in other cases it may not be so easy to check injectivity and surjectivity as directly as above. Therefore, we need a general method to determine if these two properties hold for a transformation between two abstract vector bases. The following theorem links injectivity and surjectivity with their basis vectors, but it requires the transformation to be linear (and here is where the linearity comes to play).

**Theorem 7.1.5.** A linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  (between two finite-dimensional vector spaces) is one-to-one if and only if given any basis  $\mathcal{B} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(n)}\}$  for  $\mathcal{U}$ ,  $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \dots, T(\vec{u}^{(n)}) \in \mathcal{V}$  are linearly independent.

**Theorem 7.1.6.** A linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  (between two finite-dimensional vector spaces) is onto if and only if given any basis  $\mathcal{H} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(m)}\}$  for  $\mathcal{V}$ , we can find a vector  $\vec{w}^{(i)} \in \mathcal{U}$  such that  $T(\vec{w}^{(i)}) = \vec{v}^{(i)}$  for each of the  $\vec{v}_i$ .

*Proof.* Theorem 7.1.5: The "if" direction is proved by showing  $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \dots, T(\vec{u}^{(n)})$  are linearly independent implies that, if  $T(\vec{u}) = \mathbf{0}$  then  $\vec{u} = \mathbf{0}$  as suggested by the alternative condition in Properties 7.1.3. By Theorem 6.1.9, the equation  $c_1T(\vec{u}^{(1)}) + c_2T(\vec{u}^{(2)}) + \dots + c_nT(\vec{u}^{(n)}) = \mathbf{0}$  only has  $c_j = \mathbf{0}$  as the trivial solution. Now by linearity from Definition 7.1.1, we have

$$\begin{aligned} c_1T(\vec{u}^{(1)}) + c_2T(\vec{u}^{(2)}) + \dots + c_nT(\vec{u}^{(n)}) &= T(c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \dots + c_n\vec{u}^{(n)}) \\ &= \mathbf{0} \end{aligned}$$

Since  $c_j = 0$  is the only possibility, this means that if  $T(c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \dots + c_n\vec{u}^{(n)}) = \mathbf{0}$  then  $\vec{u} = c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \dots + c_n\vec{u}^{(n)}$  must be  $\mathbf{0}$ , hence  $T(\vec{u}) = \mathbf{0}$  implies  $\vec{u} = \mathbf{0}$  and we are done. The converse is similarly proved, having the argument goes in reverse direction.

Theorem 7.1.6: We compare Theorem 7.1.6 against Properties 7.1.4 to show the part of "if" direction. Since  $H = \{\vec{v}^{(i)}\}$ ,  $i = 1, 2, \dots, m$ , is a basis for  $\mathcal{V}$ , any  $\vec{v} \in \mathcal{V}$  can be written as a linear combination of  $\vec{v} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + \dots + c_m\vec{v}^{(m)}$ . If we can find  $\vec{w}^{(i)} \in \mathcal{U}$  such that  $T(\vec{w}^{(i)}) = \vec{v}^{(i)}$  for all  $\vec{v}^{(i)}$ , then

$$\vec{v} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + \dots + c_m\vec{v}^{(m)}$$

$$\begin{aligned}
 &= c_1 T(\vec{w}^{(1)}) + c_2 T(\vec{w}^{(2)}) + \cdots + c_m T(\vec{w}^{(m)}) \\
 &= T(c_1 \vec{w}^{(1)} + c_2 \vec{w}^{(2)} + \cdots + c_m \vec{w}^{(m)})
 \end{aligned}$$

the last equality uses linearity from Definition 7.1.1 again. This shows that  $\vec{u} = c_1 \vec{w}^{(1)} + c_2 \vec{w}^{(2)} + \cdots + c_m \vec{w}^{(m)}$  is readily one possible vector in  $\mathcal{U}$  such that  $T(\vec{u}) = \vec{v}$  and the desired result is established. The converse is trivial as we take  $\vec{v} = \vec{v}^{(i)}$  in Properties 7.1.4 for all possible  $i$ .  $\square$

**Example 7.1.4.** Given a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  where  $\mathcal{U}$  and  $\mathcal{V}$  have a dimension of 3 and 4 respectively, if its matrix representation corresponding to some bases  $\mathcal{B}$  and  $\mathcal{H}$  is

$$[T]_B^H = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

determine whether it is (a) one-to-one, as well as (b) onto, or not.

*Solution.* (a) By Theorem 7.1.5, we need to check if  $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), T(\vec{u}^{(3)})$  are linearly independent, where  $\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}$  are the basis vectors from  $\mathcal{B}$ . Their coordinate representation in the  $\mathcal{B}$  system is trivially  $[\vec{u}^{(1)}]_B = (e^{(1)})_B = (1, 0, 0)^T_B$ ,  $[\vec{u}^{(2)}]_B = (e^{(2)})_B = (0, 1, 0)^T_B$  and  $[\vec{u}^{(3)}]_B = (e^{(3)})_B = (0, 0, 1)^T_B$ , and hence

$$\begin{aligned}
 [T(\vec{u}^{(1)})]_H &= [T]_B^H (e^{(1)})_B \\
 &= \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}_B^H \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_B = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}_H
 \end{aligned}$$

which is just the first column of  $[T]_B^H$ . Similarly,  $[T(\vec{u}^{(2)})]_H = (-1, 1, 0, 1)^T_H$ ,  $[T(\vec{u}^{(3)})]_H = (0, 1, -1, 0)^T_H$  are then the second/third column of  $[T]_B^H$ . From this we see that in general, the coordinates in  $\mathcal{H}$  after transformation

$[T(\vec{u}^{(j)})]_H$  is just the  $j$ -th column of  $[T]_B^H$ . (Actually, this has been observed when we are deriving the matrix representation of linear transformations in the beginning of this chapter.) So the problem is reduced to decide whether the column vectors constituting  $[T]_B^H$  are linearly independent or not. By Theorem 6.1.9, we can accomplish this by showing if the solution  $[T]_B^H \vec{u} = \mathbf{0}$  is consisted of the trivial solution only, and we have

$$\begin{array}{c}
 \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 2 & 0 & 0 \end{array} \right] \quad R_3 - R_1 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & -2 & 0 \end{array} \right] \quad R_4 - R_1 \rightarrow R_4 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 \end{array} \right] \quad R_3 - R_2 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 \end{array} \right] \quad R_4 - R_2 \rightarrow R_4 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 \end{array} \right] \quad -\frac{1}{2}R_3 \rightarrow R_3 \\
 \qquad\qquad\qquad \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_4 + 2R_3 \rightarrow R_4
 \end{array}$$

As every column in this homogeneous system contains a pivot, it demonstrates that  $[T]_B^H \vec{u} = \mathbf{0}$  indeed only has the trivial solution  $\vec{u} = \mathbf{0}$ , and therefore the linear transformation in question is one-to-one.

- (b) By Properties 7.1.4, it is equivalent to showing that if the  $\{T(\vec{u}^{(j)})\}$  span  $\mathcal{W}$ , or expressed in terms of the  $\mathcal{B}/\mathcal{H}$  coordinates, whether the three transformed vectors  $\{[T]_B^H [\vec{u}^{(1)}]_B, [T]_B^H [\vec{u}^{(2)}]_B, [T]_B^H [\vec{u}^{(3)}]_B\}$  span  $\mathbb{R}^4$ . However, it is apparent that three vectors can never span a four-dimensional vector space as the number of vectors is fewer than the dimension, and thus the linear transformation is not onto.

Notice that in the above arguments we never explicitly say what the vector spaces  $\mathcal{U}$  and  $\mathcal{V}$  are and only the matrix representation of the linear transformation is involved. However, some may be skeptical as we have fixed bases for the linear transformation and may ask if the results are *basis-dependent*. We will address this issue in later parts of this section.  $\square$

Accompanying injectivity and surjectivity is the ideas of *kernel* and *range*. For a (linear) transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$ , its kernel is consisted of vectors in  $\mathcal{U}$  that is mapped to the zero vector in  $\mathcal{V}$ , while its range is made up of all possible vectors in  $\mathcal{V}$  that are mapped from  $\mathcal{U}$  via  $T$ .

**Definition 7.1.7.** For a (linear) transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$ , its kernel is defined to be

$$\text{Ker}(T) = \{\vec{u} \in \mathcal{U} | T(\vec{u}) = \mathbf{0}_V\}$$

whereas its range is

$$\mathcal{R}(T) = \{\vec{v} \in \mathcal{V} | T(\vec{u}) = \vec{v} \text{ for some } \vec{u} \in \mathcal{U}\}$$

Also, notice that the kernel and range are a subspace of  $\mathcal{U}$  and  $\mathcal{V}$  respectively.<sup>6</sup> Hence it is reasonable to speak of their dimension or basis and we will discuss this matter later. For now, let's look at how to determine the kernel and range of a linear transformation first. For instance, in Example 7.1.2, the kernel is  $\text{span}(\{1\})$  since (only) the derivative of any constant vanishes, and the range is  $\text{span}(\{1, x\}) = \mathcal{V} = \mathcal{P}^1$  because we have already shown that every  $\mathcal{P}^1$  polynomial in this case have some corresponding preimage in  $\mathcal{U} = \mathcal{P}^2$ . Here the dimension of kernel/range is 1 and 2.

**Example 7.1.5.** Given another linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  where  $\mathcal{U}$  and  $\mathcal{V}$  are now both having a dimension of 3, if its matrix representation

---

<sup>6</sup>For  $\vec{u}^{(1)}, \vec{u}^{(2)} \in \text{Ker}(T) \subset \mathcal{U}$ ,  $T(a\vec{u}^{(1)} + b\vec{u}^{(2)}) = aT(\vec{u}^{(1)}) + bT(\vec{u}^{(2)}) = a\mathbf{0}_V + b\mathbf{0}_V = \mathbf{0}_V$  for any scalar  $a$  and  $b$  so  $a\vec{u}^{(1)} + b\vec{u}^{(2)} \in \text{Ker}(T)$  and by Theorem 6.1.2 it is a subspace of  $\mathcal{U}$ . We leave showing the range is a subspace of  $\mathcal{V}$  as an exercise to the readers.

corresponding to some bases  $\mathcal{B}$  and  $\mathcal{H}$  is

$$[T]_B^H = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

find its kernel and range.

*Solution.* According to Definition 7.1.7,  $\text{Ker}(T)$  is the set of  $\vec{u}$  that satisfies  $T(\vec{u}) = \mathbf{0}$ , or using basis representation (Definition 7.1.2),  $[T]_B^H[\vec{u}]_B = \mathbf{0}$ . Therefore, it is equivalent to finding the null space (Definition 6.3.8) of  $[T]_B^H$ :

$$\begin{array}{c} \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array} \right] \quad R_2 - R_1 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array} \right] \quad R_3 - R_1 \rightarrow R_3 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_2 \leftrightarrow R_3 \\ \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3 \end{array}$$

The nullity is 1 and we can let  $[u_3]_B = t$  be the free variable, and we have  $[u_1]_B = -t$  and  $[u_2]_B = 0$  from the first two rows. So the kernel takes the form of

$$\text{Ker}(T) = \begin{bmatrix} -t \\ 0 \\ t \end{bmatrix}_B = t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}_B$$

where  $-\infty < t < \infty$ , or in other words,  $\text{Ker}(T) = \text{span}(\{(-1, 0, 1)_B^T\})$  with a dimension of 1. Similarly, the range of  $T$  will be the column space of  $[T]_B^H$ . From the elimination procedure carried out above, we know that the first two column vectors are linearly independent and the third column is clearly the same as the first column (see Properties 6.1.13), and thus the range is

$\mathcal{R}(T) = \text{span}(\{(1, 1, 1)_B^T, (0, -1, 1)_B^T\})$  and has a dimension of 2 (the rank of the  $[T]_B^H$  matrix).  $\square$

Bear in mind that we approach the above problem with some bases (albeit unknown) fixed to represent the linear transformation in matrix form just like in the last example. Again, we will soon justify that the results are actually unrelated to the choices of bases, i.e. *basis-independent*, such that the (dimensions of) kernel and range are exactly the null space and column space (nullity and rank) derived using any matrix representation of the linear transformation with respect to arbitrary bases.

Finally, we can rewrite Properties 7.1.3 and 7.1.4 using the notion of kernel and range (alongside Properties 6.2.5).

**Properties 7.1.8.** A linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  is one-to-one if and only if the dimension of its kernel  $\text{Ker}(T)$  is zero, i.e.  $\text{Dim}(\text{Ker}(T)) = 0$ . Meanwhile, it is onto if and only if the dimension of range  $\mathcal{R}(T)$  (rank) is same as the dimension of  $\mathcal{V}$  (provided that they are finite).

### 7.1.3 Composition of Linear Transformations

In high-school mathematics we have learnt about the composite of two functions  $(g \circ f)(x) = g(f(x))$  that applies the first function  $f$  first on the input  $x$  and then the second function  $g$  on the intermediate product  $f(x)$ . It does not seem far-fetched that we can also create a composite of two (or even more) linear transformations, each of which takes an input and returns an (intermediate) output (as vectors). We now formally introduce this idea below.

**Definition 7.1.9** (Composition of Linear Transformations). For two (linear) transformations  $T : \mathcal{U} \rightarrow \mathcal{V}$  and  $S : \mathcal{V} \rightarrow \mathcal{W}$ , their composition is defined as  $S \circ T : \mathcal{U} \rightarrow \mathcal{W}$  where for  $\vec{u} \in \mathcal{U}$ ,  $(S \circ T)(\vec{u}) = S(T(\vec{u})) \in \mathcal{W}$ .

From the perspective of compositing linear transformations, we can actually go back to and understand why the usual matrix multiplication is defined in that way.

Since every linear transformation has a matrix representation, according to Definition 7.1.2, we can write  $T(\vec{u}^{(j)}) = \sum_{k=1}^r a_k^{(j)} \vec{v}^{(k)}$  and  $S(\vec{v}^{(k)}) = \sum_{i=1}^m b_i^{(k)} \vec{w}^{(i)}$  where  $\mathcal{U}, \mathcal{V}, \mathcal{W}$  are  $n/r/m$ -dimensional with bases  $\{\vec{u}^{(j)}\}, \{\vec{v}^{(k)}\}, \{\vec{w}^{(i)}\}$  and  $A_{kj} = a_k^{(j)}, B_{ik} = b_i^{(k)}$ . Hence

$$\begin{aligned}
 (S \circ T)(\vec{u}^{(j)}) &= S(T(\vec{u}^{(j)})) = S\left(\sum_{k=1}^r a_k^{(j)} \vec{v}^{(k)}\right) \\
 &= \sum_{k=1}^r a_k^{(j)} S(\vec{v}^{(k)}) \quad (\text{Linearity}) \\
 &= \sum_{k=1}^r a_k^{(j)} \left( \sum_{i=1}^m b_i^{(k)} \vec{w}^{(i)} \right) \\
 &= \sum_{i=1}^m \left( \left( \sum_{k=1}^r b_i^{(k)} a_k^{(j)} \right) \vec{w}^{(i)} \right)
 \end{aligned}$$

which allows us to identify that the matrix representation of  $S \circ T$  with  $\sum_{k=1}^r b_i^{(k)} a_k^{(j)}$ , that is, the matrix product  $BA$  as defined conventionally in Definition 1.1.1 (with the positions of  $A$  and  $B$  switched since  $T$  is applied first, followed by  $S$ ). One of the most frequent scenarios is that  $T : \mathcal{U} \rightarrow \mathcal{V}$  and  $S : \mathcal{V} \rightarrow \mathcal{U}$  are two mappings between two vector spaces, one in the "forward" and another in the "backward" direction. Given their back-and-forth nature, it is curious to ask if they are "invertible" so that applying one after another on a vector will return the same vector, i.e.  $(S \circ T)(\vec{u}) = \text{id}_U(\vec{u}) = \vec{u} \in \mathcal{U}$  and  $(T \circ S)(\vec{v}) = \text{id}_V(\vec{v}) = \vec{v} \in \mathcal{V}$ , where  $\text{id} : \mathcal{U} \rightarrow \mathcal{U}$  (or  $\mathcal{V} \rightarrow \mathcal{V}$ ) denotes the ***identity transformation/identity mapping*** that simply returns the input vector as the output.

**Definition 7.1.10.** For a (linear) transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$ , if there is another linear transformation  $S : \mathcal{V} \rightarrow \mathcal{U}$  such that  $S \circ T = \text{id}_U$  and  $T \circ S = \text{id}_V$ , then  $T$ , and also  $S$ , are known as *invertible* where  $S = T^{-1}$  and  $T = S^{-1}$  are the inverse of each other.

Again, assume that  $\mathcal{U}$  and  $\mathcal{V}$  are  $n/m$ (finite)-dimensional with bases  $\mathcal{B}$  and  $\mathcal{H}$ .

Then both the linear mappings  $T \equiv [T]_B^H$  and  $S \equiv [S]_H^B$ <sup>7</sup> have a corresponding matrix representation with a shape of  $m \times n$  and  $n \times m$ . It is also not hard to accept that  $\text{id}_U \equiv [\text{id}_U]_B^B = I_n$  and  $\text{id}_V \equiv [\text{id}_V]_H^H = I_m$ . Subsequently, for  $S$  to be the inverse of  $T$ ,  $S \circ T = \text{id}_U \equiv I_n = [S]_H^B [T]_B^H$ , and similarly we need  $[T]_B^H [S]_H^B = I_m$ . We claim that  $m = n$ <sup>8</sup>, and thus invertible linear transformations between finite-dimensional vector spaces must require that the vector spaces have the equal dimension and they have a square matrix representation. This suggests that the results in Section 2.2.1 are also applicable for invertible linear transformations here and we restate some of them below.

**Properties 7.1.11.** For an invertible linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  where  $\mathcal{U}$  and  $\mathcal{V}$  are finite-dimensional vector spaces, it is necessary that  $\mathcal{U}$  and  $\mathcal{V}$  have the same number of dimensions  $n$ . Denote its inverse by  $T^{-1} = S : \mathcal{V} \rightarrow \mathcal{U}$ , then  $T^{-1}$  will be unique. The matrix representation of this inverse would be  $[T^{-1}]_H^B = ([T]_B^H)^{-1}$ . Moreover,  $(T^{-1})^{-1} = T$  and  $(X \circ T)^{-1} = T^{-1} \circ X^{-1}$  if  $X : \mathcal{V} \rightarrow \mathcal{W}$  is another invertible linear transformation where  $\mathcal{W}$  is also  $n$ -dimensional.

### 7.1.4 Vector Space Isomorphism to $\mathbb{R}^n$

A linear transformation where both injectivity and surjectivity hold is known as *bijective/isomorphic*. As we will immediately see, this property is very central in relating finite-dimensional real vector spaces to the real  $n$ -space. Combining Properties 7.1.3 and 7.1.4, for a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  to be bijective, every vector  $\vec{v} \in \mathcal{V}$  must be an image which there is one and only one preimage  $\vec{u} \in \mathcal{U}$  is mapped onto, i.e. there is a unique  $\vec{u} \in \mathcal{U}$  that satisfies  $T(\vec{u}) = \vec{v}$  for every  $\vec{v} \in \mathcal{V}$ , which also means that it is *invertible* in the sense that every  $\vec{v} \in \mathcal{V}$  can be traced back to one and only one  $\vec{u} \in \mathcal{U}$  via the inverse transformation  $T^{-1}$  as introduced in the last subsection so that  $T^{-1}(\vec{v}) = \vec{u}$ . Hence it makes sense to

<sup>7</sup>" $\equiv$ " will also be used to associate any linear transformation to its matrix representation.

<sup>8</sup>Assume without loss of generality,  $m < n$ , then by Properties 6.3.14,  $\text{rank}([S]_H^B [T]_B^H) \leq \text{rank}([S]_H^B)$  or  $\text{rank}([T]_B^H) \leq m < n$ , but  $\text{rank}(I_n) = n$  so it is impossible that  $[S]_H^B [T]_B^H = I_n$ .

say a transformation is bijective *between* two same-dimensional vector spaces. There are two major results regarding invertibility here. The first one is

**Theorem 7.1.12.** There always exists a bijective linear mapping between  $\mathcal{V}$  itself, i.e.  $T : \mathcal{V} \rightarrow \mathcal{V}$ , that transforms the coordinates of any fixed vector in  $\mathcal{V}$  between two different bases (denote them by  $\mathcal{B}$  and  $\mathcal{B}'$ ) of its. Such a change of coordinates in  $\mathcal{V}$  has a matrix representation  $[T]_B^{B'} = P_B^{B'}$  that is invertible.

*Proof.* Since it is the same vector space  $\mathcal{V}$  but just represented in different bases, the number of dimension will stay the same, let's say  $n$ , and the bases  $\mathcal{B}$  and  $\mathcal{B}'$  both are made up of  $n$  basis vectors (Properties 6.2.4). Denote them by  $\mathcal{B} = \{\vec{v}_B^{(1)}, \vec{v}_B^{(2)}, \dots, \vec{v}_B^{(n)}\}$  and  $\mathcal{B}' = \{\vec{v}_{B'}^{(1)}, \vec{v}_{B'}^{(2)}, \dots, \vec{v}_{B'}^{(n)}\}$ . The desired mapping  $T : \mathcal{V} \rightarrow \mathcal{V}$  is in fact

$$T(\vec{v}) = \text{id}(\vec{v}) = \vec{v}$$

the *identity mapping* as it is just a change of coordinates where the actual vector stays identical. This transformation is then trivially bijective because any vector is just mapped into itself, and is described by  $[\vec{v}]_{B'} = [T]_B^{B'} [\vec{v}]_B$  following Definition 7.1.2 with  $\mathcal{U} = \mathcal{V}$  and  $\vec{u} = \vec{v}$ . Now note that  $[\vec{v}]_{B'} = [T]_B^{B'} [\vec{v}]_B$  has a unique solution  $[\vec{v}]_B$  for any  $[\vec{v}]_{B'}$  as  $T$  is bijective and by definition, together with the fact that the coordinate representation of a vector in any basis is unique, each of  $[\vec{v}]_{B'}$  is mapped onto by one and only one  $[\vec{v}]_B$ . Part (d) to (a) of Theorem 3.2.1 then shows that  $[T]_B^{B'}$  is an invertible matrix. According to the discussion prior to Definition 7.1.2,  $[T]_B^{B'}$  takes the form of

$$\begin{aligned} P_B^{B'} &= [T]_B^{B'} = \left[ [\text{id}(\vec{v}_B^{(1)})]_{B'} | [\text{id}(\vec{v}_B^{(2)})]_{B'} | \cdots | [\text{id}(\vec{v}_B^{(n)})]_{B'} \right] \\ &= \left[ [\vec{v}_B^{(1)}]_{B'} | [\vec{v}_B^{(2)}]_{B'} | \cdots | [\vec{v}_B^{(n)}]_{B'} \right] \end{aligned}$$

So we have to find how each of the basis vectors in  $\mathcal{B}$  is expressed in the  $\mathcal{B}'$  system. Conversely,

$$P_{B'}^B = ([T]_B^{B'})^{-1} = [T]_{B'}^B = \left[ [\vec{v}_{B'}^{(1)}]_B | [\vec{v}_{B'}^{(2)}]_B | \cdots | [\vec{v}_{B'}^{(n)}]_B \right]$$

□

Be aware that despite  $\text{id}$  being an identity mapping, the exact matrix representation is dependent on the bases (but the effect of such an identity mapping is obviously basis-independent) and will usually not be an identity matrix. Nevertheless, such bijectivity between any two coordinate systems of the same vector space, which is always accompanied by some invertible transformation matrix, implies that all vectors in any vector space  $\vec{u} \in \mathcal{U}$  (or  $\vec{v} \in \mathcal{V}$ ) and linear mappings  $T : \mathcal{U} \rightarrow \mathcal{V}$  from one vector space to another, together with its (dimensions of) kernel or range, are independent of the choices of bases for either  $\mathcal{U}$  or  $\mathcal{V}$  and we can pick whatever bases that suit the situation better. This also means that statements (both previous and to be derived in the future) for finite-dimensional linear transformations are applicable for corresponding matrices (when converted properly) and vice versa. The only thing that is dependent on the coordinate systems will be their coordinate representation and we will see how it unfolds in the next part. This justify our fixing of bases during several arguments in the last subsection.

**Example 7.1.6.** Show that  $\mathcal{B} = \{(1, 0, 1)^T, (0, 2, 1)^T, (-1, 1, 2)^T\}$  and  $\mathcal{B}' = \{(0, 0, 1)^T, (2, 0, 1)^T, (1, -1, 0)^T\}$  are both bases for  $\mathcal{V} = \mathbb{R}^3$  and find the matrix representation of coordinate conversion between them.

*Solution.* Just like in Example 6.2.1, we need to check whether the determinants of

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

are non-zero or not. A simple computation shows that  $\det(B) = 5$  and  $\det(B') = -2$  and thus both  $\mathcal{B}$  and  $\mathcal{B}'$  are bases for  $\mathbb{R}^3$ . By Theorem 7.1.12, the matrix representation for the change of basis abides

$$[\text{id}]_B^{B'} = [T]_B^{B'} = \left[ [\vec{v}_B^{(1)}]_{B'} | [\vec{v}_B^{(2)}]_{B'} | \cdots | [\vec{v}_B^{(n)}]_{B'} \right]$$

where each of  $[\vec{v}_B^{(j)}]_{B'}$  is found via the equation

$$[(\vec{v}_B^{(j)})_1]_{B'} (\vec{v}_{B'}^{(1)}) + [(\vec{v}_B^{(j)})_2]_{B'} (\vec{v}_{B'}^{(2)}) + [(\vec{v}_B^{(j)})_3]_{B'} (\vec{v}_{B'}^{(3)}) = \vec{v}_B^{(j)}$$

just as in Example 6.2.1 with  $[(\vec{v}_B^{(j)})_i]_{B'}$  being the  $i$ -th component of  $\vec{v}_B^{(j)}$  in the  $B'$  frame, or equivalently,

$$\begin{aligned} \left[ \vec{v}_{B'}^{(1)} | \vec{v}_{B'}^{(2)} | \vec{v}_{B'}^{(3)} \right] \begin{bmatrix} [(\vec{v}_B^{(j)})_1]_{B'} \\ [(\vec{v}_B^{(j)})_2]_{B'} \\ [(\vec{v}_B^{(j)})_3]_{B'} \end{bmatrix} &= \vec{v}_B^{(j)} \\ [\vec{v}_B^{(j)}]_{B'} &= \begin{bmatrix} [(\vec{v}_B^{(j)})_1]_{B'} \\ [(\vec{v}_B^{(j)})_2]_{B'} \\ [(\vec{v}_B^{(j)})_3]_{B'} \end{bmatrix} = \left[ \vec{v}_{B'}^{(1)} | \vec{v}_{B'}^{(2)} | \vec{v}_{B'}^{(3)} \right]^{-1} \vec{v}_B^{(j)} \\ &= B'^{-1} \vec{v}_B^{(j)} \end{aligned}$$

Subsequently,

$$\begin{aligned} [T]_B^{B'} &= \left[ [\vec{v}_B^{(1)}]_{B'} | [\vec{v}_B^{(2)}]_{B'} | \cdots | [\vec{v}_B^{(n)}]_{B'} \right] \\ &= \left[ B'^{-1} \vec{v}_B^{(1)} | B'^{-1} \vec{v}_B^{(2)} | B'^{-1} \vec{v}_B^{(3)} \right] \\ &= B'^{-1} \left[ \vec{v}_B^{(1)} | \vec{v}_B^{(2)} | \vec{v}_B^{(3)} \right] \\ &= B'^{-1} B \end{aligned}$$

The readers should verify that we can indeed factor out the  $B'^{-1}$  from the columns and put it to the left in the third line (see Footnote 10 of Chapter 6), and the required matrix representation for the coordinate change is

$$\begin{aligned} P_B^{B'} &= [T]_B^{B'} = B'^{-1} B = \begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 2 \\ \frac{1}{2} & 1 & 0 \\ 0 & -2 & -1 \end{bmatrix} \end{aligned}$$

Let's take  $(2, 2, 3)^T = 2(1, 0, 1)^T + 1(0, 2, 1)^T + 0(-1, 1, 2)^T = (2, 1, 0)_B^T$  for

double-checking:

$$P_B^{B'}(2, 1, 0)_B^T = \begin{bmatrix} \frac{1}{2} & 0 & 2 \\ \frac{1}{2} & 1 & 0 \\ 0 & -2 & -1 \end{bmatrix}_B^{B'} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}_B = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}_{B'}^T$$

and indeed  $(2, 2, 3)^T = 1(0, 0, 1)^T + 2(2, 0, 1)^T + (-2)(1, -1, 0)^T = (1, 2, -2)_{B'}^T$ .

□

More generally, the relation  $P_B^{B'} = [\text{id}]_B^{B'} = B'^{-1}B$  still remains valid where  $B$  and  $B'$  are matrices composed by the basis vectors from the  $\mathcal{B}$  and  $\mathcal{B}'$  systems, relative to a third basis (without loss of generality we have assumed it is the standard basis  $\mathcal{S}$ <sup>9</sup>, but the readers are advised to extend this for any other arbitrary basis), that are arranged in columns. To see this from another perspective, take any vector  $\vec{v}$  that is expressed in the  $\mathcal{B}$  coordinates,  $[\vec{v}]_B$ . We can view the change in coordinates from  $\mathcal{B}$  to  $\mathcal{B}'$  in two steps: first from  $\mathcal{B}$  to  $\mathcal{S}$ , and then from  $\mathcal{S}$  to  $\mathcal{B}'$ . From Section 6.2.1, we already know that the former constitutes  $[\vec{v}]_S = B[\vec{v}]_B$ , and the latter is done by  $[\vec{v}]_{B'} = B'^{-1}[\vec{v}]_S$ . Combining these two operations together we have  $[\vec{v}]_{B'} = B'^{-1}[\vec{v}]_S = B'^{-1}B[\vec{v}]_B$  and hence  $[\text{id}]_B^{B'} = B'^{-1}B$ .

The second major result in this subsection is

**Theorem 7.1.13.** There is always a bijective linear mapping between  $\mathcal{V}$  and  $\mathbb{R}^n$  where  $\mathcal{V}$  is any  $n$ -dimensional real vector space. In this sense we say  $\mathcal{V}$ /such a mapping is *isomorphic*/an *isomorphism* to  $\mathbb{R}^n$  that has an invertible matrix representation. Conversely, if a matrix representation of a linear transformation is invertible, the linear transformation must be bijective.

*Proof.* We construct such a mapping explicitly. Note that  $\mathcal{V}$  and  $\mathbb{R}^n$  are both  $n$ -dimensional vector spaces and any of their bases will contain  $n$  basis vectors.

<sup>9</sup>Unfortunately, as you may notice, there is actually no satisfying "standard" of what really is a standard basis for (real) finite-dimensional vector space other than the real  $n$ -space since any basis can be regarded to be one with respect to itself. Here we just pretend it is available for the sake of reasoning.

Denote the basis chosen for  $\mathcal{V}$  by  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \dots, \vec{v}^{(n)}\}$  and we use the standard basis  $\mathcal{S} = \{\hat{e}^{(1)}, \hat{e}^{(2)}, \dots, \hat{e}^{(n)}\}$  for  $\mathbb{R}^n$ . Then the linear mapping  $T : \mathcal{V} \rightarrow \mathbb{R}^n$  that abides

$$T(\vec{v}^{(j)}) = \hat{e}^{(j)}$$

where  $j = 1, 2, \dots, n$ , is bijective as desired. To see this, by Theorem 7.1.5, as for every  $\vec{v}^{(j)}$ ,  $T(\vec{v}^{(j)}) = \hat{e}^{(j)}$  leads to the standard unit vectors that are linearly independent,  $T$  is one-to-one. Meanwhile, a direct use of Theorem 7.1.6 over the defined association  $T(\vec{v}^{(j)}) = \hat{e}^{(j)}$  for each of the  $\hat{e}^{(j)}$  immediately shows that  $T$  is onto. Since  $T$  is now one-to-one and onto, it is bijective, and has an inverse  $T^{-1}$ . Again, the bijectivity, in addition to the uniqueness of basis coordinates, implies that for any  $\vec{x} \in \mathbb{R}^n$ ,  $[\vec{x}]_S = [T]_B^S [\vec{v}]_B$  has a unique solution  $[\vec{v}]_B$ , and part (d) to (a) of Theorem 3.2.1 then shows that the matrix representation  $[T]_B^S$  is invertible where  $T^{-1} \equiv ([T]_B^S)^{-1}$ . The converse follows the same argument running in opposite direction.  $\square$

This theorem enables us to identify and treat any finite-dimensional real vector space  $\mathcal{V}$  as the real  $n$ -space  $\mathbb{R}^n$  with  $n$  being the dimension of  $\mathcal{V}$ . **Thus we can work with  $\mathcal{V}$  as if it is  $\mathbb{R}^n$  and the results for  $\mathbb{R}^n$  derived in this and the last chapter are all applicable on other  $n$ -dimensional real vector spaces with an appropriate transformation.** Actually, we have been implicitly utilizing this isomorphism relation in many of our previous examples, e.g. writing out the coordinates of a vector from an  $n$ -dimensional vector space with  $n$  components like an  $\mathbb{R}^n$  vector. As a corollary,

**Properties 7.1.14.** Any two (real) finite-dimensional vector spaces are isomorphic such that there exists a bijective linear transformation between them, if and only if they have the same number of dimension. Otherwise, there will be no isomorphism between those with different dimensions.

The "if" direction is easy to see because they are both isomorphic to  $\mathbb{R}^n$  by Theorem 7.1.13 and bijectivity is transitive.<sup>10</sup> For the "only if" direction, let

---

<sup>10</sup>Let's say  $S: \mathcal{U} \rightarrow \mathbb{R}^n$  and  $T: \mathcal{V} \rightarrow \mathbb{R}^n$  are the two respective isomorphisms. Then  $T^{-1} \circ S: \mathcal{U} \rightarrow \mathcal{V}$  will be the required bijective linear transformation.

the two vector spaces  $\mathcal{U}$  and  $\mathcal{V}$  have dimensions of  $m$  and  $n$  respectively, and without loss of generality  $m < n$ . Then they can never be isomorphic since given any transformation  $T: \mathcal{U} \rightarrow \mathcal{V}$  the  $m$  transformed basis vectors  $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \dots, T(\vec{u}^{(m)})$  which will be unable to span the  $n$ -dimensional  $\mathcal{V}$  and by Properties 7.1.4 all of them are not surjective.

**Example 7.1.7.** Explicitly show that  $\mathcal{U} = \mathcal{P}^3$  and  $\mathcal{V} = \text{span}(\mathcal{H})$ , where  $\mathcal{H} = \{e^x, xe^x, x^2e^x, x^3e^x\}$ , are isomorphic by considering  $T: \mathcal{U} \rightarrow \mathcal{V}$ ,  $T[p(x)] = \int_{-\infty}^x e^x p(x) dx$ .

*Solution.* It is clear that both  $\mathcal{U}$  and  $\mathcal{V}$  are four-dimensional and by the above corollary they are isomorphic. Take  $\mathcal{B} = \{1, x, x^2, x^3\}$  the standard polynomial basis for  $\mathcal{U} = \mathcal{P}^3$  and the linearly independent  $\mathcal{H}$  is automatically the basis for  $\mathcal{V}$ . Now we compute the matrix representation  $[T]_B^H$  as follows. By elementary calculus,

$$\begin{aligned} T(1) &= \int_{-\infty}^x e^x dx = e^x \\ T(x) &= \int_{-\infty}^x xe^x dx = xe^x - e^x \\ T(x^2) &= \int_{-\infty}^x x^2 e^x dx = x^2 e^x - 2xe^x + 2e^x \\ T(x^3) &= \int_{-\infty}^x x^3 e^x dx = x^3 e^x - 3x^2 e^x + 6xe^x - 6e^x \end{aligned}$$

and thus

$$[T]_B^H = \begin{bmatrix} 1 & -1 & 2 & -6 \\ 0 & 1 & -2 & 6 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is an upper-triangular matrix and its determinant is simply the product of diagonal entries  $(1)^4 = 1 \neq 0$ . Therefore, by Theorem 3.2.1,  $[T]_B^H$  is invertible

and the given transformation, as well as  $\mathcal{U}$  and  $\mathcal{V}$  themselves, is/are isomorphic according to Theorem 7.1.13.  $\square$

Short Exercise: Redo the above example by considering  $T[p(x)] = e^x p(x)$  this time.<sup>11</sup>

## 7.2 Additional Discussions about Coordinate Bases

### 7.2.1 Linear Change of Coordinates

In previous parts we have already mentioned about change of coordinates between bases for several times, where such a mapping are confined to be linear just like other transformations discussed. In this section we will dive deeper into the details and address two distinct scenarios: change of coordinates for vectors and linear transformations (matrices).

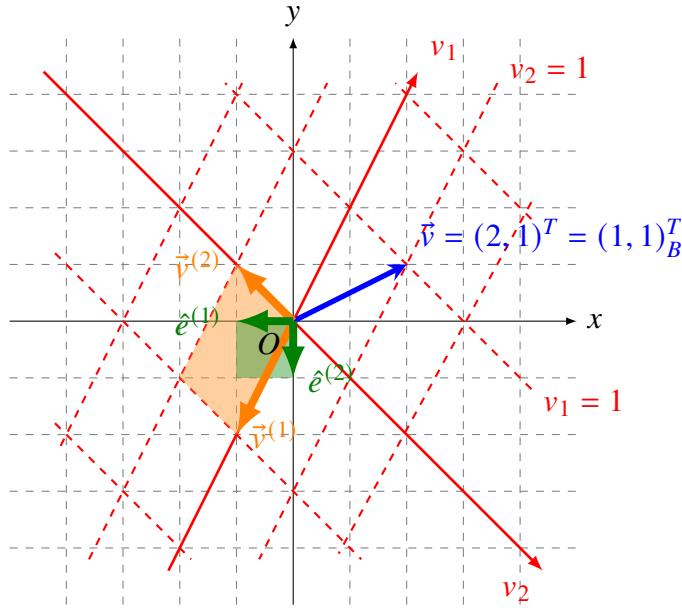
#### Change of Coordinates for Vectors

The procedure about change of coordinates for vectors have been discussed substantially in Examples 6.2.1, 7.1.6 and explained through Theorem 7.1.12. Here we will focus on its geometric interpretation instead, which will be illustrated by the small example below.

**Example 7.2.1.** Consider the vector space of  $\mathbb{R}^2$  as the  $x$ - $y$  plane. Given a basis  $\mathcal{B}$  for  $\mathbb{R}^2$  that is consisted of two vectors  $\vec{v}^{(1)} = (1, 2)^T$  and  $\vec{v}^{(2)} = (1, -1)^T$ , convert the coordinates of the vector  $\vec{v} = (2, 1)^T$  from the standard basis  $\mathcal{S}$  to  $\mathcal{B}$ .

---

<sup>11</sup>It becomes trivial and the matrix representation is simply the identity matrix.



*Solution.* As before,  $P_B^S = [\vec{v}^{(1)} | \vec{v}^{(2)}]$ , and it can be seen that

$$P_B^S = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \quad P_S^B = (P_B^S)^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}$$

Hence the coordinates of  $\vec{v}$  in the  $\mathcal{B}$  system is

$$[\vec{v}]_B = P_S^B [\vec{v}]_S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}_S^B \begin{bmatrix} 2 \\ 1 \end{bmatrix}_S = \begin{bmatrix} 1 \\ 1 \end{bmatrix}_B$$

The geometry of this problem is shown in the figure on the next page where each grid line separation represents one unit length of the axis vectors.  $\square$

In this example, we can see that in the two bases  $\mathcal{S}$ ,  $\mathcal{B}$ , their axis vectors (reversed in the figure) can be converted via  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  with  $T(\hat{e}^{(j)}) = \vec{v}^{(j)}$ . The corresponding matrix representation is  $\vec{v}^{(j)} = [\vec{v}^{(1)} | \vec{v}^{(2)}] \hat{e}^{(j)} = P_B^S \hat{e}^{(j)}$ . Meanwhile the coordinate transformation of a fixed vector follows  $[\vec{v}]_B = (P_B^S)^{-1} [\vec{v}]_S = P_S^B [\vec{v}]_S$ , where the matrix involved is the inverse of the former.

The former case actually alters the vectors themselves and is sometimes known as an ***active (coordinate) transformation***. In contrast, the latter only changes the coordinate frame but keep the vector unchanged and is hence called a ***passive (coordinate) transformation*** (in fact, it is just the identity transformation with a change of basis). We can see that in the example above, after the active transformation the area of square formed by the new two basis vectors is enlarged by a factor of  $|\det(P_B^S)| = 3$ . It is a result of Properties 5.3.4 and the similar holds for cases of any dimension. Oppositely, with the passive transformation we can say that the value of area of an identical square is shrinked to  $|\det(P_S^B)| = |\det((P_B^S)^{-1})| = |\det(P_B^S)|^{-1} = \frac{1}{3}$  of the original, expressed in the new basis units. Therefore, the appropriate factors in the two scenarios are the reciprocal of each other.

### Change of Coordinates for Linear Transformations/Matrices

It is also possible to do a change of coordinates for linear transformations and hence the matrices that represent them. Consider a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  that has a matrix representation of  $[\vec{v}]_H = [T]_B^H [\vec{u}]_B$  where  $\mathcal{B}$  and  $\mathcal{H}$  are bases for  $\mathcal{U}$  and  $\mathcal{V}$  respectively. If we want to change the basis for  $\mathcal{U}$  from  $\mathcal{B}$  to some other basis  $\mathcal{B}'$  (and similarly  $\mathcal{H}'$  for  $\mathcal{V}$ ), then the new matrix representation of the linear transformation would be  $[\vec{v}]_{H'} = [T]_{B'}^{H'} [\vec{u}]_{B'}$ . Since they are the same transformation but only expressed in different coordinate systems, these two matrix equations have to be equivalent. Now, the vectors on both sides of the original equation themselves can undergo changes of coordinates according to the previous Theorem 7.1.12 with  $[\vec{u}]_B = [\text{id}]_{B'}^B [\vec{u}]_{B'} = P_{B'}^B [\vec{u}]_{B'}$  and  $[\vec{v}]_H = [\text{id}]_{H'}^H [\vec{v}]_{H'} = Q_{H'}^H [\vec{v}]_{H'}$ , where we denote the change of coordinates matrices from  $\mathcal{B}'$  to  $\mathcal{B}$  by  $P_{B'}^B$  (and similarly  $\mathcal{H}'$  to  $\mathcal{H}$  by  $Q_{H'}^H$ ). Subsequently,

$$\begin{aligned} [\vec{v}]_H &= [T]_B^H [\vec{u}]_B \\ Q_{H'}^H [\vec{v}]_{H'} &= [T]_B^H P_{B'}^B [\vec{u}]_{B'} \\ [\vec{v}]_{H'} &= ((Q_{H'}^H)^{-1} [T]_B^H P_{B'}^B) [\vec{u}]_{B'} \end{aligned}$$

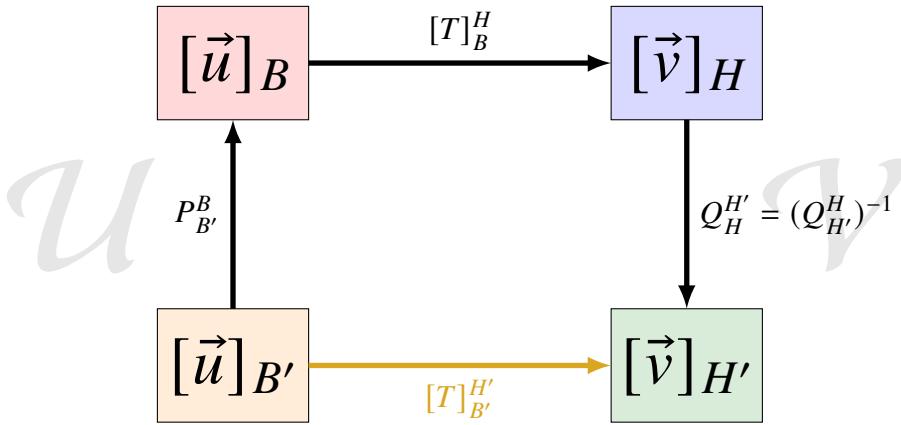


Figure 7.1: A schematic showing how the change of coordinate bases works for linear transformation.

Comparing with the assumed form of  $[\vec{v}]_{H'} = [T]_{B'}^{H'} [\vec{u}]_{B'}$ , we can identify  $[T]_{B'}^{H'}$  with  $(Q_{H'}^H)^{-1} [T]_B^H P_{B'}^B$ , and this is the desired formula for change of coordinates over the matrix form of a linear transformation.

**Properties 7.2.1.** The change of coordinates for the matrix representation of a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  from bases  $\mathcal{B}$  and  $\mathcal{H}$  to  $\mathcal{B}'$  and  $\mathcal{H}'$  for  $\mathcal{U}$  and  $\mathcal{V}$  respectively follows the relation

$$[T]_{B'}^{H'} = (Q_{H'}^H)^{-1} [T]_B^H P_{B'}^B$$

where  $P_{B'}^B$  and  $Q_{H'}^H$  are matrices for change of coordinates on vectors from  $\mathcal{B}'$  to  $\mathcal{B}$  and  $\mathcal{H}'$  to  $\mathcal{H}$  individually.

Another way to derive the above formula is to consider the linear transformation with respect to the basis  $\mathcal{B}'$  to  $\mathcal{H}'$  as three smaller steps: firstly, convert the input vector from the basis  $\mathcal{B}'$  back to  $\mathcal{B}$  ( $P_{B'}^B$ ); subsequently, carry out the transformation in terms of  $\mathcal{B}$  and  $\mathcal{H}$  ( $[T]_B^H$ ); finally, map the vector from the basis  $\mathcal{H}$  to  $\mathcal{H}'$  ( $Q_H^{H'} = (Q_{H'}^H)^{-1}$ ). This flow is illustrated in the schematic of Figure 7.1.

**Example 7.2.2.** Use Properties 7.2.1 to redo Example 7.1.2 with respect to new bases  $\mathcal{B}' = \{1, x - 1, (x - 1)^2\}$  and  $\mathcal{H}' = \{1, x + 1\}$ .

*Solution.* First it is instructive to find  $P_{B'}^B$  and  $Q_{H'}^H$ . We leave to the readers to verify that

$$P_{B'}^B = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \quad Q_{H'}^H = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and hence by Properties 7.2.1,

$$\begin{aligned} [T]_{B'}^{H'} &= (Q_{H'}^H)^{-1} [T]_B^H P_{B'}^B \\ &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -4 \\ 0 & 0 & 2 \end{bmatrix} \end{aligned}$$

We use a test case to check the answer. Let  $p(x) = x^2 - 3x + 1 = (x - 1)^2 - (x - 1) - 1$ . Then its coordinates in the  $\mathcal{B}'$  basis is  $(-1, -1, 1)_{B'}^T$ , and the transformation can be described by

$$[T]_{B'}^{H'} (1, -1, -1)_{B'}^T = \begin{bmatrix} 0 & 1 & -4 \\ 0 & 0 & 2 \end{bmatrix}_{B'}^{H'} \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}_{B'} = \begin{bmatrix} -5 \\ 2 \end{bmatrix}_{H'}$$

which corresponds to  $-5(1) + 2(x + 1) = 2x - 3$ , which is consistent with the usual calculation  $T(p(x)) = p'(x) = (x^2 - 3x + 1)' = 2x - 3$  from elementary calculus.  $\square$

In most of the times, we are interested in the type of linear transformations that are an **endomorphism** (sometimes also referred to as a **(linear) operator**) in

which the mapping is from a vector space  $\mathcal{V}$  to itself, i.e.  $T : \mathcal{V} \rightarrow \mathcal{V}$ <sup>12</sup>. Often we also use the same basis  $\mathcal{B}$  for the input and output vector. Subsequently, to change the basis for both of them at the same time, let's say  $\mathcal{B}'$ , if the matrix for change of coordinates on vectors from  $\mathcal{B}'$  to  $\mathcal{B}$  is denoted as  $P = P_{\mathcal{B}'}^{\mathcal{B}}$ , then Properties 7.2.1 is reduced to  $[T]_{\mathcal{B}'}^{B'} = (P_{\mathcal{B}'}^{\mathcal{B}})^{-1}[T]_B^B P_{\mathcal{B}'}^{\mathcal{B}} = P^{-1}AP$  where  $A = [T]_B^B$  is the original matrix representation of the endomorphism. When it is clear from the context, we will simply write  $[T]_B^B ([T]_{\mathcal{B}'}^{B'})$  as  $[T]_B ([T]_{\mathcal{B}'})$ .

**Properties 7.2.2.** For a linear endomorphism/operator  $T : \mathcal{V} \rightarrow \mathcal{V}$ , the change of coordinates for its matrix representation from the old basis  $\mathcal{B}$  to the new one  $\mathcal{B}'$  is described by the formula

$$[T]_{\mathcal{B}'} = (P_{\mathcal{B}'}^{\mathcal{B}})^{-1}[T]_B^B P_{\mathcal{B}'}^{\mathcal{B}}$$

Speaking loosely, the change of coordinates for a matrix in general takes the form of

$$A' = P^{-1}AP$$

In this case,  $A'$  and  $A$  (or  $[T]_{\mathcal{B}'}$  and  $[T]_B$ ) are said to be *similar*.

**Example 7.2.3.** For a two-dimensional vector space  $\mathcal{V}$  with a basis  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}\}$ , if a linear endomorphism  $T : \mathcal{V} \rightarrow \mathcal{V}$  is defined by  $T(\vec{v}^{(1)}) = \vec{v}^{(1)}$ ,  $T(\vec{v}^{(2)}) = \vec{v}^{(1)} + \vec{v}^{(2)}$ , finds its matrix representation with respect to  $\mathcal{B}$ . Subsequently, if a new basis  $\mathcal{B}'$  is formed by  $\{\vec{v}^{(1)'}, \vec{v}^{(2)'}\}$  where  $\vec{v}^{(1)'} = 2\vec{v}^{(1)} - \vec{v}^{(2)}$  and  $\vec{v}^{(2)'} = -\vec{v}^{(1)} + \vec{v}^{(2)}$ , use Properties 7.2.2 to compute the matrix representation of the endomorphism with respect to the new basis.

*Solution.* By Definition 7.1.2, the linear transformation has a matrix representation of

$$[T]_B = [[T(\vec{v}^{(1)})]_B | [T(\vec{v}^{(2)})]_B] = [[\vec{v}^{(1)}]_B | [\vec{v}^{(1)} + \vec{v}^{(2)}]_B]$$

---

<sup>12</sup>An endomorphism that is at the same time an isomorphism is known as an *automorphism*, e.g. the linear transformation in Example 7.2.3.

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

with respect to the old basis  $\mathcal{B}$ . The appropriate  $P_{B'}^B$  matrix that will be used for Properties 7.2.2, by Theorem 7.1.12, is

$$\begin{aligned} P_{B'}^B &= [[\vec{v}^{(1)'}]_B | [\vec{v}^{(2)'}]_B] = [[2\vec{v}^{(1)} - \vec{v}^{(2)}]_B | [-\vec{v}^{(1)} + \vec{v}^{(2)}]_B] \\ &= \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \end{aligned}$$

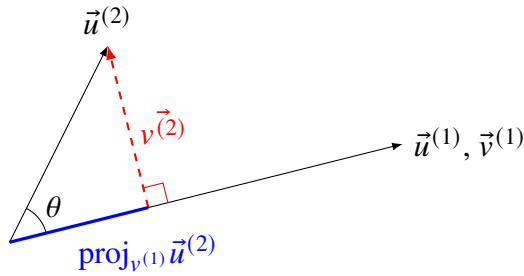
and thus the desired new matrix representation of the endomorphism with respect to  $\mathcal{B}'$  is

$$\begin{aligned} [T]_{B'} &= (P_{B'}^B)^{-1} [T]_B P_{B'}^B \\ &= \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \end{aligned}$$

□

## 7.2.2 Gram-Schmidt Orthogonalization, QR Decomposition

Sometimes the coordinate basis consists of vectors that are linearly independent but not orthogonal to each other, unlike the standard basis. A common way to create an orthogonal basis from the set is to apply the so-called **Gram-Schmidt Orthogonalization**. Basically, it is an iterative method. At each step it constructs a vector that are orthogonal to all the previously processed vectors by removing the parallel components projected onto them (blue in the figure below) while retaining the orthogonal part (red).



**Definition 7.2.3** (Algorithm for Gram-Schmidt Orthogonalization). Given a coordinate basis consisted of  $\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \dots, \vec{u}^{(n)} \in \mathbb{R}^m$ , Gram-Schmidt Orthogonalization transforms them into  $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(n)} \in \mathbb{R}^m$  ( $m$  and  $n$  are not necessarily equal, specifically  $n \leq m$ ) according to the following formulae:

$$\vec{v}^{(1)} = \vec{u}^{(1)}$$

$$\vec{v}^{(2)} = \vec{u}^{(2)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(2)} = \vec{u}^{(2)} - \frac{\vec{u}^{(2)} \cdot \vec{v}^{(1)}}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)}$$

$$\vec{v}^{(3)} = \vec{u}^{(3)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(3)} - \text{proj}_{\vec{v}^{(2)}} \vec{u}^{(3)} = \vec{u}^{(3)} - \frac{\vec{u}^{(3)} \cdot \vec{v}^{(1)}}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\vec{u}^{(3)} \cdot \vec{v}^{(2)}}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)}$$

$$\vdots$$

$$\begin{aligned} \vec{v}^{(n)} &= \vec{u}^{(n)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(n)} - \text{proj}_{\vec{v}^{(2)}} \vec{u}^{(n)} - \cdots - \text{proj}_{\vec{v}^{(n-1)}} \vec{u}^{(n)} \\ &= \vec{u}^{(n)} - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(1)}}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(2)}}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)} - \cdots - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(n-1)}}{\|\vec{v}^{(n-1)}\|^2} \vec{v}^{(n-1)} \end{aligned}$$

In general, for  $j \geq 2$ , the  $j$ -th new vector is computed by

$$\vec{v}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \text{proj}_{\vec{v}^{(k)}} \vec{u}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \frac{\vec{u}^{(j)} \cdot \vec{v}^{(k)}}{\|\vec{v}^{(k)}\|^2} \vec{v}^{(k)}$$

where the expression of vector projection, Properties 5.2.1, is used.

A variant of Gram-Schmidt Orthogonalization is to normalize every vector at each step immediately, such that  $\|\hat{v}^{(j)}\| = 1$  for all  $j$ , and the resulted basis is said to be **orthonormal** (both orthogonal and of unit length). The formulae in Definition 7.2.3 are then reduced to

**Definition 7.2.4** (Gram-Schmidt Orthogonalization with Normalization).

$$\begin{aligned}\hat{v}^{(1)} &= \frac{\vec{u}^{(1)}}{\|\vec{u}^{(1)}\|} \\ \hat{v}^{(2)} &= \frac{\vec{u}^{(2)} - (\vec{u}^{(2)} \cdot \hat{v}^{(1)})\hat{v}^{(1)}}{\|(\vec{u}^{(2)} - \vec{u}^{(2)} \cdot \hat{v}^{(1)})\hat{v}^{(1)}\|} \\ \hat{v}^{(3)} &= \frac{\vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)})\hat{v}^{(2)}}{\|(\vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)})\hat{v}^{(2)})\|} \\ &\vdots \\ \hat{v}^{(n)} &= \frac{\vec{u}^{(n)} - (\vec{u}^{(n)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(n)} \cdot \hat{v}^{(2)})\hat{v}^{(2)} - \dots - (\vec{u}^{(n)} \cdot \hat{v}^{(n-1)})\hat{v}^{(n-1)}}{\|(\vec{u}^{(n)} - (\vec{u}^{(n)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(n)} \cdot \hat{v}^{(2)})\hat{v}^{(2)} - \dots - (\vec{u}^{(n)} \cdot \hat{v}^{(n-1)})\hat{v}^{(n-1)})\|}\end{aligned}$$

For  $j \geq 2$ , the general formula is

$$\hat{v}^{(j)} = \frac{\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}}{\left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} \right\|}$$

**Example 7.2.4.** Perform Gram-Schmidt Orthogonalization with normalization on the coordinate basis for  $\mathbb{R}^3$  that is consisted of  $\vec{u}^{(1)} = (1, 2, 2)^T$ ,  $\vec{u}^{(2)} = (1, -1, 0)^T$ ,  $\vec{u}^{(3)} = (3, -1, 1)^T$ , using the formula in Definition 7.2.4.

*Solution.* The first vector is

$$\hat{v}^{(1)} = \frac{1}{\sqrt{1^2 + 2^2 + 2^2}} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$$

The second vector can be found via

$$\begin{aligned}
 \vec{u}^{(2)} - (\vec{u}^{(2)} \cdot \hat{v}^{(1)}) \hat{v}^{(1)} &= \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} - [(1)(\frac{1}{3}) + (-1)(\frac{2}{3}) + (0)(\frac{2}{3})] \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \\
 &= \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} - (-\frac{1}{3}) \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} = \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix} \\
 \hat{v}^{(2)} &= \frac{1}{\sqrt{(\frac{10}{9})^2 + (-\frac{7}{9})^2 + (\frac{2}{9})^2}} \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix} = \frac{3}{\sqrt{17}} \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix} = \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix}
 \end{aligned}$$

By the same essence, we have the third vector as

$$\begin{aligned}
 \vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)}) \hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)}) \hat{v}^{(2)} &= \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} - [(3)(\frac{1}{3}) + (-1)(\frac{2}{3}) + (1)(\frac{2}{3})] \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \\
 &\quad - [(3)(\frac{10}{3\sqrt{17}}) + (-1)(-\frac{7}{3\sqrt{17}}) + (1)(\frac{2}{3\sqrt{17}})] \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix} \\
 &= \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} - \frac{13}{\sqrt{17}} \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix} = \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix} \\
 \hat{v}^{(3)} &= \frac{1}{\sqrt{(\frac{2}{17})^2 + (\frac{2}{17})^2 + (-\frac{3}{17})^2}} \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix} = \sqrt{17} \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{17}} \\ \frac{2}{\sqrt{17}} \\ -\frac{3}{\sqrt{17}} \end{bmatrix}
 \end{aligned}$$

□

Short Exercise: Verify that  $\hat{v}^{(1)}, \hat{v}^{(2)}, \hat{v}^{(3)}$  are pairwise orthogonal.<sup>13</sup>

A major application of the Gram-Schmidt Orthogonalization is the ***QR Decomposition***, which factors a matrix into two matrices, one as the orthogonal basis vectors acquired from the Gram-Schmidt process arranged in columns and another one as a upper-triangular matrix (non-zero elements only found along or above the main diagonal) where the elements take the form of  $\vec{u}^{(j)} \cdot \hat{v}^{(i)}$  as shown below. This is very useful in the processing of large matrices and least-square error fitting.

**Properties 7.2.5.** For a matrix  $A = [\vec{u}^{(1)} | \vec{u}^{(2)} | \vec{u}^{(3)} | \dots | \vec{u}^{(n)}]$ , and the matrix  $Q = [\hat{v}^{(1)} | \hat{v}^{(2)} | \hat{v}^{(3)} | \dots | \hat{v}^{(n)}]$ , where the  $\hat{v}_j \in \mathbb{R}^m$  are orthonormal vectors that come from carrying out Gram-Schmidt orthogonalization on the basis vectors  $\vec{u}_j \in \mathbb{R}^m$ ,  $m \leq n$  and  $j = 1, 2, \dots, n$ , according to the Definition 7.2.4, we have  $A = QR$ , where

$$R = \begin{bmatrix} \vec{u}^{(1)} \cdot \hat{v}^{(1)} & \vec{u}^{(2)} \cdot \hat{v}^{(1)} & \vec{u}^{(3)} \cdot \hat{v}^{(1)} & \dots & \vec{u}^{(n)} \cdot \hat{v}^{(1)} \\ 0 & \vec{u}^{(2)} \cdot \hat{v}^{(2)} & \vec{u}^{(3)} \cdot \hat{v}^{(2)} & & \vec{u}^{(n)} \cdot \hat{v}^{(2)} \\ 0 & 0 & \vec{u}^{(3)} \cdot \hat{v}^{(3)} & & \vec{u}^{(n)} \cdot \hat{v}^{(3)} \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \vec{u}^{(n)} \cdot \hat{v}^{(n)} \end{bmatrix}$$

i.e.  $R_{ij} = \begin{cases} \vec{u}^{(j)} \cdot \hat{v}^{(i)} & i \leq j \\ 0 & i > j \end{cases}$  for  $1 \leq i, j \leq n$

is an upper-triangular  $n \times n$  invertible matrix.

*Proof.* We will show that every column of  $A$  and  $QR$  coincides. The  $j$ -th column of  $A$  is simply the  $j$ -th vector in the starting basis,  $\vec{u}^{(j)}$ . Meanwhile, the

---

<sup>13</sup>We will only check  $\hat{v}^{(1)}$  and  $\hat{v}^{(3)}$  are orthogonal to each other and leave the remaining two pairs to the readers.  $\hat{v}^{(1)} \cdot \hat{v}^{(3)} = (\frac{1}{3}, \frac{2}{3}, \frac{2}{3})^T \cdot (\frac{2}{\sqrt{17}}, \frac{2}{\sqrt{17}}, -\frac{3}{\sqrt{17}})^T = (\frac{1}{3})(\frac{2}{\sqrt{17}}) + (\frac{2}{3})(\frac{2}{\sqrt{17}}) + (\frac{2}{3})(-\frac{3}{\sqrt{17}}) = 0$ .

$j$ -th column of  $QR$  is  $Q$  times the  $j$ -th column of  $R$ , which is

$$\begin{aligned}
 QR^{(j)} &= [\hat{v}^{(1)} | \hat{v}^{(2)} | \dots | \hat{v}^{(j)} | \dots | \hat{v}^{(n)}] \begin{bmatrix} \vec{u}^{(j)} \cdot \hat{v}^{(1)} \\ \vec{u}^{(j)} \cdot \hat{v}^{(2)} \\ \vdots \\ \vec{u}^{(j)} \cdot \hat{v}^{(j)} \\ \vdots \\ 0 \end{bmatrix} \\
 &= (\vec{u}^{(j)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} + (\vec{u}^{(j)} \cdot \hat{v}^{(2)})\hat{v}^{(2)} + \dots + (\vec{u}^{(j)} \cdot \hat{v}^{(j)})\hat{v}^{(j)} + 0 \\
 &= \sum_{k=1}^j (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} = \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} + (\vec{u}^{(j)} \cdot \hat{v}^{(j)})\hat{v}^{(j)}
 \end{aligned}$$

By Definition 7.2.4, we have

$$\hat{v}^{(j)} = \frac{\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}}{\left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} \right\|}$$

<sup>14</sup> which after rearrangement, becomes

$$\vec{u}^{(j)} = \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} + \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} \right\| \hat{v}^{(j)}$$

Therefore, in order to show that  $\vec{u}^{(j)} = QR^{(j)}$ , by comparing the two expressions, we need to check if

$$\vec{u}^{(j)} \cdot \hat{v}^{(j)} = \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} \right\|$$

---

<sup>14</sup> Some may ask if  $\left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} \right\|$  can be 0 (or  $\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}$  be the zero vector) and  $\hat{v}^{(j)}$  is not well-defined. However, this will contradict the linear independence of the basis vectors  $\vec{u}^{(k)}$ , as each of  $\hat{v}^{(k)}$  are a linear combination of  $\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(k)}$ , deduced inductively from  $k = 1$  to  $k = j - 1$  by looking at Definition 7.2.4. Hence  $\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)} = \mathbf{0}$  implies a non-trivial solution to  $c_1\vec{u}^{(1)} + \dots + c_{j-1}\vec{u}^{(j-1)} + c_j\vec{u}^{(j)} = \mathbf{0}$  where particularly  $c_j = 1 \neq 0$ .

Consider

$$\begin{aligned} [\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)}] \cdot \hat{v}^{(j)} &= \vec{u}^{(j)} \cdot \hat{v}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) (\hat{v}^{(k)} \cdot \hat{v}^{(j)}) \\ &= \vec{u}^{(j)} \cdot \hat{v}^{(j)} \end{aligned}$$

as  $\hat{v}^{(k)} \cdot \hat{v}^{(j)} = 0$  for  $k \neq j$  due to the orthogonality enforced by the Gram-Schmidt process. On the other hand, by Definition 7.2.4 again,

$$\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} = \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} \right\| \hat{v}^{(j)}$$

Therefore,

$$\begin{aligned} [\vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)}] \cdot \hat{v}^{(j)} &= \left( \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} \right\| \hat{v}^{(j)} \right) \cdot \hat{v}^{(j)} \\ &= \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} \right\| (\hat{v}^{(j)} \cdot \hat{v}^{(j)}) \\ &= \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} \right\| \end{aligned}$$

as  $\hat{v}^{(j)} \cdot \hat{v}^{(j)} = \|\hat{v}^{(j)}\|^2 = 1$ . The required equality is then established and the result follows. The invertibility of  $R$  can be shown by noting that all diagonal elements  $\vec{u}^{(j)} \cdot \hat{v}^{(j)} = \left\| \vec{u}^{(j)} - \sum_{k=1}^{j-1} (\vec{u}^{(j)} \cdot \hat{v}^{(k)}) \hat{v}^{(k)} \right\|$  of the upper-triangular  $R$  matrix are non-zero (see Footnote 14).  $\square$

**Example 7.2.5.** Construct a QR decomposition for the case in Example 7.2.4.

*Solution.* The matrix  $Q$  is simply consisted of the orthonormal basis vectors:

$$Q = \begin{bmatrix} \frac{1}{3} & \frac{10}{3\sqrt{17}} & \frac{2}{\sqrt{17}} \\ \frac{2}{3} & -\frac{7}{3\sqrt{17}} & \frac{2}{\sqrt{17}} \\ \frac{2}{3} & \frac{2}{3\sqrt{17}} & -\frac{3}{\sqrt{17}} \end{bmatrix}$$

And by Properties 7.2.5, the entries in  $R$  are

$$R = \begin{bmatrix} \vec{u}^{(1)} \cdot \hat{v}^{(1)} & \vec{u}^{(2)} \cdot \hat{v}^{(1)} & \vec{u}^{(3)} \cdot \hat{v}^{(1)} \\ 0 & \vec{u}^{(2)} \cdot \hat{v}^{(2)} & \vec{u}^{(3)} \cdot \hat{v}^{(2)} \\ 0 & 0 & \vec{u}^{(3)} \cdot \hat{v}^{(3)} \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -\frac{1}{3} & 1 \\ 0 & \frac{\sqrt{17}}{3} & \frac{13}{\sqrt{17}} \\ 0 & 0 & \frac{1}{\sqrt{17}} \end{bmatrix}$$

whose values can be readily inferred from the steps during the orthogonalization process itself in Example 7.2.4 (highlighted in red/blue). The readers are encouraged to compute the matrix product  $QR$  to see if the original matrix  $A$  is recovered.

We conclude this section with a small remark related to the concept of orthogonal complement discussed in the end of Section 6.3.2.

**Properties 7.2.6.** For an orthogonal(-normal) basis  $\mathcal{B} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots, \vec{v}^{(n)}\}$  for a finite,  $n$ -dimensional vector space  $\mathcal{V}$ , the subspaces  $\mathcal{V}_G$  and  $\mathcal{V}_H$  formed by  $\mathcal{G} = \{\vec{v}^{(I)}\}$  and  $\mathcal{H} = \{\vec{v}^{(J)}\}$  respectively, where  $I$  and  $J$  are mutually exclusive indices that together count all integers from 1 to  $n$ , are the orthogonal complement to each other with respect to  $\mathcal{V}$ , such that  $\mathcal{V}_G^\perp = \mathcal{V}_H$  ( $\mathcal{V}_H^\perp = \mathcal{V}_G$ ) and  $\mathcal{V}_G \oplus \mathcal{V}_H = \mathcal{V}$ .

□

## 7.3 Python Programming

We can define a function to a change in coordinates for vectors or matrices. Let's first write a helper function to produce the change of coordinates matrix  $P$  proposed in Theorem 7.1.12, which equals to  $B'^{-1}B$  as discussed in the end of Example 7.1.6:

```

import numpy as np
from scipy import linalg

def P_matrix(B, B_prime):
    """ Computes the P matrix of change in coordinates. """
    P = linalg.inv(B_prime) @ B
    return(P)

```

Then we use Example 7.2.1 as an illustration for coordinate change for vectors, where regarding  $\mathcal{B}$  we have

$$B = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}$$

and  $B' = I$  as  $\mathcal{B}' = S$  implicitly in this case. We define another function for transforming the coordinates of any given vector as

```

def coord_trans_vector(vec, P):
    """ Transforms the coordinates of a vector. """
    trans_vec = linalg.inv(P) @ vec
    return(trans_vec)

```

Then Example 7.2.1 can be proceeded as follows.

```

B = np.array([[1., 1.],
              [2., -1.]])

P = P_matrix(B, np.identity(2))
old_v = np.array([2., 1.])
new_v = coord_trans_vector(old_v, P)
print(new_v)

```

which returns  $[1. 1.]$  correctly. Similarly, according to Properties 7.2.2, we can make a function to carry out the change of coordinates for matrices through

```

def coord_trans_matrix(A, P):
    """ Transforms the coordinates of a matrix. """
    trans_matrix = linalg.inv(P) @ A @ P
    return(trans_matrix)

```

Let's use this to redo Example 7.2.3, where

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

Subsequently,

```
B = np.array([[2., -1.],
             [-1., 1.]])
old_A = np.array([[1., 1.],
                  [0., 1.]])
```

```
P = P_matrix(B, np.identity(2))
new_A = coord_trans_matrix(old_A, P)
print(new_A)
```

gives

```
[[ 0.  1.]
 [-1.  2.]]
```

as expected. Meanwhile, to apply Gram-Schmidt Orthogonalization for a basis, in addition to deriving the corresponding QR decomposition, we can use the function `qr` in `scipy.linalg`. Let's we use Examples 7.2.4 and 7.2.5 as a demonstration:

```
A = np.array([[1., 1., 3.],
             [2., -1., -1.],
             [2., 0., 1.],])
Q, R = linalg.qr(A)
print("Q = ", Q)
print("R = ", R)
```

which yields

```
Q = [[-0.33333333  0.80845208 -0.48507125]
      [-0.66666667 -0.56591646 -0.48507125]
      [-0.66666667  0.16169042  0.72760688]]
R = [[-3.          0.33333333 -1.          ]
      [ 0.          1.37436854  3.15296313]
      [ 0.          0.           -0.24253563]]
```

The columns in Q form the desired orthonormal basis. Notice that the signs of the first/third column vectors in Q are flipped when compared to that in Example 7.2.5, which leads to corresponding sign switches in R as well.

## 7.4 Exercises

**Exercise 7.1** Let  $\mathcal{V} = \mathcal{W} = \mathbb{R}^3$ , and take  $\mathcal{B} = \{(1, 1, 1)^T, (1, 1, 0)^T, (1, 0, 0)^T\}$  and  $\mathcal{H} = \{(1, 2, 3)^T, (1, -1, 0)^T, (2, -1, -1)^T\}$  as bases for  $\mathcal{V}$  and  $\mathcal{W}$ . If a linear transformation  $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  is defined by  $T(x, y, z)^T = (x + y + z, 2x + y, x - y - 3z)^T$ , find its matrix representation and decide if it is one-to-one, onto and hence bijective.

**Exercise 7.2** Let  $\mathcal{V}$  be the real vector space generated by the basis  $\mathcal{B} = \{\cos x, \sin x, x \cos x, x \sin x\}$  and  $T : \mathcal{V} \rightarrow \mathcal{V}, T[f(x)] = f'(x)$  be the differentiation operator over  $\mathcal{V}$ . Find the matrix representation of  $T$  with respect to  $\mathcal{B}$ , and determine if  $T$  is injective, surjective and hence bijective. If it is bijective, find the inverse (matrix) of  $T$ .

**Exercise 7.3** Let  $\mathcal{V} = \mathcal{P}^2, \mathcal{W} = \mathcal{P}^3$  be the polynomial spaces of degree 2 and 3 respectively. Define  $T : \mathcal{V} \rightarrow \mathcal{W}$  by

$$T[p(x)] = \int_1^x p(x)dx$$

find its matrix representation with respect to the standard bases and decide if the transformation is isomorphic.

**Exercise 7.4** If  $\mathcal{U}$  and  $\mathcal{V}$  are isomorphic to an  $m/n$ -dimensional (real) vector space, show that the direct sum  $\mathcal{U} \oplus \mathcal{V}$  is isomorphic to an  $m + n$ -dimensional (real) vector space.

**Exercise 7.5** Show that every identity transformation  $T : \mathcal{V} \rightarrow \mathcal{V}, T(\vec{v}) = \text{id}(\vec{v}) = \vec{v}$  for a finite-dimensional vector space  $\mathcal{V}$  with respect to a fixed basis  $\mathcal{B}$  throughout always has a matrix representation of an identity matrix such that  $[T]_B = I$ .

**Exercise 7.6** For a linear operator  $T : \mathcal{P}^2 \rightarrow \mathcal{P}^2$ ,  $T(p(x)) = p(x - 1)$ , find its matrix representation with respect to the standard basis first, and then use the change of coordinates formula in Properties 7.2.2 where we do a variable substitution of  $x' = x + 2$  to compute the new matrix representation.

**Exercise 7.7** For two linear transformations  $T : \mathcal{U} \rightarrow \mathcal{V}$  and  $S : \mathcal{V} \rightarrow \mathcal{W}$ , where  $\mathcal{U}, \mathcal{V}, \mathcal{W}$  are finite-dimensional with bases  $\mathcal{B}, \mathcal{H}, \mathcal{K}$ , find the formula for change of coordinates to new bases  $\mathcal{B}', \mathcal{H}', \mathcal{K}'$  for the matrix representation of the composition  $S \circ T$ .

**Exercise 7.8** Apply Gram-Schmidt Orthogonalization on the following set of vectors, and then write down their QR Decomposition.

- (a)  $\vec{u}_1 = (1, 2)^T, \vec{u}_2 = (3, 8)^T,$
- (b)  $\vec{u}_1 = (1, 2, 1)^T, \vec{u}_2 = (1, 4, 4)^T, \vec{u}_3 = (2, 2, 5)^T$ , and
- (c)  $\vec{u}_1 = (1, -2, 2, 1)^T, \vec{u}_2 = (1, 1, 0, 2)^T, \vec{u}_3 = (2, 3, -1, 0)^T.$

## Chapter 8

# Complex Vectors/Matrices and Block Matrices

---

In this chapter, we will take a detour to talk about two auxiliary topics. The first one is the generalization of vectors and matrices to having complex numbers as entries. Eventually, we will mention about the *complex vector space*, and compare it to the real vector space that we just learnt in the previous chapters. The second one is about *block form* of a matrix (or simply referred to as a *block matrix*) that is composed of smaller *submatrices* as the building blocks. Writing a matrix in block form enables efficient manipulation for many situations that we will encounter in the remaining parts of this book.

## 8.1 Definition and Operations of Complex Numbers

### 8.1.1 Basic Structure of Complex Numbers

The idea of complex numbers initially came from some algebra problems that lead to the square root of negative quantities, which was undefined back in the days. Later, mathematicians addressed this issue by introducing the **imaginary number**  $i = \sqrt{-1}$ , and  $i^2 = -1$ . For any positive number  $b$ , we have  $\sqrt{-b^2} = \sqrt{b^2}\sqrt{-1} = bi$ . **Complex numbers** are then quantities in the form of  $a + bi$ , where  $a$  and  $b$  themselves are real. Here  $a$  and  $b$  are called the **real** and

**imaginary part** respectively. As a small example of how complex numbers arise, note that the solutions to the quadratic equation  $(3x + 2)^2 = -1$ , are  $-2 \pm \frac{1}{3}i$ .

**Definition 8.1.1.** Complex numbers are scalars in the form of  $z = a + bi$ , where  $a$  and  $b$  are some real numbers. Their real and imaginary part are denoted by  $\text{Re}\{z\} = a$  and  $\text{Im}\{z\} = b$ .

We also need to consider when two complex numbers are equal. This happens when their real parts, as well as imaginary parts, are equal to each other respectively.

**Properties 8.1.2.** Two complex numbers  $z_1 = a + bi$ , and  $z_2 = c + di$ , where  $a, b, c, d$  are real numbers, are equal if and only if,  $\text{Re}\{z_1\} = a = c = \text{Re}\{z_2\}$  and  $\text{Im}\{z_1\} = b = d = \text{Im}\{z_2\}$ .

For every complex number, there exists another corresponding complex number known as the **(complex) conjugate** associated to it, formed by flipping the sign of its imaginary part.

**Definition 8.1.3.** For a complex number  $z = a + bi$ , its complex conjugate is defined as  $\bar{z} = a - bi$ .

For example, the conjugate of  $2 - 5i$  is  $2 + 5i$ .

## 8.1.2 Complex Number Operations

Below are some rules about usual operations on two complex numbers.

### Addition and Subtraction

Addition and subtraction between two complex numbers is carried out over the real parts and the imaginary parts separately.

**Definition 8.1.4.** For two complex numbers  $z_1 = a + bi$ , and  $z_2 = c + di$ , we have  $z_1 \pm z_2 = (a + bi) \pm (c + di) = (a \pm c) + (b \pm d)i = (\operatorname{Re}\{z_1\} \pm \operatorname{Re}\{z_2\}) + (\operatorname{Im}\{z_1\} \pm \operatorname{Im}\{z_2\})i$ .

For instance, adding  $1 + 3i$  to  $2 - 4i$  results in  $(1 + 2) + (3 - 4)i = 3 - i$ .

### Multiplication and Division

Multiplication of two complex numbers simply works like the usual distributive law where we pretend that  $i$  is a variable.

**Definition 8.1.5.** Given two complex numbers  $a + bi$ , and  $c + di$ , their product is

$$\begin{aligned}(a + bi)(c + di) &= a(c + di) + bi(c + di) \\ &= ac + adi + bci + bdi^2 \\ &= (ac - bd) + (ad + bc)i \quad (i^2 = -1)\end{aligned}$$

**Example 8.1.1.** Evaluate  $(1 + 2i)(3 - 4i)$ .

*Solution.*

$$\begin{aligned}(1 + 2i)(3 - 4i) &= ((1)(3) - (2)(-4)) + ((1)(-4) + (2)(3))i \\ &= 11 + 2i\end{aligned}$$

□

Dividing something by a complex number  $a + bi$  can be viewed as multiplication by its complex conjugate  $a - bi$ , as

$$\begin{aligned}\frac{1}{a + bi} &= \frac{1}{a + bi} \frac{a - bi}{a - bi} \\ &= \frac{a - bi}{a^2 - (-b^2) - abi + bai}\end{aligned}$$

$$= \frac{a - bi}{a^2 + b^2}$$

with an additional factor of  $\frac{1}{a^2+b^2}$ . It is interesting that this  $a^2 + b^2$  term coming from multiplying the complex number by its conjugate over the denominators looks like the square of hypotenuse as in the *Pythagoras' Theorem*. Later on we will see more when we discuss the geometric meaning of complex numbers.

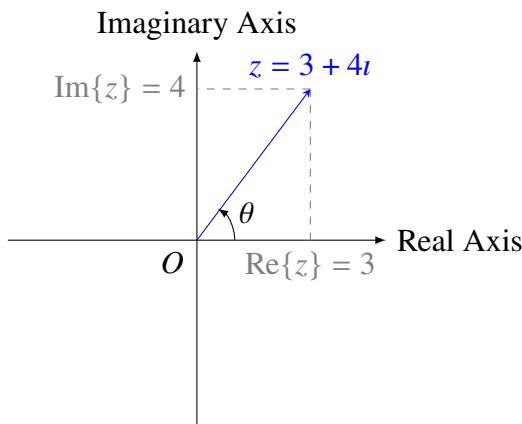
**Example 8.1.2.** Compute  $\frac{1+4i}{2+3i}$ .

*Solution.* Following the idea outlined above, we have

$$\begin{aligned} \frac{1+4i}{2+3i} &= \frac{1+4i}{2+3i} \frac{2-3i}{2-3i} \\ &= \frac{(1+4i)(2-3i)}{2^2 + 3^2} \\ &= \frac{((1)(2) - (4)(-3)) + ((1)(-3) + (4)(2))i}{13} = \frac{14}{13} + \frac{5}{13}i \end{aligned}$$

□

### 8.1.3 Geometric Meaning of Complex Numbers



A complex number  $z = 3 + 4i$  represented in the complex plane.

## 8.1 Definition and Operations of Complex Numbers

A complex number can be visualized as a two-dimensional vector, in the so-called **complex plane** (or sometimes referred to as the **Argand plane**), where the *x*-axis represents the real part and the *y*-axis represents the imaginary part. These two axes are referred to as the **real axis** and **imaginary axis** respectively.

It is obvious that the length of such vector is  $|z| = \sqrt{\operatorname{Re}\{z\}^2 + \operatorname{Im}\{z\}^2}$ , which is called the **modulus** of the corresponding complex number. In the diagram above, the modulus of  $z$  is easily seen to be  $|z| = 5$ .

The angle between the real axis and the complex number is called the **argument**, shown as  $\theta = \arctan(\operatorname{Im}\{z\}/\operatorname{Re}\{z\})$  in the same figure. Since its complex conjugate  $\bar{z}$  has the sign of the imaginary part flipped while the real part remains the same, the argument of the complex conjugate is simply the negative of that of the original complex number  $z$ . Also, the modulus will be unchanged.

Moreover, from elementary trigonometry, we know that  $\operatorname{Re}\{z\} = |z| \cos \theta$  and  $\operatorname{Im}\{z\} = |z| \sin \theta$ . Hence  $z$  can be represented as  $z = \operatorname{Re}\{z\} + i \operatorname{Im}\{z\} = |z|(\cos \theta + i \sin \theta)$ . We also have the famous **Euler's Formula**, relating the geometry of any complex number with an exponential raised to an imaginary power.

**Definition 8.1.6** (Euler's Formula). An exponential raised to an imaginary power is a complex number such that

$$e^{i\theta} = \cos \theta + i \sin \theta$$

where  $\theta$  is taken to be real.

Hence  $z$  can be further written as  $z = |z|e^{i\theta}$ , and  $\bar{z} = \operatorname{Re}\{z\} - i \operatorname{Im}\{z\} = |z|(\cos \theta - i \sin \theta) = |z|(\cos(-\theta) + i \sin(-\theta)) = |z|e^{-i\theta}$ . Conversely, the quantity  $e^{i\theta}$  can be regarded as a complex number that has a modulus of 1 and an argument of  $\theta$ . Additionally, this provides formulae to express sines and cosines with complex exponentials.

**Properties 8.1.7.** For any  $\theta$  which is confined to be real,

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2}$$

$$\sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$

*Proof.* By Definition 8.1.6,

$$\begin{aligned}\frac{e^{i\theta} + e^{-i\theta}}{2} &= \frac{1}{2}((\cos \theta + i \sin \theta) + (\cos(-\theta) + i \sin(-\theta))) \\ &= \frac{1}{2}((\cos \theta + i \sin \theta) + (\cos \theta - i \sin \theta)) \\ &= \cos \theta\end{aligned}$$

The derivation for  $\sin \theta$  is left as an exercise.  $\square$

Now we can go back to investigate complex multiplication and division. Multiplication of a complex number  $z_1$  by another complex number  $z_2$ , can be viewed as  $z_1 z_2 = (|z_1|e^{i\theta_1})(|z_2|e^{i\theta_2}) = |z_1||z_2|e^{i(\theta_1+\theta_2)}$ .<sup>1</sup> This can be interpreted as, starting with the complex number  $z_1 = |z_1|e^{i\theta_1}$  on the complex plane, rotating it anti-clockwise (i.e. in the positive direction) by an angle of  $\theta_2$ , and scaling its modulus by a factor of  $|z_2|$ .

Similarly, division of  $z_1$  by  $z_2$ , is  $z_1/z_2 = (|z_1|/|z_2|)e^{i(\theta_1-\theta_2)}$ . Notice that for a fraction like  $1/z = 1/(a+bi)$ , it can be rewritten as

$$\begin{aligned}\frac{1}{z} &= \frac{1}{|z|e^{i\theta}} = \frac{1}{|z|} \frac{e^{-i\theta}}{e^{i\theta}e^{-i\theta}} = \frac{1}{|z|} \frac{e^{-i\theta}}{e^{i(\theta-\theta)}} = \frac{1}{|z|} \frac{e^{-i\theta}}{e^0} = \frac{1}{|z|} e^{-i\theta} \\ &= \frac{1}{|z|^2} (|z|e^{-i\theta}) \\ &= \frac{1}{|z|^2} \bar{z}\end{aligned}$$

---

<sup>1</sup>We take it for granted that  $e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$ .

## 8.1 Definition and Operations of Complex Numbers

which is consistent with the discussion about complex division in the last section, where  $|z|^2 = a^2 + b^2$  arises in the denominator. In addition, we can observe that  $|z|^2 = z\bar{z}$ . This is not coincidence, as

$$\begin{aligned} z\bar{z} &= |z|e^{i\theta}|z|e^{-i\theta} \\ &= |z|^2 e^{i(\theta-\theta)} = |z|^2 e^0 = |z|^2 \end{aligned} \quad (8.1)$$

Geometrically, we can think of it as starting with 1 along the real axis in the complex plane, then we scale it by  $|z|$  and rotate it by  $\theta$ , and finally scale it again by  $|z|$  but rotate it by  $-\theta$ , the same angle but in opposite direction. The results will be a real number  $|z|^2$  (the length of  $z$  squared), since the two opposite rotations cancel out each other.

Below are some properties of modulus and complex conjugate to be remembered.

**Properties 8.1.8.** For two complex numbers  $z_1$  and  $z_2$ , we have

- (a)  $\overline{z_1 \pm z_2} = \overline{z_1} \pm \overline{z_2}$ ,
- (b)  $\overline{z_1 z_2} = \overline{z_1} \overline{z_2}$ ,
- (c)  $\overline{z_1/z_2} = \overline{z_1}/\overline{z_2}$ ,
- (d)  $\overline{\overline{z}} = z$ ,
- (e)  $\overline{\overline{z_1} z_2} = z_1 \overline{z_2}$ ,
- (f)  $|\overline{z}| = |z|$ ,
- (g)  $|z_1 z_2| = |z_1||z_2|$ ,
- (h)  $|z_1/z_2| = |z_1|/|z_2|$ .

Another very useful result is the ***De Moivre's Formula*** that builds up on the Euler's formula, expressing  $e^{i\theta}$  raised to an integer power  $n$ .

**Theorem 8.1.9** (De Moivre's Formula). Given  $n$  as an integer, then

$$(e^{i\theta})^n = e^{i(n\theta)}$$

$$(\cos \theta + i \sin \theta)^n = \cos(n\theta) + i \sin(n\theta)$$

## 8.2 Complex Vectors and Complex Matrices

Our discussion about vectors and matrices in previous chapters is limited to those with real entries. However, we can extend the ideas to include complex elements. A complex vector is simply a vector that have complex number as components. An  $n$ -dimensional complex vector can be somehow viewed as a real vector that is  $2n$ -dimensional, as each complex entry can be expressed in two parts, real and imaginary. This equivalence will be further clarified in the end of this section. A complex matrix is similarly a matrix with complex elements, or from another perspective, formed by complex column vectors.

### 8.2.1 Operations and Properties of Complex Vectors

Addition and Subtraction for complex vectors are the same as the real counterpart, carried out element-wise. Multiplication by a scalar is also similar, applied to all elements. However, the form of complex dot product is slightly different from the real dot product, as defined below.

**Definition 8.2.1** (Complex Dot Product). The dot product of two complex vectors  $\vec{u}$  and  $\vec{v}$  is computed as the sum of products between each pair of elements, but additionally with the conjugate operation applied on the second complex vector beforehand.

$$\begin{aligned} \vec{u} \cdot \vec{v} &= \mathbf{u}^T \bar{\mathbf{v}} \\ &= u_1 \bar{v}_1 + u_2 \bar{v}_2 + \cdots + u_n \bar{v}_n = \sum_{k=1}^n u_k \bar{v}_k \end{aligned}$$

The bar on  $\bar{\mathbf{v}}$  means carrying out conjugate on every entry of  $\mathbf{v}$ . If  $\mathbf{v} = \operatorname{Re}\{\mathbf{v}\} + i \operatorname{Im}\{\mathbf{v}\}$ , where  $\operatorname{Re}\{\mathbf{v}\}$  and  $\operatorname{Im}\{\mathbf{v}\}$  are the vectors consisted of the real/imaginary part of every element in  $\mathbf{v}$ , then  $\bar{\mathbf{v}} = \operatorname{Re}\{\mathbf{v}\} - i \operatorname{Im}\{\mathbf{v}\}$ .

The Euclidean norm, or length of a complex vector, is defined in a similar fashion.

**Definition 8.2.2.** The length  $\|\vec{v}\|$  of a complex vector  $\vec{v}$  is calculated by

$$\begin{aligned}\|\vec{v}\| &= \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\mathbf{v}^T \bar{\mathbf{v}}} \\ &= \sqrt{v_1 \bar{v}_1 + v_2 \bar{v}_2 + \cdots + v_n \bar{v}_n} \\ &= \sqrt{|v_1|^2 + |v_2|^2 + \cdots + |v_n|^2} \quad (\text{Equation 8.1}) \\ &= \sqrt{\sum_{k=1}^n |v_k|^2}\end{aligned}$$

Properties of complex dot product hence also vary slightly from its real counterpart, Properties 4.2.3.

**Properties 8.2.3.** For two complex vectors  $\vec{u}$  and  $\vec{v}$ , we have

$$\begin{array}{ll}\vec{u} \cdot \vec{v} = \overline{\vec{v} \cdot \vec{u}} & \text{Anti-symmetric Property} \\ \vec{u} \cdot (\vec{v} \pm \vec{w}) = \vec{u} \cdot \vec{v} \pm \vec{u} \cdot \vec{w} & \text{Distributive Property} \\ (\vec{u} \pm \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} \pm \vec{v} \cdot \vec{w} & \text{Distributive Property} \\ (a\vec{u}) \cdot (b\vec{v}) = a\bar{b}(\vec{u} \cdot \vec{v}) & \text{where } a, b \text{ are some complex constants}\end{array}$$

There is no cross product analogous for complex vectors.

**Example 8.2.1.** Show the anti-symmetric property in Properties 8.2.3 holds for  $\vec{u} = (1 + 2i, 3 + i)^T$ ,  $\vec{v} = (2 - 5i, 1 + 4i)^T$ .

*Solution.*

$$\vec{u} \cdot \vec{v} = (1 + 2i)(\overline{2 - 5i}) + (3 + i)(\overline{1 + 4i})$$

$$\begin{aligned}
 &= (1 + 2i)(2 + 5i) + (3 + i)(1 - 4i) \\
 &= (-8 + 9i) + (7 - 11i) \\
 &= -1 - 2i
 \end{aligned}$$

$$\begin{aligned}
 \vec{v} \cdot \vec{u} &= (2 - 5i)(\overline{1 + 2i}) + (1 + 4i)(\overline{3 - i}) \\
 &= (2 - 5i)(1 - 2i) + (1 + 4i)(3 - i) \\
 &= (-8 - 9i) + (7 + 11i) \\
 &= -1 + 2i
 \end{aligned}$$

Hence  $\vec{u} \cdot \vec{v} = \overline{\vec{v} \cdot \vec{u}}$ . □

Short Exercise: Find the norm  $\|\vec{u}\|$  and  $\|\vec{v}\|$  respectively.<sup>2</sup>

### 8.2.2 Operations and Properties of Complex Matrices

Matrix multiplication between two complex matrices is carried out in the same way as we have been always doing, according to Definition 1.1.1. However, due to the difference in definition of dot product for real and complex vectors, we can no longer claim like in the discussion below Definition 4.2.1 that the elements resulted from a complex matrix product are complex vector dot products between the corresponding rows and columns. To make the statement work again, a minor modification is needed as we will see soon.

#### Conjugate Transpose

Transpose can be similarly defined for complex matrices. However, there exists a more useful operation that combines transpose and conjugate.

---

<sup>2</sup>  $\|\vec{u}\| = \sqrt{(1 + 2i)(1 - 2i) + (3 + i)(3 - i)} = \sqrt{(1^2 + 2^2) + (3^2 + 1^2)} = \sqrt{15}$ . Similarly,  $\|\vec{v}\| = \sqrt{46}$ .

**Definition 8.2.4** (Conjugate Transpose). The *conjugate transpose* of a matrix  $A$ , denoted as  $A^* = \overline{A^T}$ , has elements  $A_{pq}^* = \overline{A}_{qp}$ , where  $\overline{A}$  is the conjugate of the matrix  $A$  produced by changing every element in  $A$  to its complex conjugate. Sometimes  $A^*$  is called the *adjoint* or *Hermitian transpose* of  $A$ , and alternatively denoted as  $A^H$ .

It means that conjugate transpose is done by conjugating all elements of the matrix and flipping them about its main diagonal. A *Hermitian matrix* is a complex matrix whose conjugate transpose equals to itself.

**Definition 8.2.5.** A complex square matrix  $A$  is called Hermitian if  $A^* = A$ .

Properties of conjugate transpose are alike to those for real transpose stated in Properties 2.1.4. Related, for complex dot product, there are also something parallel to the second half of Properties 4.2.3.

**Properties 8.2.6.** For two complex matrices  $A$  and  $B$ , we have

1.  $(cA)^* = \overline{c}A^*$ , where  $c$  is any complex scalar,
2.  $(A^*)^* = A$ ,
3.  $(A \pm B)^* = A^* \pm B^*$ , if  $A$  and  $B$  have the same shape,
4.  $(AB)^* = B^*A^*$ , if  $A$  and  $B$  have compatible shapes.

**Properties 8.2.7.** For two complex vectors  $\vec{u}$  and  $\vec{v}$  and a complex matrix  $A$ , we have

$$\begin{aligned}\vec{u} \cdot (A\vec{v}) &= \mathbf{u}^T \overline{(A\vec{v})} = (\overline{A^T \mathbf{u}})^T \overline{\vec{v}} = (A^* \vec{u}) \cdot \vec{v} \\ (A\vec{u}) \cdot \vec{v} &= (A\mathbf{u})^T \overline{\vec{v}} = \mathbf{u}^T (A^T \overline{\vec{v}}) = \vec{u} \cdot (A^* \vec{v})\end{aligned}$$

where Properties 2.1.4 and Definitions 8.2.1, 8.2.4 are used.

With complex conjugate of a matrix defined in passing, we can now say that the complex vector dot products between each of the row and column vectors

in a matrix  $A$  and another matrix  $B$  respectively, are encoded in the elements of the complex matrix product  $A\bar{B}$  where a conjugate is applied on the second matrix.

**Example 8.2.2.** For two complex matrices

$$A = \begin{bmatrix} 1 & i \\ -i & 0 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 2+3i \\ 1+i & 1-i \end{bmatrix}$$

Verify that  $(AB)^* = B^*A^*$ .

*Solution.*

$$\begin{aligned} A^* &= \begin{bmatrix} 1 & i \\ -i & 0 \end{bmatrix} \\ B^* &= \begin{bmatrix} 1 & 1-i \\ 2-3i & 1+i \end{bmatrix} \\ B^*A^* &= \begin{bmatrix} 1 & 1-i \\ 2-3i & 1+i \end{bmatrix} \begin{bmatrix} 1 & i \\ -i & 0 \end{bmatrix} \\ &= \begin{bmatrix} (1)(1) + (1-i)(-i) & (1)(i) + (1-i)(0) \\ (2-3i)(1) + (1+i)(-i) & (2-3i)(i) + (1+i)(0) \end{bmatrix} = \begin{bmatrix} -i & i \\ 3-4i & 3+2i \end{bmatrix} \end{aligned}$$

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

□

## Determinants and Inverses for complex matrices

Complex matrices also have determinants and inverses, and are calculated in the exact same ways outlined in Sections 2.3 and 2.2. We provide a few examples here.

**Example 8.2.3.** Calculate the determinant for

$$A = \begin{bmatrix} 1 - i & 3 & 2 \\ 1 + i & 0 & i \\ 2 & -2i & 1 \end{bmatrix}$$

*Solution.* We apply Cofactor Expansion along the middle row in the way outlined in Properties 2.3.3, the result is

$$\begin{aligned} \det(A) &= -(1+i) \begin{vmatrix} 3 & 2 \\ -2i & 1 \end{vmatrix} + (0) \begin{vmatrix} 1-i & 2 \\ 2 & 1 \end{vmatrix} - (i) \begin{vmatrix} 1-i & 3 \\ 2 & -2i \end{vmatrix} \\ &= -(1+i)(3+4i) - (i)(-8-2i) \\ &= -1+i \end{aligned}$$

□

**Example 8.2.4.** Find the inverse of the matrix  $A$  in the last example.

*Solution.* The computation of inverse follows Properties 2.3.11. First, we note that

$$\begin{aligned} \frac{1}{\det(A)} &= \frac{1}{-1+i} \\ &= \frac{1}{-1+i} \frac{-1-i}{-1-i} \\ &= \frac{-1-i}{1+1} = -\frac{1+i}{2} \end{aligned}$$

Then, we proceed to compute the cofactor matrix for  $A$ , which is

$$C = \begin{bmatrix} \left| \begin{array}{cc} 0 & \iota \\ -2\iota & 1 \end{array} \right| & -\left| \begin{array}{cc} 1+\iota & \iota \\ 2 & 1 \end{array} \right| & \left| \begin{array}{cc} 1+\iota & 0 \\ 2 & -2\iota \end{array} \right| \\ -\left| \begin{array}{cc} 3 & 2 \\ -2\iota & 1 \end{array} \right| & \left| \begin{array}{cc} 1-\iota & 2 \\ 2 & 1 \end{array} \right| & -\left| \begin{array}{cc} 1-\iota & 3 \\ 2 & -2\iota \end{array} \right| \\ \left| \begin{array}{cc} 3 & 2 \\ 0 & \iota \end{array} \right| & -\left| \begin{array}{cc} 1-\iota & 2 \\ 1+\iota & \iota \end{array} \right| & \left| \begin{array}{cc} 1-\iota & 3 \\ 1+\iota & 0 \end{array} \right| \end{bmatrix}$$

$$= \begin{bmatrix} -2 & -1+\iota & 2-2\iota \\ -3-4\iota & -3-\iota & 8+2\iota \\ 3\iota & 1+\iota & -3-3\iota \end{bmatrix}$$

Thus, by Properties 2.3.11, the inverse of  $A$  is

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A) = \frac{1}{\det(A)} C^T$$

$$= -\frac{1+\iota}{2} \begin{bmatrix} -2 & -3-4\iota & 3\iota \\ -1+\iota & -3-\iota & 1+\iota \\ 2-2\iota & 8+2\iota & -3-3\iota \end{bmatrix}$$

$$= \begin{bmatrix} 1+\iota & -\frac{1}{2} + \frac{7}{2}\iota & \frac{3}{2} - \frac{3}{2}\iota \\ 1 & 1+2\iota & -\iota \\ -2 & -3-5\iota & 3\iota \end{bmatrix}$$

□

Short Exercise: Find  $A^{-1}$  via Gaussian Elimination.<sup>3</sup>

Below are some useful properties of determinant and inverse for complex matrices that can be compared to Properties 2.3.9 and 2.2.3.

---

<sup>3</sup>Notice that we will now need to multiply rows with complex constants instead when doing elementary row operations. You should be able to get the same answer. Possible first few steps are multiplying the first row by  $\frac{1+\iota}{2}$  to create a leading 1 and subtract  $1+\iota$  and 2 times the new first row from the second and third row respectively.

**Properties 8.2.8.** If  $A$  is a complex matrix, then

1.  $\det(A^T) = \det(A)$ ,
2.  $\det(A^*) = \overline{\det(A)}$ ,
3.  $\det(kA) = k^n \det(A)$ , for any complex constant  $k$ ,
4.  $\det(AB) = \det(A)\det(B)$ , and
5.  $\det(A^{-1}) = \frac{1}{\det(A)}$ , given  $A$  is invertible ( $\det(A) \neq 0$  as the same as before).

Additionally, if  $A$  is non-singular, then

1.  $(cA)^{-1} = \frac{1}{c}A^{-1}$ , for any complex scalar  $c \neq 0$ ,
2.  $(A^{-1})^{-1} = A$ ,
3.  $(A^n)^{-1} = (A^{-1})^n$ , for any positive integer  $n$ ,
4.  $(AB)^{-1} = B^{-1}A^{-1}$ , provided that  $B$  is invertible too, plus  $A$  and  $B$  are conformable,
5.  $(A^T)^{-1} = (A^{-1})^T$ ,
6.  $(A^*)^{-1} = (A^{-1})^*$ .

### 8.2.3 The Complex $n$ -space $\mathbb{C}^n$

Similar to the real  $n$ -space  $\mathbb{R}^n$  brought up in Definition 4.1.2, the set of all complex vectors, now with  $n$  complex components, forms the **complex  $n$ -space**  $\mathbb{C}^n$  as follows.

**Definition 8.2.9** (The Complex  $n$ -space  $\mathbb{C}^n$ ). The complex  $n$ -space  $\mathbb{C}^n$  is defined as the set of all possible  $n$ -tuples in the form of  $\vec{v} = (v_1, v_2, v_3, \dots, v_n)^T$ , where  $v_i$  can be any complex numbers, for  $i = 1, 2, 3, \dots, n$ . They are known as  $n$ -dimensional complex vectors.

A very interesting (and perhaps quite confusing) fact about the complex  $n$ -space  $\mathbb{C}^n$ , or an  $n$ -dimensional complex vector, is that it can be considered as  $2n$ -dimensional when put in the frame of a real vector space. The key lies in Definition 6.1.1 where if the underlying scalar is set to  $\mathbb{R}$  or  $\mathbb{C}$  so that it becomes a real/complex vector space. Notice the subtle difference between a real/complex vector (that is indicative of its components being real/complex) and real/complex vector space (concerning *the underlying scalar used in scalar multiplication*). We take  $\mathbb{C}$  as a vector space here for illustration. If  $\mathbb{C}$  is treated as a complex vector space, i.e. over  $\mathbb{C}$  itself, then  $\{1\}$  is a basis for  $\mathbb{C}$  since the scalar multiplication of 1 by any arbitrary *complex scalar* can generate all complex numbers as the scalar itself. Hence, the dimension of  $\mathbb{C}$  is 1 over  $\mathbb{C}$  (Properties 6.2.4 still holds for complex vector spaces). Otherwise, if  $\mathbb{C}$  is taken as a real vector space (over  $\mathbb{R}$ ), then  $\{1\}$  is not sufficient to be a basis for  $\mathbb{C}$  since multiplication of 1 by only any *real scalar*  $a$  can never produce complex numbers with a non-zero imaginary part. On the other hand,  $\{1, i\}$  can instead be a basis for  $\mathbb{C}$  over  $\mathbb{R}$  as linear combinations of 1 and  $i$  with purely real coefficients can produce all complex numbers. So by Properties 6.2.4, the dimension of  $\mathbb{C}$  over  $\mathbb{R}$  is 2, and with Theorem 7.1.13 it is isomorphic to  $\mathbb{R}^2$  in this situation. An explicit isomorphism between  $\mathbb{C}$  and  $\mathbb{R}^2$  over  $\mathbb{R}$  is simply

$$T(a + bi) = (a, b)^T$$

Extending this observation,  $\mathbb{C}^n$  can either be treated as  $n$ -dimensional over  $\mathbb{C}$  or  $2n$ -dimensional over  $\mathbb{R}$  (and is isomorphic to  $\mathbb{R}^{2n}$ ). However, unless mentioned otherwise, we consider any  $\mathbb{C}^n$  vector (or complex matrix) is taken over  $\mathbb{C}$  (the former case) onwards. All results from the last two chapters are still valid if we replace  $\mathbb{R}$  and  $\mathbb{R}^n$  by  $\mathbb{C}$  and  $\mathbb{C}^n$ .<sup>4</sup>

**Properties 8.2.10.** (Angle between complex vectors, WIP)

---

<sup>4</sup>For example, a linear combination of complex vectors (Definition 6.1.3)  $\vec{v}^{(j)} \in \mathbb{C}^n$  is still in the form of  $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \cdots + c_q\vec{v}^{(q)}$  but the coefficients  $c_j \in \mathbb{C}$  are complex numbers now.

## 8.3 Manipulating Block Matrices

Moving to our second topic, a ***block matrix*** is a matrix written in smaller *submatrices* as if they are ordinary entries. For example, a  $2 \times 2$  block matrix has the form of

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where  $A, B, C, D$  are themselves matrices having the shapes of  $m \times p, m \times q, n \times p, n \times q$ , and  $m, n, p, q$  can be any positive integer. As a more concrete example, we have

$$M = \left[ \begin{array}{ccc|cc} 1 & 0 & 3 & 0 & 4 \\ 0 & 1 & 1 & 2 & -1 \\ \hline 2 & -1 & 0 & 1 & -2 \end{array} \right]$$

being a  $3 \times 5$  matrix at the same time a  $2 \times 2$  block matrix where

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

are of the shapes  $2 \times 3, 2 \times 2, 1 \times 3$  and  $1 \times 2$ . We can extend this for block matrices of any partition. For instance, a  $3 \times 4$  block matrix will be in the form of

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \end{bmatrix}$$

where the  $M_{ij}$ s are submatrices, and for a fixed  $i$  ( $j$ ),  $M_{ij}$  has the same number of rows (columns).

### 8.3.1 Block Matrix Multiplication

With the structure of a block matrix explained, we can now examine how matrix multiplication between two block matrices is done. Let's take a look at the easiest case of two  $2 \times 2$  block matrices:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad N = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$$

Of course, from the very beginning (Section 1.1), we know that  $M$  and  $N$  themselves have to be of the shapes  $m \times r$  and  $r \times n$  as an ordinary matrix, but how about the submatrices? In fact, we just carry out the multiplication as if each of them are a single entry, such that

$$MN = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{bmatrix}$$

Then, for each resulting block to be valid, the number of columns in  $A$  and  $C$  ( $B$  and  $D$ ) must be the same as that of rows in  $X$  and  $Y$  ( $Z$  and  $W$ ). So that  $A$  and  $C$  will have the shapes of  $m_1 \times r_1$ ,  $m_2 \times r_1$ ,  $X$  and  $Y$  will have the shapes of  $r_1 \times n_1$ ,  $r_1 \times n_2$ ,  $m_1 + m_2 = m$ ,  $n_1 + n_2 = n$ . Similarly,  $B$  and  $D$  need to have the shapes of  $m_1 \times r_2$ ,  $m_2 \times r_2$ ,  $Z$  and  $W$  need to have the shapes of  $r_2 \times n_1$ ,  $r_2 \times n_2$ , and  $r = r_1 + r_2$ . In short, the position of vertical cuts along the column direction of  $M$  must coincide with that of horizontal cuts along the row direction of  $N$ . Below is a walk-through example.

**Example 8.3.1.** Given

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad N = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$$

as a  $3 \times 3$  and  $3 \times 2$  matrix respectively, with

$$\begin{aligned} A &= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} & B &= \begin{bmatrix} 1 \\ -2 \end{bmatrix} \\ C &= \begin{bmatrix} 0 & -1 \end{bmatrix} & D &= [1] \end{aligned}$$

$$X = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

$$Z = \begin{bmatrix} -1 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$W = \begin{bmatrix} 1 \end{bmatrix}$$

such that the partitions look like

$$M = \left[ \begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 1 & -2 \\ \hline 0 & -1 & 1 \end{array} \right] \quad N = \left[ \begin{array}{c|c} 0 & 1 \\ 2 & -1 \\ \hline -1 & 1 \end{array} \right]$$

Use block matrix multiplication to compute  $MN$ .

*Solution.* Note that the cuts along the column/row direction in  $M$  and  $N$  are both located in-between the 2nd-3rd index. Consequentially, we can use the formula above:

$$MN = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{bmatrix}$$

which requires us to compute

$$\begin{aligned} AX &= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} & BZ &= \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \\ AY &= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} & BW &= \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \\ CX &= \begin{bmatrix} 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \end{bmatrix} & DZ &= \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} = \begin{bmatrix} -1 \end{bmatrix} \\ CY &= \begin{bmatrix} 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} & DW &= \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} \end{aligned}$$

Hence

$$MN = \begin{bmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{bmatrix}$$

$$= \left[ \begin{array}{c|c} \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 2 \end{bmatrix} & \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \end{bmatrix} \\ \hline \begin{bmatrix} -2 \\ -1 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{array} \right] = \left[ \begin{array}{c|c} 3 & 0 \\ \hline 4 & -3 \\ \hline -3 & 2 \end{array} \right]$$

The readers can check the answer by computing the matrix product in the usual way.  $\square$

For multiplication involving block matrices with more blocks, the two block matrices  $M$  and  $N$  must have a partition of  $m \times r$  and  $r \times n$  blocks and the block multiplication is carried out as if they are individual entries in usual matrix multiplication as well. Particularly, the positions where the  $r$  column (row) partition of  $M$  ( $N$ ) occurs must align exactly. Given

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & \cdots & M_{1r} \\ M_{21} & M_{22} & M_{23} & & M_{2r} \\ M_{31} & M_{32} & M_{33} & & M_{3r} \\ \vdots & & & \ddots & \vdots \\ M_{m1} & M_{m2} & M_{m3} & & M_{mr} \end{bmatrix} \quad \text{and } N = \begin{bmatrix} N_{11} & N_{12} & N_{13} & \cdots & N_{1n} \\ N_{21} & N_{22} & N_{23} & & N_{2n} \\ N_{31} & N_{32} & N_{33} & & N_{3n} \\ \vdots & & & \ddots & \vdots \\ N_{r1} & N_{r2} & N_{r3} & & N_{rn} \end{bmatrix}$$

this means that the numbers of columns and rows in  $M_{ik}$  and  $N_{kj}$  for any fixed  $k$  should be equal, such that  $M_{ik}$  and  $N_{kj}$  are of the shapes  $m_i \times r_k$  and  $r_k \times n_j$ .

### 8.3.2 Inverse and Determinant of a Block Matrix

To properly utilize block matrices, we also need to know how to compute some basic quantities related to them, like inverse and determinant. Since most of the situations involve  $2 \times 2$  block matrices only, we will handle them exclusively. Specifically, we consider  $2 \times 2$  block matrices in the form of

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where  $A, B, C, D$  are submatrices of the shapes  $p \times p$ ,  $p \times q$ ,  $q \times p$  and  $q \times q$ , such that  $A, D$  and thus  $M$  are square. To proceed, we need the following observations.

**Properties 8.3.1.** Denote the  $p \times p$  and  $q \times q$  identity matrices by  $I_p$  and  $I_q$ . Then the matrix

$$\begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ -C & I_q \end{bmatrix}$$

is invertible, particularly having a determinant of 1. If furthermore,  $A$  is invertible, then

$$\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix}$$

is also invertible with a determinant of  $\det(A^{-1}) = (\det(A))^{-1}$ .

*Proof.* For the first matrix, simply note that it is a lower-triangular matrix with all diagonal elements being 1 and therefore it has a determinant of 1. Therefore, by Properties 2.3.8, it is invertible. Similarly, by repeated cofactor expansions along the bottommost row for  $q$  times, the determinant of the second matrix can be seen to be  $\det(A^{-1}) = (\det(A))^{-1}$  (Properties 2.3.9). If  $A$  is invertible, then  $\det(A^{-1}) = (\det(A))^{-1}$  is nonzero by Properties 2.3.8 again, and

$$\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix}$$

will also be invertible. □

The above properties imply that these two matrices are the results from elementary row operations (Properties 2.2.11), and therefore, their product

$$\begin{aligned} \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ -C & I_q \end{bmatrix} \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} &= \begin{bmatrix} I_p A^{-1} + [\mathbf{0}]_{p \times q} [\mathbf{0}]_{q \times p} & I_p [\mathbf{0}]_{p \times q} + [\mathbf{0}]_{p \times q} I_q \\ (-C)A^{-1} + I_q [\mathbf{0}]_{q \times p} & -C[\mathbf{0}]_{p \times q} + I_q I_q \end{bmatrix} \\ &= \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \end{aligned}$$

can also be arrived via elementary row operations and is invertible as well (Properties 2.2.4). By multiplying this matrix to  $M$ , we have

$$\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A^{-1}A + [\mathbf{0}]_{p \times q}C & A^{-1}B + [\mathbf{0}]_{p \times q}D \\ -CA^{-1}A + I_q C & -CA^{-1}B + I_q D \end{bmatrix}$$

$$= \begin{bmatrix} I_p & A^{-1}B \\ -C + C & -CA^{-1}B + D \end{bmatrix}$$

$$= \begin{bmatrix} I_p & A^{-1}B \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}$$

The bottom right block,  $D - CA^{-1}B$ , is known as the **Schur complement** of block  $A$  in  $M$ , denoted as  $M/A$  and has the same shape  $q \times q$  as  $D$ . The above block multiplication constitutes a *block Gaussian Elimination* over the matrix  $M$  to make it *block upper-triangular*. It is not hard to see that

$$\begin{bmatrix} I_p & A^{-1}B \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} = \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}$$

Therefore,

$$\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix}$$

$$= \begin{bmatrix} I_p & A^{-1}B \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} = \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}$$

According to the above equation, if the Schur complement  $M/A = D - CA^{-1}B$  is also invertible, then the inverse of  $M$  will exist, because

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \left( \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \right)^{-1} \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix} \left( \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} \right)^{-1}$$

where the three matrices on R.H.S. are all invertible.<sup>5</sup> By Properties 2.2.3, we

---

<sup>5</sup>The invertibility of the first and last matrix follows the same arguments in Properties 8.3.1, while for the matrix in the middle we have required that  $D - CA^{-1}B$  has to be invertible, and its inverse can be readily seen to be

$$\begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}^{-1} = \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

arrive at

$$\begin{aligned}
 M^{-1} &= \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \\
 &= \left( \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix}^{-1} \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix}^{-1} \right)^{-1} \\
 &= \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \\
 &= \begin{bmatrix} I_p & -A^{-1}B(D - CA^{-1}B)^{-1} \\ [\mathbf{0}]_{q \times p} & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \\
 &= \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \\
 &= \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{bmatrix}
 \end{aligned}$$

To summarize, we have the following statements.

**Properties 8.3.2.** For the  $2 \times 2$  block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where  $A$  and  $D$  are square submatrices, if  $A$  and its Schur complement  $M/A = D - CA^{-1}B$  of block  $A$  are both invertible, then  $M$  is invertible with

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{bmatrix}$$

**Properties 8.3.3.** The determinant of the  $2 \times 2$  block matrix in Properties 8.3.2 is

$$\det(M) = \det(A) \det(D - CA^{-1}B) = \det(A) \det(M/A)$$

if  $A^{-1}$  is well-defined.

*Proof.* From the derivation above, we have

$$\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix} = \begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}$$

Evaluating the determinants of both sides leads to

$$\begin{aligned} & \det\left(\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p \times q} \\ -CA^{-1} & I_q \end{bmatrix}\right) \det(M) \det\left(\begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q \times p} & I_q \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} I_p & [\mathbf{0}]_{p \times q} \\ [\mathbf{0}]_{q \times p} & D - CA^{-1}B \end{bmatrix}\right) \end{aligned}$$

In the same vein of Properties 8.3.1, we then have

$$\begin{aligned} (\det(A))^{-1} \det(M)(1) &= \det(D - CA^{-1}B) \\ \det(M) &= \det(A) \det(D - CA^{-1}B) = \det(A) \det(M/A) \end{aligned}$$

□

**Example 8.3.2.** Use Properties 8.3.2 and 8.3.3 to compute the inverse and determinant of the following matrix

$$M = \left[ \begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & 1 \\ \hline -2 & -1 & 2 \end{array} \right]$$

via the partition above.

*Solution.* To use Properties 8.3.2, we need to first compute  $A^{-1}$  and  $M/A = D - CA^{-1}B$ . We leave to the readers for verifying that

$$\begin{aligned} A^{-1} &= \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \\ M/A &= D - CA^{-1}B = [2] - [-2 \quad -1] \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = [-1] \end{aligned}$$

Then by Properties 8.3.2, we have

$$\begin{aligned}
 M^{-1} &= \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{bmatrix} \\
 &= \frac{\left[ \begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 1 & 1 \end{array} \right] + \left[ \begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 1 & 1 \end{array} \right] \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} -2 & -1 \end{bmatrix} \left[ \begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 1 & 1 \end{array} \right] - \left[ \begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 1 & 1 \end{array} \right] \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} -2 & -1 \end{bmatrix}}{\left[ \begin{array}{cc|c} 1 & -2 & 0 \\ -1 & -2 & -1 \end{array} \right] \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}} \\
 &= \left[ \begin{array}{cc|c} -3 & 4 & -2 \\ 2 & -2 & 1 \\ -2 & 3 & -1 \end{array} \right]
 \end{aligned}$$

Meanwhile, by Properties 8.3.3,

$$\begin{aligned}
 \det(M) &= \det(A) \det(M/A) \\
 &= \det\left(\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}\right) \det([-1]) = (1)(-1) = -1
 \end{aligned}$$

□

Similar results are also available in terms of the Schur complement using block  $D$  instead of  $A$ .

**Properties 8.3.4.** For the  $2 \times 2$  block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where  $A$  and  $D$  are square submatrices, if  $D$  and its Schur complement  $M/D = A - BD^{-1}C$  of block  $D$  are both invertible, then  $M$  is invertible with

$$M^{-1} = \begin{bmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{bmatrix}$$

and its determinant can be computed by

$$\det(M) = \det(D) \det(A - BD^{-1}C) = \det(D) \det(M/D)$$

*Proof.* See Exercise 8.6. □

### 8.3.3 Restriction of a Linear Transformation, Direct Sum of a Matrix

In the last chapter we have discussed about linear transformations between two vector spaces, let's say, from  $\mathcal{U}$  to  $\mathcal{V}$ . Sometimes we only care about how the linear transformation works on some specific subspace  $\mathcal{W}$  of  $\mathcal{U}$ . This leads to the idea of **restriction** of a linear transformation as follows.

**Definition 8.3.5.** Given a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  and a proper subspace  $\mathcal{W} \subset \mathcal{U}$ , the restriction of  $T$  to  $\mathcal{W}$  is defined as

$$T|_{\mathcal{W}} : \mathcal{W} \rightarrow \mathcal{V}, T|_{\mathcal{W}}(\vec{w}) = T(\vec{w}) \text{ for any } \vec{w} \in \mathcal{W}.$$

In simpler terms,  $T|_{\mathcal{W}}$  works exactly as  $T$  but only defined on  $\mathcal{W}$ . Assume the vector spaces involved are all finite-dimensional, and  $\dim(\mathcal{W}) = r < n = \dim(\mathcal{U})$ .  $\mathcal{W}$  then has a basis  $\mathcal{B}_{\mathcal{W}}$  with  $r$  generating vectors, which by part (c) of Properties 6.2.7 can be extended to a new basis  $\mathcal{B}' = \mathcal{B}_{\mathcal{W}} \cup \mathcal{G}$  for  $\mathcal{U}$ , where  $\mathcal{G}$  contains  $n - r$  vectors and  $\mathcal{B}' = \mathcal{B}_{\mathcal{W}} \cup \mathcal{G}$  has exactly  $n$  linearly independent vectors by construction. Some may wonder why we suddenly talk about the restriction of a linear transformation here and the reason is that its related principles can be viewed from the stand point of a block matrix.

To see this, let  $\mathcal{B}_{\mathcal{W}} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(r)}\}$  and  $\mathcal{G} = \{\vec{u}^{(r+1)}, \dots, \vec{u}^{(n)}\}$ , and thus  $\mathcal{B}' = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \dots, \vec{u}^{(r)}, \vec{u}^{(r+1)}, \dots, \vec{u}^{(n)}\}$ . By Definition 6.2.9, since the vectors in  $\mathcal{B}' = \mathcal{B}_{\mathcal{W}} \cup \mathcal{G}$  are designed to be linear independent, the subspace  $\mathcal{W}^C$  generated by  $\mathcal{G}$  will be the complement of  $\mathcal{W}$  as a result of  $\mathcal{W} \oplus \mathcal{W}^C$  being a direct sum that produces  $\mathcal{U}$  (Properties 6.2.10). Any  $\vec{w} \in \mathcal{W} \subset \mathcal{U}$  will then have a coordinate representation of

$$(w_1, w_2, \dots, w_r, 0, \dots, 0)^T$$

in the  $\mathcal{B}'$  basis where components beyond the  $r$ -th index are all zeros. From the perspective of direct sum, it is the same as  $\vec{w} \oplus \mathbf{0}_{n-r} = (w_1, w_2, \dots, w_r)_{\mathcal{B}_{\mathcal{W}}}^T \oplus$

$(0, \dots, 0)_G^T$ , i.e.  $\vec{w}$  has all zero components in  $\mathcal{W}^C$ . By Definition 7.1.2, writing out the matrix representation of  $[T]_{B'}^H$  where  $H$  is an arbitrary basis for  $\mathcal{V}$  results in

$$[T]_{B'}^H = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} & a_1^{(r+1)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} & a_2^{(r+1)} & & a_2^{(n)} \\ \vdots & & & \vdots & & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} & a_m^{(r+1)} & \cdots & a_m^{(n)} \end{bmatrix}$$

Since we are only concerned about  $\vec{w} \in \mathcal{W} \subset \mathcal{U}$  (or  $\vec{w} \oplus \mathbf{0} \in \mathcal{W} \oplus \mathcal{W}^C$ ) when dealing with  $T|_W$ , when we apply  $T$  on  $\vec{w}$ , which is

$$\begin{aligned} [T]_{B'}^H [\vec{w}]_{B'} &= \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} \\ \vdots & & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} \end{bmatrix} \begin{bmatrix} a_1^{(r+1)} & \cdots & a_1^{(n)} \\ a_2^{(r+1)} & & a_2^{(n)} \\ \vdots & & \vdots \\ a_m^{(r+1)} & \cdots & a_m^{(n)} \end{bmatrix}_{B_W+G}^H \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{B_W+G} \\ &= [ T|_W \mid T|_{W^C} ]_{B_W+G}^H [\vec{w} \oplus \mathbf{0}]_{B_W+G}^T \end{aligned}$$

We can simply ignore  $[T|_{W^C}]_G^H$ , the block at the right of the  $[T]_{B'}^H$  partition as well as discard the all-zero components of  $[\vec{w}]_{B'}$  starting from the  $(r+1)$ -th index, and keep only the other block  $[T|_W]_{B_W}^H$  at the left and the  $[\vec{w}]_{B_W}$  part. The output of the truncated multiplication

$$[T|_W]_{B_W}^H [\vec{w}]_{B_W} = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} \\ \vdots & & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_r \end{bmatrix}$$

will represent the same vector in  $\mathcal{W} \subset \mathcal{U}$  as that mapped from the full form  $[T]_{B'}^H [\vec{w}]_{B'}$ .

**Properties 8.3.6.** For a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  between two finite-dimensional spaces, if a proper subspace  $\mathcal{W}$  of  $\mathcal{U}$  is generated by a basis  $\mathcal{B}_W = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \dots, \vec{w}^{(r)}\}$ , then the matrix representation of the restriction of  $T$  to  $\mathcal{W}$  with respect to  $\mathcal{B}_W$  and  $\mathcal{H}$  will be given by

$$[T|_W]_{B_W}^H = [[T(\vec{w}^{(1)})]_H | [T(\vec{w}^{(2)})]_H | \cdots | [T(\vec{w}^{(r)})]_H]$$

where  $\mathcal{H}$  is any basis for  $\mathcal{V}$ . (This can be compared to Definition 7.1.2.)

In general, the effect of a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  applied to  $\vec{u} \in \mathcal{U}$  is equivalent to the sum of responses from the restrictions of  $T$  to a set of subspaces  $\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_s$  where they constitute a direct sum  $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \cdots \oplus \mathcal{W}_s = \mathcal{U}$ , applied on the corresponding components  $\vec{w}_1 \in \mathcal{W}_1, \vec{w}_2 \in \mathcal{W}_2, \dots, \vec{w}_s \in \mathcal{W}_s$  of  $\vec{u} = \vec{w}_1 \oplus \vec{w}_2 \oplus \cdots \oplus \vec{w}_s$  in these smaller subspaces:  $T(\vec{u}) = T(\vec{w}_1) \oplus T(\vec{w}_2) \oplus \cdots \oplus T(\vec{w}_s)$ .

**Example 8.3.3.** Given a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  that has a matrix representation of

$$[T]_B^H = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}$$

with respect to some bases  $\mathcal{B} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}\}$  and  $\mathcal{H} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$  for  $\mathcal{U}$  and  $\mathcal{V}$ , find the restriction of  $T$  to  $\mathcal{W}$ , where  $\mathcal{W} \subset \mathcal{U}$  has a basis of  $B_W = \{\vec{w}^{(1)}, \vec{w}^{(2)}\}$ , with  $\vec{w}^{(1)} = \vec{u}^{(1)} + \vec{u}^{(2)}$  and  $\vec{w}^{(2)} = \vec{u}^{(1)} + \vec{u}^{(2)} + \vec{u}^{(3)}$ .

*Solution.* We will take an indirect approach of reconstructing the basis first by finding a third vector generating  $\mathcal{W}^C$  and producing the direct sum  $\mathcal{W} \oplus \mathcal{W}^C = \mathcal{U}$ . The change of coordinates matrix  $P_{B_W}^B$  from  $B_W$  to  $B$  as devised in Theorem 7.1.12 appropriate in this situation is a  $3 \times 2$  matrix instead since there are only two basis vectors in  $\mathcal{B}_W$ , and it can be easily seen to be

$$\begin{aligned} P_{B_W}^B &= [[\vec{w}^{(1)}]_B | [\vec{w}^{(2)}]_B] \\ &= [[\vec{u}^{(1)} + \vec{u}^{(2)}]_B | [\vec{u}^{(1)} + \vec{u}^{(2)} + \vec{u}^{(3)}]_B] \end{aligned}$$

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

and we are to find  $[\vec{w}^{(3)}]_B$  to complete a basis  $\mathcal{B}' = \mathcal{B}_W \cup \{\vec{w}^{(3)}\}$  and hence  $P_{B'}^B$ . An algorithm to do so, motivated by Footnote 12 in Chapter 6, is to apply Gaussian Elimination to  $P_{B_W}^B$  first and then append  $(0, 0, 1)^T$  to the right of the rref to make an identity matrix, and reverse the entire reduction procedure as follows.

$$\begin{array}{c} \left[ \begin{array}{cc} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{cc} 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{array} \right] \\ R_2 - R_1 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{cc} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{array} \right] \\ R_2 \leftrightarrow R_3 \\ \rightarrow \left[ \begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{array} \right] \\ R_1 - R_2 \rightarrow R_1 \end{array}$$

$$\begin{array}{c} \left[ \begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] \\ R_1 + R_2 \rightarrow R_1 \\ \rightarrow \left[ \begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array} \right] \\ R_2 \leftrightarrow R_3 \\ \rightarrow \left[ \begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{array} \right] \\ R_2 + R_1 \rightarrow R_2 \end{array}$$

So  $[\vec{w}^{(3)}]_B = (0, 1, 0)_B^T$  is a possible choice. This simple algorithm remains straight-forward when the number of dimensions and vectors to be appended

become much larger.<sup>6</sup> Now

$$P_{B'}^B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

and by Properties 7.2.1

$$\begin{aligned} [T]_{B'}^H &= [T]_B^H P_{B'}^B \\ &= \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 4 & 1 \\ 2 & 3 & 0 \\ 0 & 1 & -1 \end{bmatrix} \end{aligned}$$

The matrix representation of the restriction of  $T$  to  $\mathcal{W}$  with respect to  $\mathcal{B}_W$  agrees with the first two columns of  $[T]_{B'}^H$ . The third column of  $[T]_{B'}^H$  that characterizes the action of  $T|_{W_C}$ , is removed. These lead to

$$[T|_W]_{B_W}^H = \begin{bmatrix} 2 & 4 \\ 2 & 3 \\ 0 & 1 \end{bmatrix}$$

□

**Short Exercise:** Directly apply Properties 8.3.6 to redo the example above.<sup>7</sup>

With the concept of restriction, we can now introduce the matrix analogous of a direct sum. For a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$ , if the vector spaces

<sup>6</sup>In fact, we only need to keep track of the row swapping operations with full-zero rows.

<sup>7</sup> $[T(\vec{w}_1)]_H = [T(\vec{u}_1 + \vec{u}_2)]_H = [T]_B^H (1, 1, 0)_B^T = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}_B^H \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}_B = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}_H$  and this will be the first column of  $[T|_W]_{B_W}^H$ . The second column is derived similarly by evaluating  $[T(\vec{w}_2)]_H$ .

(finite-dimensional) involved are direct sums such that  $\mathcal{U} = \mathcal{W} \oplus \mathcal{W}^C$  and  $\mathcal{V} = \mathcal{Y} \oplus \mathcal{Y}^C$ , and the ranges

$$R(T|_W) \in \mathcal{Y} \quad R(T|_{W^C}) \in \mathcal{Y}^C$$

of the two restrictions are such that vectors in  $\mathcal{W}$  and  $\mathcal{W}^C$  are mapped by  $T$  to vectors in  $\mathcal{Y}$  and  $\mathcal{Y}^C$  separately, then  $T = T|_W \oplus T|_{W^C}$  is a **matrix direct sum** in the sense that the linear transformation  $T$  maps each of the complement subspaces in a direct sum of  $\mathcal{U}$  into the corresponding complement subspaces in a direct sum of  $\mathcal{V}$ . If we write the input vector  $\vec{u} = \vec{w} \oplus \vec{w}^C$  as a direct sum where  $\vec{w} \in \mathcal{W}$  and  $\vec{w}^C \in \mathcal{W}^C$ , then the output vector will also be a direct sum  $\vec{v} = \vec{y} \oplus \vec{y}^C$  where  $\vec{y} = T|_W(\vec{w}) = T(\vec{w})$  and  $\vec{y}^C = T|_{W^C}(\vec{w}^C) = T(\vec{w}^C)$ , which are obtained by first computing  $T(\vec{w})$  and  $T(\vec{w}^C)$  individually, and then directly concatenating them together.

**Definition 8.3.7** (Matrix Direct Sum). The direct sum of two matrices acting as linear transformations  $T_1 : \mathcal{U}_1 \rightarrow \mathcal{V}_1$  and  $T_2 : \mathcal{U}_2 \rightarrow \mathcal{V}_2$  is  $T = T_1 \oplus T_2$  such that for any vector direct sum  $\vec{u} = \vec{u}_1 \oplus \vec{u}_2$  in  $\mathcal{U} = \mathcal{U}_1 \oplus \mathcal{U}_2$ , applying  $T$  on  $\vec{u}$  will yield an output of a vector direct sum  $T(\vec{u}) = \vec{v} = \vec{v}_1 \oplus \vec{v}_2$  in  $\mathcal{V}_1 \oplus \mathcal{V}_2$  as well, where  $\vec{v}_1 = T_1(\vec{u}_1) = T|_{\mathcal{U}_1}(\vec{u}_1) \in \mathcal{V}_1$  and  $\vec{v}_2 = T_2(\vec{u}_2) = T|_{\mathcal{U}_2}(\vec{u}_2) \in \mathcal{V}_2$ . The matrix direct sum is then the matrix representation of  $T = T_1 \oplus T_2$  with respect to the direct sum bases for  $\mathcal{U}_1 \oplus \mathcal{U}_2$  and  $\mathcal{V}_1 \oplus \mathcal{V}_2$ .

Using the above definition, if  $\mathcal{U}_1$  and  $\mathcal{U}_2$  has a basis  $\mathcal{B}_1 = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \dots, \vec{w}^{(r)}\}$  and  $\mathcal{B}_2 = \{\vec{w}^{(r+1)}, \vec{w}^{(r+2)}, \dots, \vec{w}^{(n)}\}$ , and  $\mathcal{V}_1$  and  $\mathcal{V}_2$  has a basis  $\mathcal{H}_1 = \{\vec{y}^{(1)}, \vec{y}^{(2)}, \dots, \vec{y}^{(s)}\}$  and  $\mathcal{H}_2 = \{\vec{y}^{(s+1)}, \vec{y}^{(s+2)}, \dots, \vec{y}^{(m)}\}$ , where  $r, n, s, m$  are some integers, then  $T = T_1 \oplus T_2$  will have a *block diagonal* matrix representation of

$$[T]_{B_1+B_2}^{H_1+H_2} \equiv \begin{bmatrix} ([T_1]_{B_1}^{H_1})_{s \times r} & [\mathbf{0}]_{s \times (n-r)} \\ [\mathbf{0}]_{(m-s) \times r} & ([T_2]_{B_2}^{H_2})_{(m-s) \times (n-r)} \end{bmatrix}_{B_1+B_2}^{H_1+H_2}$$

with respect to  $\mathcal{B}_1 \oplus \mathcal{B}_2$  and  $\mathcal{H}_1 \oplus \mathcal{H}_2$ . To see this, let  $\vec{u} = \vec{u}^{(1)} \oplus \vec{u}^{(2)}$ ,  $\vec{u}^{(1)} \in \mathcal{U}_1$  and  $\vec{u}^{(2)} \in \mathcal{U}_2$ , then

$$T(\vec{u}) = [T]_{B_1+B_2}^{H_1+H_2} [\vec{u}]_{B_1+B_2}$$

$$\begin{aligned}
 &= \begin{bmatrix} ([T_1]_{B_1}^{H_1})_{s \times r} & [\mathbf{0}]_{s \times (n-r)} \\ [\mathbf{0}]_{(m-s) \times r} & ([T_2]_{B_2}^{H_2})_{(m-s) \times (n-r)} \end{bmatrix}_{B_1+B_2}^{H_1+H_2} \begin{bmatrix} ([\vec{u}_1]_{B_1})_r \\ ([\vec{u}_2]_{B_2})_{n-r} \end{bmatrix}_{B_1+B_2} \\
 &= \begin{bmatrix} ([T_1]_{B_1}^{H_1})_{s \times r}([\vec{u}_1]_{B_1})_r + [\mathbf{0}]_{s \times (n-r)}([\vec{u}_2]_{B_2})_{n-r} \\ [\mathbf{0}]_{(m-s) \times r}([\vec{u}_1]_{B_1})_r + ([T_2]_{B_2}^{H_2})_{(m-s) \times (n-r)}([\vec{u}_2]_{B_2})_{n-r} \end{bmatrix}_{H_1+H_2} \\
 &= \begin{bmatrix} ([T_1]_{B_1}^{H_1}[\vec{u}_1]_{B_1})_s \\ ([T_2]_{B_2}^{H_2}[\vec{u}_2]_{B_2})_{m-s} \end{bmatrix}_{H_1+H_2} \\
 &\equiv T_1(\vec{u}_1) \oplus T_2(\vec{u}_2)
 \end{aligned}$$

where the image is a direct sum composed of  $T_1(\vec{u}_1) \equiv [T_1]_{B_1}^{H_1}[\vec{u}_1]_{B_1} \in \mathcal{V}_1$  and  $T_2(\vec{u}_2) \equiv [T_2]_{B_2}^{H_2}[\vec{u}_2]_{B_2} \in \mathcal{V}_2$  from applying  $T_1$  and  $T_2$  separately to the preimages  $\vec{u}_1 \in \mathcal{U}_1$  and  $\vec{u}_2 \in \mathcal{U}_2$  in the two subspaces. For example, the matrix direct sum of  $A \oplus B$  given

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 8 \\ 1 & 1 \\ 4 & 0 \end{bmatrix}$$

is

$$A \oplus B = \left[ \begin{array}{ccc|cc} 1 & 2 & 3 & 0 & 0 \\ 4 & 5 & 6 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 8 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 4 & 0 \end{array} \right]$$

in which  $A$  and  $B$  are matrices representing linear transformations of  $\mathcal{U}_1 \rightarrow \mathcal{V}_1$  and  $\mathcal{U}_2 \rightarrow \mathcal{V}_2$ , where  $\mathcal{U}_1, \mathcal{U}_2, \mathcal{V}_1, \mathcal{V}_2$  have dimensions of  $3, 2, 2, 3$ . Subsequently,  $A \oplus B$  is a matrix corresponding to a mapping from  $\mathcal{U} = \mathcal{U}_1 \oplus \mathcal{U}_2$  to  $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$ . Finally, the matrix direct sum of more than two matrices  $A_1, A_2, A_3, \dots, A_{n-1}, A_n$  are defined recursively just like a vector direct sum as

$$A_1 \oplus A_2 \oplus A_3 \oplus \cdots \oplus A_{n-1} \oplus A_n$$

$$= (\cdots ((A_1 \oplus A_2) \oplus A_3) \oplus \cdots A_{n-1}) \oplus A_n$$

As another example, sometimes we may regard a matrix that does not look like a direct sum to be effectively one with respect to appropriate coordinate systems in a broader sense.

**Example 8.3.4.** For a linear transformation  $T : \mathcal{U} \rightarrow \mathcal{V}$  that has a matrix representation of

$$[T]_B^H = \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix}$$

with respect to some bases  $\mathcal{B} = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \vec{u}^{(4)}\}$ ,  $\mathcal{H} = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$ , show that it can turn into a matrix direct sum if the coordinate systems are changed according to  $\mathcal{B}' = \{\vec{u}^{(1)'}, \vec{u}^{(2)'}, \vec{u}^{(3)'}, \vec{u}^{(4)'}\}$ ,  $\mathcal{H}' = \{\vec{v}^{(1)'}, \vec{v}^{(2)'}, \vec{v}^{(3)'}\}$ , where

$$\begin{aligned} \vec{u}^{(1)'} &= \vec{u}^{(1)} & \vec{v}^{(1)'} &= \vec{v}^{(1)} + \vec{v}^{(2)} \\ \vec{u}^{(2)'} &= \vec{u}^{(3)} & \vec{v}^{(2)'} &= -\vec{v}^{(2)} + \vec{v}^{(3)} \\ \vec{u}^{(3)'} &= \vec{u}^{(1)} + \vec{u}^{(2)} & \vec{v}^{(3)'} &= \vec{v}^{(1)} - \vec{v}^{(3)} \\ \vec{u}^{(4)'} &= \vec{u}^{(1)} + \vec{u}^{(4)} \end{aligned}$$

*Solution.* The change of coordinate matrices suggested by Properties 7.2.1 are

$$P_{B'}^B = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_B^B \quad Q_{H'}^H = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}_{H'}^H$$

and the new matrix representation of  $T$  is

$$[T]_{B'}^{H'} = (Q_{H'}^H)^{-1} [T]_B^H P_{B'}^B$$

$$\begin{aligned}
 &= \left( \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}_{H'}^H \right)^{-1} \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix}_B^H \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{B'}^B \\
 &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}_H^H \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix}_B^H \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{B'}^B \\
 &= \begin{bmatrix} 1 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}_{B'}^{H'}
 \end{aligned}$$

where

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

□

## 8.4 Python Programming

Complex numbers in Python are written as `a+bj`. For example,

```

z_1 = 1 - 2j
z_2 = 3 + 1j
print(z_1, z_2)

```

returns  $(1-2j)$   $(3+1j)$  (1 in front of `j` is needed). Conjugate, modulus and argument can be found by

```

import numpy as np
from scipy import linalg

print(np.conjugate(z_1))
print(np.abs(z_2))
print(np.angle(z_1))

```

which yields  $(1+2j)$ ,  $3.162278(\sqrt{10})$ , and  $-1.10715$  (in radians). Addition, subtraction, multiplication and division of complex numbers in Python are coded just like if they are ordinary numbers.

```
print(3*z_1 + z_2) # (6-5j)
print(z_1 - 2*z_2) # (-5-4j)
print(z_1 * z_2) # (5-5j)
print(z_1 / z_2) # (0.1-0.7000000000000001j), floating-point
                  error
```

The same goes for complex matrices and their multiplication. As an example,

```
A = np.array([[4., 2.+1.j],
             [-3.j, 1-2.j]])
B = np.array([[3-1.j, 0],
              [2-5.j, 4.j]])
print(A @ B)
```

produces

```
[[ 21.-12.j -4. +8.j]
 [-11.-18.j  8. +4.j]]
```

Conjugate and Hermitian transpose of a complex matrix is retrieved by

```
print(np.conjugate(A))
print(np.conjugate(B).T) # Hermitian transpose = conjugate +
                        transpose, or just .H if np.matrix is used instead of np.
                        array
```

resulting in

```
[[ 4.-0.j  2.-1.j]
 [-0.+3.j  1.+2.j]]
 [[3.+1.j  2.+5.j]
 [0.-0.j  0.-4.j]]
```

The usual functions for inverse and determinant also work on complex matrices. The lines

```
print(linalg.det(A))
print(linalg.inv(B))
```

give

```
(1-2j)
[[3.00000000e-01+0.1j  0.00000000e+00+0.j    ]
 [3.25000000e-01+0.275j 1.38777878e-17-0.25j]] # Again, round
-off error
```

We can use 1D complex matrices as complex vectors.

```
u = np.array([1+1.j, -3.j, 2])
v = np.array([5, 1+2.j, 1-4.j])
```

The complex dot product between two complex vectors are then found by vdot

```
print(np.vdot(u,v).conj())
```

notice that a conjugate is needed since numpy defines complex dot product with a different convention such that the first complex vector is conjugated instead of the second one. It then outputs the correct answer of  $(1+10j)$ . The norm function still works fine, e.g. `print(linalg.norm(u))` gives  $3.87298$  ( $\sqrt{(1+i)(1-i) + (-3i)(3i) + (2)^2} = \sqrt{15}$ ). Finally, for the discussion in the last section, to build a block matrix using submatrices, we can use the block function as

```
C = np.array([1, 3+2.j])
D = np.array([-1.j, 2])

print(np.block([[A,B],
               [C,D]]))
```

which outputs

```
[[ 4.+0.j  2.+1.j  3.-1.j  0.+0.j]
 [-0.-3.j  1.-2.j  2.-5.j  0.+4.j]
 [ 1.+0.j  3.+2.j -0.-1.j  2.+0.j]]
```

Another example of constructing a block diagonal matrix is

```
print(np.block([[A, np.zeros([2,3])],
               [np.zeros([3,2]), np.identity(3)]]))
```

generating

```
[[ 4.+0.j  2.+1.j  0.+0.j  0.+0.j  0.+0.j]
 [-0.-3.j  1.-2.j  0.+0.j  0.+0.j  0.+0.j]
 [ 0.+0.j  0.+0.j  1.+0.j  0.+0.j  0.+0.j]
 [ 0.+0.j  0.+0.j  0.+0.j  1.+0.j  0.+0.j]]
```

```
[ 0.+0.j  0.+0.j  0.+0.j  0.+0.j  1.+0.j]]
```

## 8.5 Exercises

**Exercise 8.1** By considering Euler's formula stated in Definition 8.1.6, we have for any  $\theta, \phi$

$$\begin{aligned} e^{i\theta} &= \cos \theta + i \sin \theta \\ e^{i\phi} &= \cos \phi + i \sin \phi \\ e^{i(\theta+\phi)} &= \cos(\theta + \phi) + i \sin(\theta + \phi) \end{aligned}$$

If we take the product of the first two equations, we also have

$$e^{i(\theta+\phi)} = (\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi)$$

By equating the two expressions of  $e^{i(\theta+\phi)}$ , expand and compare the real and imaginary parts, prove the famous angle sum identities, which are

$$\begin{aligned} \cos(\theta + \phi) &= \cos \theta \cos \phi - \sin \theta \sin \phi \\ \sin(\theta + \phi) &= \sin \theta \cos \phi + \cos \theta \sin \phi \end{aligned}$$

Hence, by either using the results above, or the De Moivre's Formula, prove the double angle formula shown below.

$$\begin{aligned} \cos(2\theta) &= \cos^2 \theta - \sin^2 \theta \\ \sin(2\theta) &= 2 \sin \theta \cos \theta \end{aligned}$$

**Exercise 8.2** Evaluate

- (a)  $(1 + i)(3 - 2i)$ ,
- (b)  $\overline{(2 - i)/(4 + i)}$ ,

(c)  $(3 + 5i)\overline{(1+i)/(2-3i)}$

as well as their modulus and argument.

**Exercise 8.3** For  $\vec{u} = (1+i, 2-i, 3)^T$ ,  $\vec{v} = (2+i, 1-2i, i)^T$ , and  $\vec{w} = (-i, 3, 1-i)^T$ , find

- (a)  $\vec{u} \cdot \vec{v}$ ,
- (b)  $(\vec{u} + \vec{v}) \cdot (\vec{u} - \vec{w})$ ,
- (c)  $\|\vec{u}\|\vec{v} - \|\vec{v}\|\vec{w}$ .

**Exercise 8.4** For the two complex matrices below,

$$A = \begin{bmatrix} 1+i & -i & 3 \\ 0 & 2-i & 1 \\ -1 & i & 2 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 2-i & i \\ -i & 3+i & 1-i \\ 0 & 1 & 2i \end{bmatrix}$$

compute  $AB$ , and verify  $(AB^*)^* = BA^*$ .

**Exercise 8.5** For the matrix

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

find its determinant and inverse.

**Exercise 8.6** Prove the formulae in Properties 8.3.4, by noting that

$$\begin{bmatrix} I_p & 0_{p \times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} = \begin{bmatrix} I_p & 0_{p \times q} \\ 0_{q \times p} & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & 0_{p \times q} \\ -C & I_q \end{bmatrix}$$

and

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & 0_{p \times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ 0_{q \times p} & I_q \end{bmatrix}$$

**Exercise 8.7** Write down the direct sum of the following three matrices.

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

**Exercise 8.8** Show that given two bases  $\mathcal{B} = \{\cos x, \sin x, 1, x, x^2\}$  and  $\mathcal{H} = \{\cos x, \sin x, 1, x\}$  which generate vector spaces  $\mathcal{U}$  and  $\mathcal{V}$  respectively, the differentiation operator  $T(f(x)) = f'(x) : \mathcal{U} \rightarrow \mathcal{V}$  has a  $2 \times 2$  block matrix direct sum representation.



## Chapter 9

# Eigenvalues and Eigenvectors

---

In this section we will discuss a very important topic in Linear Algebra, the *eigenvalue-eigenvector* problem. By finding the eigenvectors of a square matrix which span subspaces that are *invariant* under the corresponding linear operator, it is sometimes possible to obtain a coordinate basis such that the matrix can be *diagonalized*, i.e. become a diagonal matrix under that particular change of coordinates. One of the practical usages of *diagonalization* is to solve systems of linear ordinary differential equations (ODEs) which are also commonly seen in many areas of Earth Science.

## 9.1 Eigenvalues and Eigenvectors of a Square Matrix

### 9.1.1 Definition of Eigenvalues and Eigenvectors

Consider a linear operator/endomorphism  $T : \mathcal{V} \rightarrow \mathcal{V}$ , an interesting question is about if a vector  $\vec{v} \in \mathcal{V}$  under this mapping will remain stationary in direction such that the image  $T(\vec{v}) = \lambda v$  is a scalar multiple of the original vector, or in other words, the effect of  $T$  on  $\vec{v}$  is simply a rescaling. In this situation, the vector  $\vec{v}$  is known as an **eigenvector** of  $T$  and the factor  $\lambda$  is the corresponding **eigenvalue**. Since a linear operator is a mapping between a vector space itself,

it has a square matrix representation under any basis. This fact extends the ideas of eigenvalues and eigenvectors to square matrices.

**Definition 9.1.1.** Given a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$ , we call  $\lambda$  and  $\vec{v}_\lambda$  its eigenvalue and eigenvector if

$$T(\vec{v}_\lambda) = \lambda \vec{v}_\lambda$$

Similarly, given an  $n \times n$  square matrix  $A$ ,  $\lambda$  and  $\vec{v}_\lambda$  will be an eigenvalue and eigenvector for it when

$$A\vec{v}_\lambda = \lambda \vec{v}_\lambda$$

This is a special case in which a vector space  $\mathcal{V}$  is finite-dimensional,  $\dim(\mathcal{V}) = n$ , and  $A = [T]_B$  is just the matrix representation of  $T$  with respect to some basis  $\mathcal{B}$ .

Notice that there can be more than one eigenvalues and eigenvectors. An example is given by the matrix

$$A = \begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix}$$

It can be seen that the vector  $\vec{v}_1 = (1, 2)^T$  is an eigenvector of  $A$ , as

$$\begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

that corresponds to an eigenvalue of  $\lambda = 2$ . Meanwhile,  $\vec{v}_2 = (1, -2)^T$  is another eigenvector that has an eigenvalue of  $\lambda = 0$ , since

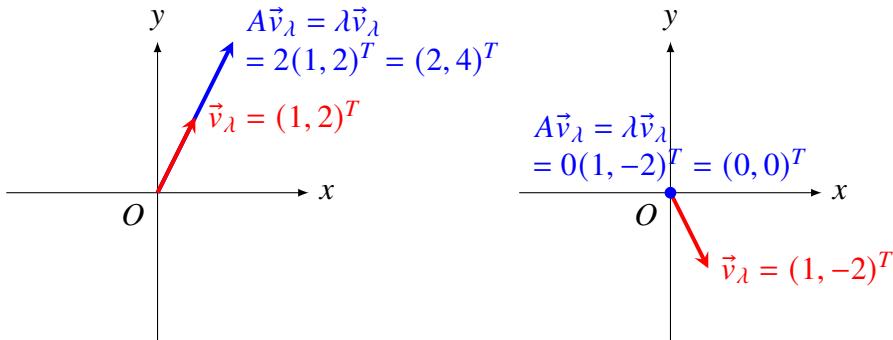
$$\begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

We emphasize that an eigenvalue of zero is perfectly valid which represents vanishing.

## 9.1 Eigenvalues and Eigenvectors of a Square Matrix

---

**Short Exercise:** Prove that all vectors in form of  $s(1, 2)^T$ , where  $s$  is any number, are eigenvectors for the matrix  $A$  above with  $\lambda = 2$ .<sup>1</sup>



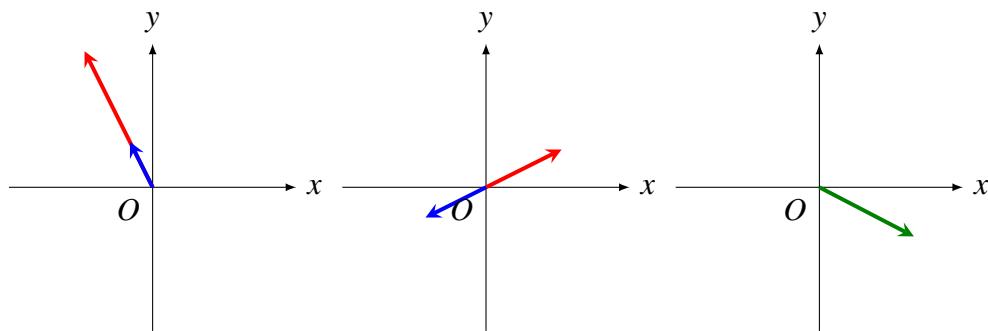
Illustrations for the example above with  $\lambda = 2 > 1$  (Extension), and  $\lambda = 0$  (Vanished). The red/blue vector is before/after applying the linear transformation.

There are infinitely many eigenvectors which are oriented in the same direction for a single eigenvalue as seen in the remark of the last short exercise. Particularly, they are actually the span of any one of these eigenvectors that is non-zero. Thus, along a single direction, only one of them is needed for representation, and its span is at the same time a subspace by Properties 6.1.6. This subspace is known as the **eigenspace** corresponding to that eigenvalue. Moreover, there may be more than one linearly independent eigenvectors for the same eigenvalue, and the dimension of eigenspace generated by them will be greater than one as well. In addition, the zero vector, technically, can be the eigenvectors of any matrix since  $A\vec{0} = \vec{0} = \lambda\vec{0}$  for any matrix  $A$  and scalar  $\lambda$ . However, it is a trivial solution, plus more importantly the zero vector is always linearly dependent by definition, and will not be taken into consideration (unlike the totally fine eigenvalue of zero).

Below is the visualization of some other possibilities of eigenvector rescaling.

---

<sup>1</sup>  $\begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix} \left( s \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right) = s \begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = s \begin{bmatrix} 2 \\ 4 \end{bmatrix} = 2 \left( s \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$ . In general, if  $A\vec{v}_\lambda = \lambda\vec{v}_\lambda$  so that  $\vec{v}_\lambda$  is some non-zero eigenvector, then  $A(s\vec{v}_\lambda) = sA\vec{v}_\lambda = s\lambda\vec{v}_\lambda = \lambda(s\vec{v}_\lambda)$  and therefore all of its non-zero scalar multiples  $s\vec{v}_\lambda$  is also an eigenvector.



Contraction ( $0 < \lambda < 1$ ), Reversal ( $\lambda < 0$ ), Unchanged ( $\lambda = 1$ ).

### 9.1.2 Finding Eigenvalues and Eigenvectors with Characteristic Polynomials

To find the eigenvalues of a square matrix (or the linear transformation it represents), rearrange the equation in Definition 9.1.1 relating eigenvalues and eigenvectors to obtain

$$\begin{aligned} A\vec{v}_\lambda &= \lambda\vec{v}_\lambda \\ A\vec{v}_\lambda &= \lambda I\vec{v}_\lambda && (I\vec{v} = \vec{v} \text{ for any } \vec{v}) \\ (A - \lambda I)\vec{v}_\lambda &= \mathbf{0} \end{aligned}$$

The last line constitutes a homogeneous linear system  $B\vec{v}_\lambda = \vec{0}$  where  $B = A - \lambda I$ . For this system to have a non-trivial solution and hence an eigenvector, it is required that  $\det(B) = \det(A - \lambda I) = 0$  from Theorem 3.1.2. The relationship  $p(\lambda) = \det(A - \lambda I) = 0$  is called the **characteristic equation**. The roots for  $\lambda$  of the *characteristic polynomial*  $p(\lambda)$  are then the desired eigenvalues.

**Definition 9.1.2** (Characteristic Equation). The characteristic equation for a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  over a vector space  $\mathcal{V}$  with  $\dim(\mathcal{V}) = n$  (or an  $n \times n$  square matrix  $A$ ) is

$$\det([T]_B - \lambda I) = 0 \quad \text{or} \quad \det(A - \lambda I) = 0$$

where  $\mathcal{B}$  is any coordinate basis for  $\mathcal{V}$ . The L.H.S. ( $p(\lambda) = \det([T]_B - \lambda I)$  or

$\det(A - \lambda I)$ ), when expanded, constitutes a **characteristic polynomial** in  $\lambda$  of degree  $n$ , the roots of which are the eigenvalues of  $T$  (or  $A$ ).

Short Exercise: By inspection, find all three eigenvalues of the matrix.<sup>2</sup>

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

For each eigenvalue there corresponds at least one eigenvectors by definition. The required linearly independent eigenvectors that generate the eigenspace for a particular eigenvalue  $\lambda_j$  can be found as the general solution to the matrix equation  $(A - \lambda_j I)\vec{v} = \mathbf{0}$ , or in other words, (the basis for) the null space of  $A - \lambda_j I$ ,  $\mathcal{N}(A - \lambda_j I)$ . The number of linearly independent eigenvectors (dimensions) in the eigenspace for  $\lambda_j$  is hence the dimension of the null space (Definition 6.3.8) of  $A - \lambda_j I$ , known as the **geometric multiplicity**. A closely related quantity is the **algebraic multiplicity** which is the number of times where  $\lambda_j$  appears as the root to the characteristic polynomial. It can be shown that the geometric multiplicity of  $\lambda_j$  is capped by its algebraic multiplicity.

**Theorem 9.1.3.** For an eigenvalue  $\lambda_j$  to a square matrix  $A$ , its geometric multiplicity is equal to  $\dim(\mathcal{N}(A - \lambda_j I))$ . Meanwhile its algebraic multiplicity is the power  $k_j$  of the factor  $(\lambda_j - \lambda)^{k_j}$  in the characteristic polynomial  $p(\lambda)$ . With these two quantities defined, we have

$$1 \leq \text{Geometric Multiplicity} \leq \text{Algebraic Multiplicity}$$

for every distinct eigenvalue  $\lambda_j$ .

We defer the proof showing that geometric multiplicity is always smaller than algebraic multiplicity until we introduce diagonalization later in this chapter.

---

<sup>2</sup>They are  $\lambda = 1, 2, 3$ .

**Example 9.1.1.** Find all eigenvalues and eigenvectors for the matrix

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

*Solution.* The characteristic equation is

$$\begin{aligned} p(\lambda) &= \det(A - \lambda I) = \begin{vmatrix} 1 - \lambda & -1 \\ 0 & 1 - \lambda \end{vmatrix} \\ &= (1 - \lambda)^2 = 0 \end{aligned}$$

Apparently, there is only one eigenvalue  $\lambda = 1$ , which has an algebraic multiplicity of 2. Possible eigenvectors are then found by solving  $A - \lambda I = \mathbf{0}$ :

$$\left[ \begin{array}{cc|c} 1 - 1 & -1 & 0 \\ 0 & 1 - 1 & 0 \end{array} \right] = \left[ \begin{array}{cc|c} 0 & -1 & 0 \\ 0 & 0 & 0 \end{array} \right]$$

where the general solution is easily seen to be  $t(1, 0)^T$ . So for the eigenvalue  $\lambda = 1$ , there is only one linearly independent eigenvector of  $(1, 0)^T$ , which implies a geometric multiplicity of 1.  $\square$

**Example 9.1.2.** For the matrix

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

find its eigenvalues and eigenvectors.

*Solution.* The characteristic polynomial is, by Sarrus' Rule (Properties 2.3.1)

$$\begin{aligned} \det(A - \lambda I) &= \begin{vmatrix} 1 - \lambda & 3 & 1 \\ 0 & 1 - \lambda & 0 \\ 1 & 0 & 2 - \lambda \end{vmatrix} = (1 - \lambda)(1 - \lambda)(2 - \lambda) - (1)(1 - \lambda)(1) \\ &= (1 - \lambda)((2 - 3\lambda + \lambda^2) - 1) \end{aligned}$$

## 9.1 Eigenvalues and Eigenvectors of a Square Matrix

---

$$= (1 - \lambda)(1 - 3\lambda + \lambda^2)$$

The roots and thus eigenvalues are  $\lambda = 1$ , as well as

$$\begin{aligned}\lambda &= \frac{-(-3) \pm \sqrt{(-3)^2 - 4(1)(1)}}{2} \\ &= \frac{3}{2} \pm \frac{\sqrt{5}}{2}\end{aligned}$$

Particularly, for the eigenvalue  $\lambda = \frac{3}{2} + \frac{\sqrt{5}}{2}$ , the eigenvector is inferred from the homogeneous system  $A - \lambda I = \mathbf{0}$

$$\begin{array}{c} \left[ \begin{array}{ccc|c} \frac{1}{2} - \frac{\sqrt{5}}{2} & 3 & 1 & 0 \\ 0 & -\frac{1}{2} - \frac{\sqrt{5}}{2} & 0 & 0 \\ 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \end{array} \right] \\ \xrightarrow{\quad} \left[ \begin{array}{ccc|c} 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{2} - \frac{\sqrt{5}}{2} & 3 & 1 & 0 \end{array} \right] \quad R_1 \leftrightarrow R_3 \\ \xrightarrow{\quad} \left[ \begin{array}{ccc|c} 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{array} \right] \quad \left( \frac{1}{-\frac{1}{2} - \frac{\sqrt{5}}{2}} \right) R_2 \rightarrow R_2 \\ \xrightarrow{\quad} \left[ \begin{array}{ccc|c} 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 - (-\frac{1}{2} - \frac{\sqrt{5}}{2})R_1 \rightarrow R_3 \\ \xrightarrow{\quad} \left[ \begin{array}{ccc|c} 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 - 3R_2 \rightarrow R_3 \end{array}$$

whose eigenvector can be seen to be  $(-\frac{1}{2} + \frac{\sqrt{5}}{2}, 0, 1)^T$  for  $\lambda = \frac{3}{2} + \frac{\sqrt{5}}{2}$  by letting  $z$  be the free variable.  $\square$

Short Exercise: Find the eigenvectors for other remaining eigenvalues.<sup>3</sup>

---

<sup>3</sup>For  $\lambda = 1$ , the eigenvector is  $(-1, -\frac{1}{3}, 1)^T$ . For  $\lambda = \frac{3}{2} - \frac{\sqrt{5}}{2}$ , it is  $(-\frac{1}{2} - \frac{\sqrt{5}}{2}, 0, 1)^T$ . Note that for all three eigenvalues their algebraic and geometric multiplicity are both 1.

**Example 9.1.3.** Find all eigenvalues and eigenvectors for the matrix

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

*Solution.* Again, the characteristic polynomial can be found by Sarrus' Rule as below.

$$\begin{aligned} p(\lambda) = \det(A - \lambda I) &= \begin{vmatrix} 1-\lambda & 1 & -1 \\ -1 & 3-\lambda & -1 \\ 1 & -1 & 3-\lambda \end{vmatrix} \\ &= [(1-\lambda)(3-\lambda)(3-\lambda) + (1)(-1)(1) + (-1)(-1)(-1)] \\ &\quad - [(-1)(3-\lambda)(1) + (1-\lambda)(-1)(-1) + (1)(-1)(3-\lambda)] \\ &= -\lambda^3 + 7\lambda^2 - 16\lambda + 12 \\ &= (2-\lambda)^2(3-\lambda) \end{aligned}$$

Hence the eigenvalues are  $\lambda = 2, 3$  with an algebraic multiplicity of 2 and 1 respectively. For  $\lambda = 2$ , its eigenvectors are retrieved via solving

$$\left[ \begin{array}{ccc|c} 1-2 & 1 & -1 & 0 \\ -1 & 3-2 & -1 & 0 \\ 1 & -1 & 3-2 & 0 \end{array} \right] = \left[ \begin{array}{ccc|c} -1 & 1 & -1 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 0 \\ -1 & 1 & -1 & 0 \end{array} \right] \quad R_1 \leftrightarrow R_3$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_2 + R_1 \rightarrow R_2$$

$$\quad \quad \quad R_3 + R_1 \rightarrow R_3$$

We have only one leading 1 and two non-pivotal columns, so we can assign two free variables. For that, let  $y = s, z = t$ , then  $x = s - t$ . The general solution is then

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s-t \\ s \\ t \end{bmatrix} = s \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

and hence the two linearly independent eigenvectors for  $\lambda = 2$  are  $(1, 1, 0)^T$  and  $(-1, 0, 1)^T$ . In this case the geometric multiplicity is 2 and equal to the algebraic multiplicity. On the other hand, for  $\lambda = 3$ , we have

$$\begin{array}{c}
 \left[ \begin{array}{ccc|c} -2 & 1 & -1 & 0 \\ -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 1 & -1 & 0 \end{array} \right] \quad R_1 \leftrightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 0 \end{array} \right] \quad R_2 + R_1 \rightarrow R_2 \\
 \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 \end{array} \right] \quad R_3 + 2R_1 \rightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad -R_2 \rightarrow R_2 \\
 \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3
 \end{array}$$

Now we have one non-pivotal column only and it is possible to take  $z = t$  as the free variable. This will eventually yield  $(-1, -1, 1)^T$  as the only eigenvector for  $\lambda = 3$ .  $\square$

**Example 9.1.4.** Show that

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

admits no real eigenvalues. How about if we work over  $\mathbb{C}$ ?

*Solution.* The characteristic equation is simply

$$\begin{aligned}
 p(\lambda) &= \det(A - \lambda I) = \begin{vmatrix} -\lambda & -1 \\ 1 & -\lambda \end{vmatrix} = 0 \\
 (-\lambda)^2 - (1)(-1) &= \lambda^2 + 1 = 0
 \end{aligned}$$

which has no real solution. So if we are confined to work over  $\mathbb{R}$  as the scalar for a real vector space, there is no possible eigenvalue  $\lambda$  such that  $A\vec{v}_\lambda = \lambda\vec{v}_\lambda$ , as the scalar multiplication on R.H.S. only allows  $\lambda$  to be a real number. Geometrically, for any (real) vector  $\vec{v} = (x, y)^T$  on a flat plane, multiplying by  $A$  leads to

$$A\vec{v} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ x \end{bmatrix}$$

which rotates the vector in an anti-clockwise sense by  $90^\circ$ . So it is impossible for a vector to remain oriented in the same direction after  $A$  is applied, consistent with the absence of any real eigenvalue. However, if we relax the constraint and work with  $\mathbb{C}$ , then the eigenvalues are clearly  $\lambda = \pm i$ . For  $\lambda = i$ , the eigenvector can be found from

$$\left[ \begin{array}{cc|c} -i & -1 & 0 \\ 1 & -i & 0 \end{array} \right]$$

We can use a trick to immediately ignore the second row since the algebraic multiplicity of  $\lambda = i$  and hence nullity of the system above is 1 so there will be one redundant row, in this system which only has two rows. Now by considering the first row we can easily find that the corresponding eigenvector is  $(i, 1)^T$ . Similarly, the eigenvector of  $\lambda = -i$  is  $(-i, 1)^T$ . We will talk more about complex eigenvalues in Section 9.2.4 and Chapter 10.  $\square$

Some notable properties of eigenvalues and eigenvectors are provided below.

**Properties 9.1.4.** Eigenvectors for distinct eigenvalues are linearly independent.

*Proof.* Let  $\vec{v}_1$  and  $\vec{v}_2$  are two eigenvectors corresponding to two different eigenvalues  $\lambda_1$  and  $\lambda_2$  of some square matrix  $A$ . By Theorem 6.1.9, we want to show that  $c_1\vec{v}_1 + c_2\vec{v}_2 = \mathbf{0}$  has the trivial solution  $c_1 = c_2 = 0$  as the only solution. Now, applying  $A$  to the left on both sides, we have

$$\begin{aligned} A(c_1\vec{v}_1 + c_2\vec{v}_2) &= \mathbf{0} \\ c_1(A\vec{v}_1) + c_2(A\vec{v}_2) &= \mathbf{0} \end{aligned}$$

$$c_1\lambda_1\vec{v}_1 + c_2\lambda_2\vec{v}_2 = \mathbf{0} \quad (\text{Definition 9.1.1})$$

But multiplying the same equation of  $c_1\vec{v}_1 + c_2\vec{v}_2 = \mathbf{0}$  by  $\lambda_1$  gives  $c_1\lambda_1\vec{v}_1 + c_2\lambda_1\vec{v}_2 = \mathbf{0}$  instead. Subtracting this from above leads to

$$\begin{aligned} (c_1\lambda_1\vec{v}_1 + c_2\lambda_2\vec{v}_2) - (c_1\lambda_1\vec{v}_1 + c_2\lambda_1\vec{v}_2) &= \mathbf{0} - \mathbf{0} \\ c_2(\lambda_2 - \lambda_1)\vec{v}_2 &= \mathbf{0} \end{aligned}$$

Since  $\lambda_1$  and  $\lambda_2$  are distinct,  $\lambda_2 - \lambda_1 \neq 0$ , and  $\vec{v}_2$ , being an eigenvector, cannot be the zero vector, the only possibility is  $c_2 = 0$ . The same argument shows that  $c_1 = 0$ , and we are done.  $\square$

**Properties 9.1.5.** For a square matrix  $A$ ,

1.  $A^T$  shares the same eigenvalues, but for each eigenvalue the eigenvector is not guaranteed to be the same,
2. The eigenvalues for the inverse  $A^{-1}$ , provided that it exists, are the reciprocals of the eigenvalues of  $A$ , but the corresponding eigenvectors are the same.

*Proof.* For the first statement, note that

$$\begin{aligned} \det(A^T - \lambda I) &= \det(A^T - \lambda I^T) \\ &= \det((A - \lambda I)^T) \quad (\text{Properties 2.1.4}) \\ &= \det(A - \lambda I) \quad (\text{Properties 2.3.9}) \end{aligned}$$

So the characteristic polynomials of  $A$  and  $A^T$  coincides and the eigenvalues (if exist) will be the same. For the second statement, multiplying both sides of Definition 9.1.1 by  $A^{-1}$  to obtain

$$\begin{aligned} A\vec{v}_\lambda &= \lambda\vec{v}_\lambda \\ A^{-1}A\vec{v}_\lambda &= I\vec{v}_\lambda = \vec{v}_\lambda = \lambda A^{-1}\vec{v}_\lambda \quad (\text{Definition 2.2.1}) \\ \frac{1}{\lambda}\vec{v}_\lambda &= A^{-1}\vec{v}_\lambda \end{aligned}$$

which explicitly shows that  $A^{-1}$  has  $\frac{1}{\lambda}$  and  $\vec{v}_\lambda$  as its eigenvalue and eigenvector.  $\square$

**Theorem 9.1.6.** A square matrix is singular if and only if it has zero as one of its eigenvalues.

*Proof.* Exercise.<sup>4</sup> Note that this is also equivalent to  $A$  having a non-trivial null space of dimension  $\geq 1$  due to Theorem 9.1.3.  $\square$

### 9.1.3 Eigenspace as an Invariant Subspace

The eigenspace actually belongs to a broader class of subspaces known as the ***invariant subspaces***. Given a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$ , a subspace  $\mathcal{W}$  of  $\mathcal{V}$  is known as ***T-invariant*** if applying  $T$  to any vector  $\vec{w} \in \mathcal{W}$  in the subspace maps it into a vector in the subspace  $\mathcal{W}$  itself.

**Definition 9.1.7** (Invariant Subspace). With respect to a linear endomorphism  $T : \mathcal{V} \rightarrow \mathcal{V}$ , a  $T$ -invariant subspace  $\mathcal{W} \subseteq \mathcal{V}$  is a subspace such that  $T(\vec{w}) \in \mathcal{W}$  for all  $\vec{w} \in \mathcal{W}$ .

The zero subspace and improper subspace ( $\mathcal{V}$  itself) are two trivial invariant subspaces for any  $T$ . A more relevant fact is that any eigenspace of a fixed eigenvalue for is  $T$ -invariant.

**Properties 9.1.8.** For a linear endomorphism  $T : \mathcal{V} \rightarrow \mathcal{V}$ , the eigenspace  $\mathcal{E}_{J_i} \subseteq \mathcal{V}$  associated with a specific eigenvalue  $\lambda_{J_i}$  is a  $T$ -invariant subspace.

*Proof.* Exercise.<sup>5</sup>  $\square$

---

<sup>4</sup>Denote the matrix by  $A$ , if it has an eigenvalue of zero, then by Definition 9.1.2,  $\det(A - 0I) = 0 = \det(A)$ . Properties 2.3.8 then implies  $A$  is singular. The converse is established by the same argument in reverse direction.

<sup>5</sup>For any  $\vec{w} \in \mathcal{E}_{J_i}$ , by Definition 9.1.1,  $T(\vec{w}) = \lambda_{J_i} \vec{w} \in \mathcal{E}_{J_i}$ . ((2) of Theorem 6.1.2)

and the same idea of an invariant subspace is applicable to any  $n \times n$  square matrix  $A$  and the real  $n$ -space  $\mathbb{R}^n$ . Another important observation is that for each of the linear transformations  $T_{J_i} : \mathcal{W}_{J_i} \rightarrow \mathcal{W}_{J_i}$  (when they happen to be endomorphisms) in their direct sum  $T = \bigoplus_{J_i} T_{J_i}$  that is formed according to Definition 8.3.7, its corresponding subspace  $\mathcal{W}_{J_i}$  (as in the vector direct sum  $\mathcal{V} = \bigoplus_{J_i} \mathcal{W}_{J_i}$ ) is ( $T_{J_i}$ - and)  $T$ -invariant.

### 9.1.4 Cayley-Hamilton Theorem

We conclude this section by introducing the **Cayley-Hamilton Theorem**, which states that every square matrix satisfies its own characteristic equation, which means that replacing the  $\lambda$  in the characteristic polynomial by the matrix results in a zero matrix.

**Theorem 9.1.9** (Cayley-Hamilton Theorem). For any  $n \times n$  square matrix  $A$ , denote its characteristic polynomial by

$$p(\lambda) = \det(A - \lambda I) = \sum_{k=0}^n c_k \lambda^k$$

then we have

$$p(A) = \sum_{k=0}^n c_k A^k = [\mathbf{0}]$$

as an  $n \times n$  zero matrix.

One may be tempted to substitute  $\lambda = A$  into  $\det(A - \lambda I)$  to prove the Cayley-Hamilton Theorem. However, since  $\lambda$  (and  $\det(A - \lambda I)$ ) is a scalar but  $A$  (and  $[\mathbf{0}]$ ) is a matrix, it is not a rigorous proof. Correct proofs require advanced knowledge, which is presented in the Appendix. Here we will briefly see how it works.

**Example 9.1.5.** With the matrix

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

verify the Cayley-Hamilton Theorem, and use the Cayley-Hamilton Theorem to evaluate  $A^2 - 7A + 6I$ .

*Solution.* The characteristic polynomial is

$$\begin{aligned} \begin{vmatrix} 1-\lambda & -1 \\ 3 & 5-\lambda \end{vmatrix} &= (1-\lambda)(5-\lambda) - (3)(-1) \\ &= 5 - 6\lambda + \lambda^2 + 3 \\ &= \lambda^2 - 6\lambda + 8 \end{aligned}$$

Replacing all  $\lambda^k$  terms in the characteristic polynomial with  $A^k$  (Notice that the constant term  $c_0$  becomes  $c_0I$ ), we have

$$\begin{aligned} A^2 - 6A + 8I &= \begin{bmatrix} 1 & -1 \\ 3 & 5 \end{bmatrix}^2 - 6 \begin{bmatrix} 1 & -1 \\ 3 & 5 \end{bmatrix} + 8 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -2 & -6 \\ 18 & 22 \end{bmatrix} + \begin{bmatrix} -6 & 6 \\ -18 & -30 \end{bmatrix} + \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

So the Cayley-Hamilton Theorem holds in this case. We can efficiently compute  $A^2 - 7A + 6I$  through

$$\begin{aligned} A^2 - 7A + 6I &= (A^2 - 6A + 8I) - (A^2 - 6A + 8I) \\ &= -A - 2I \\ &= - \begin{bmatrix} 1 & -1 \\ 3 & 5 \end{bmatrix} - 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -3 & 1 \\ -3 & -7 \end{bmatrix} \end{aligned}$$

since  $A^2 - 6A + 8I$  is a zero matrix. □

## 9.2 Diagonalization

### 9.2.1 Mathematical Ideas of Diagonalization

The existence of eigenvectors allows us to carry out *diagonalization* which helps us to simplify many linear algebra problems and derivations. A matrix  $P$  is said to *diagonalize* another matrix  $A$  if the product  $P^{-1}AP$  results in a *diagonal matrix*  $D$ , where all non-zero entries are only found along the main diagonal.

**Definition 9.2.1** (Diagonalization). A square matrix  $A$  is diagonalizable, if there exists some invertible square matrix  $P$ , such that

$$P^{-1}AP = D$$

where  $D$  is a diagonal matrix.

Recalling Properties 7.2.2, we know that the form  $P^{-1}AP$  (or  $P^{-1}[T]P$  for a linear transformation) represents a change of coordinates for the matrix. So the problem of diagonalization is actually asking if there exists a basis (expressed as the columns of  $P$ ) such that the matrix  $A$  in question is diagonal with respect to the corresponding coordinate system, or in other words, if  $A$  is similar to a diagonal matrix. We will show that in fact, the coordinate matrix  $P$  required for diagonalizing  $A$ , is formed by combining all the linearly independent eigenvectors of  $A$  column by column. This is only possible if the amount of distinct eigenvectors is equal to the extent of  $A$ . An equivalent condition is that the characteristic polynomial splits over the scalar ( $\mathbb{R}$  or  $\mathbb{C}$ )<sup>6</sup> used in the vector

---

<sup>6</sup> A (characteristic) polynomial  $p(\lambda)$  of degree  $n$  is said to *split* over  $\mathbb{C}$  ( $\mathbb{R}$ ) if it can be expressed as the product of linear factors only, i.e.

$$p(\lambda) = \prod_{j=1}^n (\lambda_j - \lambda)$$

where  $\lambda_j$  are complex (real) constants. Also, there is a fundamental result which states that every polynomial splits over  $\mathbb{C}$  so the first part of the condition is automatically satisfied for the complex case.

space and for every eigenvalue, its geometric multiplicity is equal to the algebraic multiplicity.<sup>7</sup> So, the matrix in Example 9.1.1 is non-diagonalizable.

**Properties 9.2.2.** An  $n \times n$  square matrix  $A$  can be diagonalized by another matrix  $P$  if and only if  $A$  has  $n$  linearly independent eigenvectors, such that they form a basis for  $\mathbb{R}^n$  (Properties 6.2.7) when the scalar of vector space is taken to be  $\mathbb{R}$  ( $\mathbb{C}^n$  if over  $\mathbb{C}$ ). Then the columns of  $P$  are consisted of those eigenvectors  $\vec{v}_\lambda^{(j)}$ ,  $j = 1, 2, \dots, n$ . The diagonal entries of  $P^{-1}AP = D$ , are subsequently each of the eigenvalues  $\lambda_j$  (counting repeated ones) corresponding to the eigenvector  $\vec{v}_\lambda^{(j)}$  in the same column position of  $P$ .

*Proof.* We will explicitly construct the desired form. Consider two matrix products  $AP$  and  $PD$ , where  $P = [\vec{v}_\lambda^{(1)} | \vec{v}_\lambda^{(2)} | \cdots | \vec{v}_\lambda^{(n)}]$  has the  $n$  linearly independent eigenvectors as its columns. Then

$$\begin{aligned} AP &= A[\vec{v}_\lambda^{(1)} | \vec{v}_\lambda^{(2)} | \cdots | \vec{v}_\lambda^{(n)}] \\ &= [A\vec{v}_\lambda^{(1)} | A\vec{v}_\lambda^{(2)} | \cdots | A\vec{v}_\lambda^{(n)}] \\ &= [\lambda_1\vec{v}_\lambda^{(1)} | \lambda_2\vec{v}_\lambda^{(2)} | \cdots | \lambda_n\vec{v}_\lambda^{(n)}] \end{aligned}$$

where we have used Definition 9.1.1 and the first to second line can be referred to Footnote 10 of Chapter 6. Also

$$\begin{aligned} PD &= [\vec{v}_\lambda^{(1)} | \vec{v}_\lambda^{(2)} | \cdots | \vec{v}_\lambda^{(n)}] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \\ &= [\lambda_1\vec{v}_\lambda^{(1)} | \lambda_2\vec{v}_\lambda^{(2)} | \cdots | \lambda_n\vec{v}_\lambda^{(n)}] \end{aligned}$$

---

<sup>7</sup>Note that the sum of algebraic multiplicities for all eigenvalues are less than or equal to (when the characteristic polynomial splits) the degree of the characteristic polynomial by observing Definition 9.1.2 and Theorem 9.1.3, which is the same as the extent of the matrix. For the amount of linearly independent eigenvectors to be equal to that, the only possibility is that any geometric multiplicity being strictly equal to the algebraic multiplicity while  $p(\lambda)$  splits. (Properties 9.1.4 ensures linear independence over eigenspaces of different eigenvalues.)

So  $AP = PD$ . Notice that  $P$  is invertible by (e) to (a) of Theorem 6.1.10, since the columns of  $P$  are made up of linearly independent eigenvectors, and thus  $P^{-1}AP = D$ .  $\square$

### 9.2.2 Diagonalization for Real Eigenvalues

Before getting into the properties of diagonalization, let's see how diagonalization is carried out first. For a diagonalizable matrix with real eigenvalues, diagonalization is straight-forward by the use of Properties 9.2.2. Below shows a worked example.

**Example 9.2.1.** Diagonalize the matrix

$$A = \begin{bmatrix} 1 & -2 & 1 \\ -\frac{2}{7} & 1 & -1 \\ -\frac{4}{7} & -2 & 0 \end{bmatrix}$$

*Solution.* The characteristic polynomial can be checked to be

$$\begin{aligned} p(\lambda) &= \det(A - \lambda I) \\ &= \begin{vmatrix} 1 - \lambda & -2 & 1 \\ -\frac{2}{7} & 1 - \lambda & -1 \\ -\frac{4}{7} & -2 & -\lambda \end{vmatrix} \\ &= -\lambda^3 + 2\lambda^2 + \lambda - 2 = (-1 - \lambda)(1 - \lambda)(2 - \lambda) \end{aligned}$$

Hence the eigenvalues are  $\lambda = -1, 1, 2$ . For  $\lambda = -1$ , the eigenvector is obtained from

$$\begin{array}{ccc|c} 1 - (-1) & -2 & 1 & 0 \\ -\frac{2}{7} & 1 - (-1) & -1 & 0 \\ -\frac{4}{7} & -2 & -(-1) & 0 \end{array} = \begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ -\frac{2}{7} & 2 & -1 & 0 \\ -\frac{4}{7} & -2 & 1 & 0 \end{array} \quad \begin{array}{l} \frac{1}{2}R_1 \rightarrow R_1 \\ R_2 + \frac{2}{7}R_1 \rightarrow R_2 \\ R_3 + \frac{4}{7}R_1 \rightarrow R_3 \end{array}$$

$$\begin{aligned}
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & -\frac{18}{7} & \frac{9}{7} & 0 \end{array} \right] \quad \frac{7}{12}R_2 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + \frac{18}{7}R_2 \rightarrow R_3 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_1 + R_2 \rightarrow R_1
 \end{aligned}$$

The third column is non-pivotal and can be set as the free variable, which eventually leads to an eigenvector of  $(0, 1, 2)^T$ . We leave to the readers to check that the eigenvectors for the remaining eigenvalues  $\lambda = 1, 2$  are  $(-7, 1, 2)^T, (7, -3, 1)^T$  respectively. Concatenating these three eigenvectors column by column yields

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

The diagonalization is then done according to Properties 9.2.2 as

$$\begin{aligned}
 D &= P^{-1}AP \\
 &= \begin{bmatrix} 0 & -7 & 7 \\ 1 & 1 & -3 \\ 2 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -2 & 1 \\ -\frac{2}{7} & 1 & -1 \\ -\frac{4}{7} & -2 & 0 \end{bmatrix} \begin{bmatrix} 0 & -7 & 7 \\ 1 & 1 & -3 \\ 2 & 2 & 1 \end{bmatrix} \\
 &= \begin{bmatrix} \frac{1}{7} & \frac{3}{7} & \frac{2}{7} \\ -\frac{1}{7} & -\frac{2}{7} & \frac{1}{7} \\ 0 & -\frac{2}{7} & \frac{1}{7} \end{bmatrix} \begin{bmatrix} 0 & -7 & 14 \\ -1 & 1 & -6 \\ -2 & 2 & 2 \end{bmatrix} \\
 &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}
 \end{aligned}$$

Note that we can shuffle the ordering of the eigenvalues and eigenvectors.<sup>8</sup> We say that diagonalization is *unique up to a permutation*.  $\square$

Short Exercise: Reverse the diagonalization to recover the original matrix.<sup>9</sup>

**Example 9.2.2.** Diagonalize the matrix

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

in Example 9.1.3.

*Solution.* From Example 9.1.3, we know that the characteristic polynomial splits over  $\mathbb{R}$  and for the eigenvalues of  $\lambda = 2, 3$ , their geometric multiplicities are equal to algebraic multiplicities (2 and 1). Therefore, its diagonalization is possible by Properties 9.2.2. Specifically, the eigenvectors for  $\lambda = 2$  are  $(1, 1, 0)^T, (-1, 0, 1)^T$  and that for  $\lambda = 3$  is  $(-1, -1, 1)^T$ . Thus we have

$$\begin{aligned} P &= \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{and} \\ D &= P^{-1}AP \\ &= \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & -1 \\ -1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \end{aligned}$$

---

<sup>8</sup>For example,

$$\begin{bmatrix} 0 & 7 & -7 \\ 1 & -3 & 1 \\ 2 & 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -2 & 1 \\ -\frac{2}{7} & 1 & -1 \\ -\frac{4}{7} & -2 & 0 \end{bmatrix} \begin{bmatrix} 0 & 7 & -7 \\ 1 & -3 & 1 \\ 2 & 1 & 2 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is another valid outcome where we have interchanged the second and third eigenvector.

<sup>9</sup>Simply compute  $PDP^{-1}$  which should be equal to  $A$ .

$$\begin{aligned}
 &= \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & -3 \\ 2 & 0 & -3 \\ 0 & 2 & 3 \end{bmatrix} \\
 &= \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}
 \end{aligned}$$

The first and second eigenvector can be interchanged without modifying the form of diagonalized matrix as they correspond to the same eigenvalue of  $\lambda = 2$ .  $\square$

### 9.2.3 Properties of Diagonalization

Having seen the actual procedure of diagonalization, we are ready to derive its properties. First, it is instructive to view diagonalization from the perspective of a vector/matrix direct sum. The condition in Properties 9.2.2 requires that there are  $n$  one-dimensional subspaces  $\mathcal{E}_j$  generated by each of the  $n$  linearly independent eigenvectors  $\vec{v}_\lambda^{(j)}$  for the linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  (can be treated as an  $n \times n$  square matrix  $A$ ) where  $\mathcal{V}$  is an  $n$ -dimensional vector space, and thus they form a direct sum  $\bigoplus_{j=1}^n \mathcal{E}_j = \mathcal{V}$  (or  $\mathbb{R}^n/\mathbb{C}^n$ ) following Definition 6.2.9 and Properties 6.2.7. In Properties 9.1.8, we have already observed that the eigenspace  $\mathcal{E}_{J_i} = \bigoplus_{\lambda_j=\lambda_{J_i}} \mathcal{E}_j$  of each distinct eigenvalue  $\lambda_{J_i}$ <sup>10</sup> is an invariant subspace under  $T$  and the same can be said for the individual one-dimensional  $\mathcal{E}_j$  themselves. In particular, the direction of any vector in each of the eigenspaces remains unchanged after  $T$  is applied. Therefore, by Definition 8.3.7,  $T$  can be written as a matrix direct sum  $T = \bigoplus_{j=1}^n T_j$  where  $T_j(\vec{v}_\lambda^{(j)}) : \mathcal{E}_j \rightarrow \mathcal{E}_j = T|_{\mathcal{E}_j}(\vec{v}_\lambda^{(j)}) = T(\vec{v}_\lambda^{(j)}) = \lambda_j \vec{v}_\lambda^{(j)} = \lambda_j(\text{id}(\vec{v}_\lambda^{(j)}))$ . In matrix representation it becomes  $A = \bigoplus_{j=1}^n [\lambda_j]$  where each of the  $[\lambda_j]$  is a singleton ( $1 \times 1$ ) block. If we consider repeated eigenvalues and their full eigenspaces  $\mathcal{E}_{J_i} = \bigoplus_{\lambda_j=\lambda_{J_i}} \mathcal{E}_j$  as an entirety, then we have  $T = \bigoplus_{\text{all distinct } \lambda_{J_i}} T_{J_i}$  with simply  $T_{J_i} = \lambda_{J_i}(\text{id}[\ ])$  and the matrix form is  $A = \bigoplus_{\text{all distinct } \lambda_{J_i}} \lambda_{J_i} I_{k_{J_i}}$  where  $k_{J_i}$  is the eigenvalue repetition count of  $\lambda_{J_i}$  in the characteristic polynomial (here

---

<sup>10</sup> $J_i$  refers to the set of indices those are associated to a common eigenvalue, denoted by  $\lambda_{J_i}$ .

the algebraic multiplicity is equal to the geometric counterpart as required by Properties 9.2.2). The essence of diagonalization is hence to find a basis for  $\mathcal{V}$  as a vector direct sum such that  $T : \mathcal{V} \rightarrow \mathcal{V}$  becomes a matrix direct sum that is composed by some scalar multiples of identity matrices/identity mappings with respect to this vector direct sum basis.

On the other hand, the original matrix  $A$  and its diagonalized form  $D = P^{-1}AP$  share some similarities sometimes called **invariants**. More generally, these invariants hold for any pair of similar matrices (introduced following Properties 7.2.2).

**Properties 9.2.3** (Invariants for Similar Matrices). Two similar square matrices  $A$  and  $A'$  have the

1. same determinant,
2. same characteristic equation,
3. same eigenvalues,
4. same eigenspace dimension for any fixed eigenvalue (but the eigenvectors are probably different), and
5. same trace.

*Proof.* We will briefly prove (1) here. Similar matrices are related by  $A' = P^{-1}AP$  for any invertible matrix  $P$ . As a result,  $\det(A') = \det(P^{-1}AP) = \det(P^{-1})\det(A)\det(P) = \frac{1}{\det(P)}\det(A)\det(P) = \det(A)$  due to Properties 2.3.9. ( $\det(P) \neq 0$  by Properties 2.3.8.)  $\square$

Since the diagonalized form of a matrix is clearly similar to the original one, those properties hold for them as well.

Short Exercise: Prove (2) and (3) of the above properties.<sup>11</sup>

---

<sup>11</sup> $\det(A' - \lambda I) = \det(P^{-1}AP - \lambda I) = \det(P^{-1}AP - \lambda P^{-1}P) = \det(P^{-1}AP - \lambda P^{-1}IP) = \det(P^{-1}(A - \lambda I)P) = \det(P^{-1})\det(A - \lambda I)\det(P) = \det(A - \lambda I)$  (Definition 2.2.1, Properties 2.1.2 and 2.3.9). (3) follows from (2) immediately.

With all of these, we can finally now prove Theorem 9.1.3. Denote the geometric multiplicity of  $A$  by  $n_{J_i}$  for the eigenvalue  $\lambda_{J_i}$ . Then the  $n_{J_i}$  linearly independent eigenvectors  $\vec{v}_{\lambda_{J_i}}^{(1)}, \vec{v}_{\lambda_{J_i}}^{(2)}, \dots, \vec{v}_{\lambda_{J_i}}^{(n_{J_i})}$  accordingly lead to an  $n_{J_i}$ -dimensional eigenspace of  $\mathcal{E}_{J_i} = \bigoplus_{\lambda_j=\lambda_{J_i}} \mathcal{E}_j$  as just described, that is, an invariant subspace under the multiplication by  $A$  to its left. We can always enlarge (the basis of) this eigenspace  $\mathcal{E}_{J_i}$  to reproduce the entire  $\mathbb{R}^n$  (or  $\mathbb{C}^n$ ) as suggested by part (c) of Properties 6.2.7. If we do a change of coordinates on  $A$  using this new basis, then it will becomes (Properties 7.2.2 again)

$$A' = P^{-1}AP$$

where

$$P = \left[ \vec{v}_{\lambda_{J_i}}^{(1)} | \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | \text{other vectors used to fill the basis for } \mathbb{R}^n \right]$$

Comparing with our discussion just before and according to Definition 7.1.2, we know that the new matrix  $A'$  will take the block form of

$$A' = \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix}$$

<sup>12</sup> where the top left represents scaling of the eigenvectors (now as a part of the new basis) by the eigenvalue, and the \* parts can be anything (notice the zero

---

<sup>12</sup>We can check if the first  $n_j$  columns are equal for  $PA'$  and  $AP$ , which is very similar to the derivation in Properties 9.2.2.

$$\begin{aligned} PA' &= \left[ \vec{v}_{\lambda_{J_i}}^{(1)} | \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | * \right] \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix} \\ &= \left[ \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(1)} | \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | * \right] \end{aligned}$$

and

$$\begin{aligned} AP &= A \left[ \vec{v}_{\lambda_{J_i}}^{(1)} | \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | * \right] \\ &= \left[ A \vec{v}_{\lambda_{J_i}}^{(1)} | A \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | A \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | * \right] \\ &= \left[ \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(1)} | \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(2)} | \dots | \lambda_{J_i} \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | * \right] \end{aligned}$$

the last step is by the definition of eigenvectors (Definition 9.1.1).

block at the bottom left). Its characteristic polynomial is then

$$\det(A' - \lambda I) = \begin{vmatrix} (\lambda_{J_i} - \lambda)I_{n_{J_i}} & *_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} - \lambda I_{(n-n_{J_i})} \end{vmatrix}$$

By repeated cofactor expansion along the first  $n_{J_i}$  columns, we have

$$p(\lambda) = (\lambda_{J_i} - \lambda)^{n_{J_i}} \det(*_{(n-n_{J_i}) \times (n-n_{J_i})} - \lambda I_{(n-n_{J_i})})$$

which shows that the characteristic polynomial have  $\lambda_{J_i}$  as its roots occurring for at least  $n_{J_i}$  times (the determinant after may or may not have any  $(\lambda_{J_i} - \lambda)$  factor). By Properties 9.2.3, the original matrix  $A$  will share the same characteristic polynomial and hence the algebraic multiplicity for  $\lambda_{J_i}$  (the times where the  $(\lambda_{J_i} - \lambda)$  factor appears in the characteristic polynomial) concerning  $A$  will be greater than or equal to its given geometric multiplicity  $n_{J_i}$ .

## 9.2.4 Diagonalization for Complex Eigenvalues

It is not uncommon for a real square matrix  $A$  to have complex eigenvalues (as well as complex eigenvectors). In such cases, to perform diagonalization, one possible approach is following exactly what we have done in the Section 9.2.2 to produce  $D = P^{-1}AP$  which contains the complex eigenvalues along its main diagonal (and are all zeros elsewhere) if we choose to work over  $\mathbb{C}$ . This can be useful sometimes, but the downsides are that  $D$  (as well as the  $P$  matrix that is made up of the eigenvectors as its columns) are comprised of complex entries, despite  $A$  being a real matrix. There is, indeed, another method uses the property of complex numbers introduced in the last chapter, that circumvents the appearance of complex numbers in a modified diagonalized form and allows us to stay in the world of  $\mathbb{R}$ . But first of all, we need to introduce a basic theorem about complex roots of an equation.

**Theorem 9.2.4.** For a *real* polynomial equation of order  $n$

$$p(x) = \sum_{k=0}^n c_k x^k$$

such that the coefficients  $c_k$  are all real numbers. If  $z_0 = a + bi$  is a complex root so that  $p(z_0) = \sum_{k=0}^n c_k z_0^k = 0$ , where  $a$  and  $b$  are real constants, then its conjugate  $\bar{z}_0 = a - bi$  is also another root for the equation.

*Proof.* By Properties 8.1.8, if we take the complex conjugate on both sides of the real polynomial equation with  $x = z_0$ , then

$$\begin{aligned} \overline{p(z_0)} &= \overline{\sum_{k=0}^n c_k z_0^k} = \bar{0} \\ p(\bar{z}_0) &= \sum_{k=0}^n \overline{c_k} \overline{z_0^k} = 0 \\ &= \sum_{k=0}^n c_k \bar{z}_0^k = 0 \end{aligned}$$

$\overline{c_k} = c_k$  as they are real coefficients. □

Since the characteristic polynomial must be a real polynomial for a real matrix  $A$ , by the theorem we have just proved, we know that its complex roots and hence the complex eigenvalues of  $A$  always come in a conjugate pair. For a pair of complex eigenvalues, their eigenvectors are also the conjugate of each other.

**Properties 9.2.5.** If  $\lambda_0$  is a complex eigenvalue for a real matrix  $A$  with an eigenvector of  $\vec{v}_{\lambda_0}$ , then  $A$  also has  $\bar{\lambda}_0$  and  $\bar{v}_{\lambda_0}$  as another eigenvalue and eigenvector.

*Proof.* By Definition 9.1.1,

$$A\vec{v}_{\lambda_0} = \lambda_0 \vec{v}_{\lambda_0}$$

Taking complex conjugate on both sides, we have

$$\begin{aligned}\overline{A\vec{v}_{\lambda_0}} &= \overline{\lambda_0 \vec{v}_{\lambda_0}} \\ A\overline{\vec{v}_{\lambda_0}} &= \overline{\lambda_0} \overline{\vec{v}_{\lambda_0}}\end{aligned}$$

with the use of Properties 8.1.8, and noting that  $\overline{A} = A$  as  $A$  is a real matrix.  $\square$

Now as we know that complex eigenvalues and eigenvectors always appear as a pair of conjugates, and conjugates share the same real and imaginary parts except a sign difference for the latter, we are encouraged to use them like two different eigenvalues and eigenvectors for making up. The following properties show that it is possible to take advantage of this and derive a new diagonalization procedure with some tweaks.

**Properties 9.2.6.** For a real square matrix  $A$ , the procedure in Properties 9.2.2 can be adapted for a complex eigenvalue  $\lambda_0 = \operatorname{Re}\{\lambda_0\} + i \operatorname{Im}\{\lambda_0\}$ , the corresponding eigenvector  $\vec{v}_{\lambda_0} = \operatorname{Re}\{\vec{v}_{\lambda_0}\} + i \operatorname{Im}\{\vec{v}_{\lambda_0}\}$ , as well as their complex conjugates, by replacing the corresponding columns in

$$P = [\cdots | \vec{v}_{\lambda_0} | \overline{\vec{v}_{\lambda_0}} | \cdots] \quad D = \begin{bmatrix} \ddots & 0 & 0 \\ & \lambda_0 & 0 \\ 0 & \bar{\lambda}_0 & \\ 0 & 0 & \ddots \end{bmatrix}$$

by

$$P = [\cdots | \operatorname{Re}\{\vec{v}_{\lambda_0}\} | \operatorname{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \quad D = \begin{bmatrix} \ddots & 0 & 0 \\ & \operatorname{Re}\{\lambda_0\} & \operatorname{Im}\{\lambda_0\} \\ -\operatorname{Im}\{\lambda_0\} & \operatorname{Re}\{\lambda_0\} & \\ 0 & 0 & \ddots \end{bmatrix}$$

*Proof.* As in the proof for Properties 9.2.2, we set to prove that  $AP = PD$  for the two columns concerned. To make it easier to read, denote  $\operatorname{Re}\{\lambda_0\} = a$  and

$\text{Im}\{\lambda_0\} = b$ . It is not hard to see that

$$PD = [\cdots | \text{Re}\{\vec{v}_{\lambda_0}\} | \text{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \begin{bmatrix} \ddots & 0 & 0 \\ & a & b \\ & -b & a \\ 0 & 0 & \ddots \end{bmatrix} = [\cdots | a \text{Re}\{\vec{v}_{\lambda_0}\} - b \text{Im}\{\vec{v}_{\lambda_0}\} | b \text{Re}\{\lambda_0\} + a \text{Im}\{\vec{v}_{\lambda_0}\} | \cdots]$$

(Properties 6.1.4) On the other hand, notice that the relation between the complex eigenvalue and eigenvector via Definition 9.1.1 can be expanded as

$$A\vec{v}_{\lambda_0} = \lambda_0\vec{v}_{\lambda_0}$$

$$A(\text{Re}\{\vec{v}_{\lambda_0}\} + i\text{Im}\{\vec{v}_{\lambda_0}\}) = (a + bi)(\text{Re}\{\vec{v}_{\lambda_0}\} + i\text{Im}\{\vec{v}_{\lambda_0}\})$$

$$A \text{Re}\{\vec{v}_{\lambda_0}\} + iA \text{Im}\{\vec{v}_{\lambda_0}\} = (a \text{Re}\{\vec{v}_{\lambda_0}\} - b \text{Im}\{\vec{v}_{\lambda_0}\}) + i(b \text{Re}\{\vec{v}_{\lambda_0}\} + a \text{Im}\{\vec{v}_{\lambda_0}\})$$

where  $\text{Re}\{\vec{v}_{\lambda_0}\}$  and  $\text{Im}\{\vec{v}_{\lambda_0}\}$  themselves are real vectors (plus  $A$  is a real matrix). By equating the real and imaginary parts, we have

$$A \text{Re}\{\vec{v}_{\lambda_0}\} = a \text{Re}\{\vec{v}_{\lambda_0}\} - b \text{Im}\{\vec{v}_{\lambda_0}\} \quad (9.1)$$

$$A \text{Im}\{\vec{v}_{\lambda_0}\} = b \text{Re}\{\vec{v}_{\lambda_0}\} + a \text{Im}\{\vec{v}_{\lambda_0}\} \quad (9.2)$$

Hence

$$\begin{aligned} AP &= A[\cdots | \text{Re}\{\vec{v}_{\lambda_0}\} | \text{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \\ &= [\cdots | A \text{Re}\{\vec{v}_{\lambda_0}\} | A \text{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \\ &= [\cdots | a \text{Re}\{\vec{v}_{\lambda_0}\} - b \text{Im}\{\vec{v}_{\lambda_0}\} | b \text{Re}\{\lambda_0\} + a \text{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \\ &= PD \end{aligned}$$

The only caveat is to prove that  $P$  is invertible, or equivalently  $\text{Re}\{\vec{v}_{\lambda_0}\}$  and  $\text{Im}\{\vec{v}_{\lambda_0}\}$  are linearly independent (assuming there is no problem with other columns). We can prove that by assuming they are linearly dependent, so  $\text{Re}\{\vec{v}_{\lambda_0}\} = k \text{Im}\{\vec{v}_{\lambda_0}\}$  for some real constant  $k$ , and plugging this into the expressions of  $A \text{Re}\{\vec{v}_{\lambda_0}\}$  and  $A \text{Im}\{\vec{v}_{\lambda_0}\}$  to derive a contradiction.<sup>13</sup> □

<sup>13</sup> Substitute  $\text{Re}\{\vec{v}_{\lambda_0}\} = k \text{Im}\{\vec{v}_{\lambda_0}\}$  into Equation 9.1

$$A \text{Re}\{\vec{v}_{\lambda_0}\} = a \text{Re}\{\vec{v}_{\lambda_0}\} - b \text{Im}\{\vec{v}_{\lambda_0}\}$$

The block in the form

$$\begin{bmatrix} \ddots & 0 & 0 \\ a & b & \\ -b & a & \\ 0 & 0 & \ddots \end{bmatrix}$$

generally represents a rotation by a degree of  $\theta = \arctan(b/a)$ . (see Example 9.1.4) It means that a complex eigenvalue entails a rotation. This will be discussed more thoroughly in the next chapter.

**Example 9.2.3.** Convert

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

into the modified real "diagonalized" form  $D = P^{-1}AP$  as in Properties 9.2.6.

*Solution.* The characteristic equation is

$$\begin{vmatrix} 1 - \lambda & 0 & 1 \\ 0 & 1 - \lambda & 1 \\ 0 & -2 & 1 - \lambda \end{vmatrix} = (1 - \lambda)((1 - \lambda)^2 + (1)(-2))$$

$$Ak \operatorname{Im}\{\vec{v}_{\lambda_0}\} = ak \operatorname{Im}\{\vec{v}_{\lambda_0}\} - b \operatorname{Im}\{\vec{v}_{\lambda_0}\} = (ak - b) \operatorname{Im}\{\vec{v}_{\lambda_0}\}$$

and multiplying Equation 9.2 by  $k$

$$\begin{aligned} A \operatorname{Im}\{\vec{v}_{\lambda_0}\} &= b \operatorname{Re}\{\vec{v}_{\lambda_0}\} + a \operatorname{Im}\{\vec{v}_{\lambda_0}\} \\ Ak \operatorname{Im}\{\vec{v}_{\lambda_0}\} &= bk(k \operatorname{Im}\{\vec{v}_{\lambda_0}\}) + ak \operatorname{Im}\{\vec{v}_{\lambda_0}\} = (bk^2 + ak) \operatorname{Im}\{\vec{v}_{\lambda_0}\} \end{aligned}$$

Comparing these two equations ( $\operatorname{Re}\{\vec{v}_{\lambda_0}\}$  and  $\operatorname{Im}\{\vec{v}_{\lambda_0}\}$  are non-zero, real vectors) gives

$$bk^2 + ak = ak - b \implies bk^2 = -b \implies k = \pm i$$

which contradicts the assumption where  $k$  has to be real.

$$= (1 - \lambda)(\lambda^2 - 2\lambda + 3) = 0$$

The roots are  $\lambda = 1$  and

$$\lambda = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(3)}}{2}$$

$$= 1 \pm \sqrt{2}i$$

It is very easy to see that the eigenvector for  $\lambda = 1$  is  $(1, 0, 0)^T$ . For the remaining two complex eigenvalues ( $\lambda = 1 \pm \sqrt{2}i$ ) and eigenvectors, since by Properties 9.2.5 they occur in a conjugate pair, we only need to compute one of them. We can find the eigenvector for  $\lambda = 1 - \sqrt{2}i$ , by solving the system

$$\left[ \begin{array}{ccc|c} \sqrt{2}l & 0 & 1 & 0 \\ 0 & \sqrt{2}l & 1 & 0 \\ 0 & -2 & \sqrt{2}l & 0 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & -2 & \sqrt{2}l & 0 \end{array} \right] \quad -\frac{1}{\sqrt{2}}lR_1 \rightarrow R_1$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad -\frac{1}{\sqrt{2}}lR_2 \rightarrow R_2$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}}l & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 + 2R_1 \rightarrow R_3$$

So the eigenvector for  $\lambda = 1 - \sqrt{2}\iota$  is  $(\iota, \iota, \sqrt{2})^T$ , and the eigenvector for  $\lambda = 1 + \sqrt{2}\iota$  is  $(-\iota, -\iota, \sqrt{2})^T$ . Applying Properties 9.2.6, with  $a = 1$ ,  $b = \sqrt{2}$ ,  $\text{Re}\{\vec{v}_{\lambda_0}\} = (0, 0, \sqrt{2})^T$ ,  $\text{Im}\{\vec{v}_{\lambda_0}\} = (-1, -1, 0)^T$ , we have

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

□

Short Exercise: Repeat the diagonalization but keep complex entries without the modification.<sup>14</sup>

14

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 + \sqrt{2}\iota & 0 \\ 0 & 0 & 1 - \sqrt{2}\iota \end{bmatrix} = \begin{bmatrix} 1 & -\iota & \iota \\ 0 & -\iota & \iota \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\iota & \iota \\ 0 & -\iota & \iota \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}$$

## 9.3 System of Ordinary Differential Equations

**Ordinary Differential Equations (ODEs)** arise frequently in the areas of Physics and Earth Science. They involve the derivatives of a dependent variable (that represents the state of the system, e.g. temperature, velocity) with respect to one independent variable only (usually the displacement  $x$  or time  $t$ ). The simplest class of ODEs is *first-order, linear, homogeneous* described as follows.

**Definition 9.3.1.** A *first-order, linear ODE* is a differential equation written in the form of

$$\frac{dy}{dx} + P(x)y = Q(x)$$

where  $x$  denotes the independent variable,  $y(x)$  is the dependent variable,  $P(x)$  and  $Q(x)$  is some function of  $x$ .

*First-order* means that the highest order of derivatives involved in the equation is the first derivative only, such that there are  $y$  and  $\frac{dy}{dx}$ , but no  $\frac{d^2y}{dx^2}$  or  $\frac{d^3y}{dx^3}$ . *Linearity* means that there are no product terms between the dependent variable or its derivatives themselves, e.g.  $y\frac{dy}{dx}, y^2, (\frac{dy}{dx})^2$ . An example will be deriving terminal speed of a falling hydrometeor under gravity:

$$\frac{dw}{dt} - \alpha w = g$$

where  $w$  is the downward velocity as the dependent variable and the independent variable is now the time  $t$ ,  $Q(t) = g$  is the acceleration due to gravity, and  $P(t) = -\alpha$  represents a linear air resistance that is proportional to the downward speed. Finally, if it is of *constant coefficients*, and *homogeneous* further, then it implies that  $P(x)$  is a constant and  $Q(x) = 0$  respectively.

**Properties 9.3.2.** For a first-order, linear, constant coefficients, homogeneous

ODE in the form of

$$\frac{dy}{dx} = \beta y$$

where  $\beta$  is a constant, the general solution is

$$y(x) = ce^{\beta x}$$

with  $c$  as an integration constant to be determined.

A direct substitution can verify the solution given above.<sup>15</sup> A famous example will be the first-order decay of radioactive isotopes whose concentration is  $N$ :

$$\frac{dN}{dt} = -kN$$

where  $\beta = -k$  is the decay constant.

However, for many real-life problems, there are multiple ODEs, each of which is coupled to others in the system. This frequently happens in Earth System Science where we have to consider the interactions between different variables, for instance, mass flows and chemical reactions between the reservoirs of various substances. This naturally leads to a **system of ODEs**. Again the simplest case would be all of the ODEs being first-order, linear, having constant coefficients, and homogeneous.

**Definition 9.3.3.** A first-order, linear system of  $n$  ODEs that are of constant coefficients, and homogeneous takes the form of

$$\begin{cases} dy_1/dx &= \beta_1^{(1)} y_1 + \beta_2^{(1)} y_2 + \beta_3^{(1)} y_3 + \cdots + \beta_n^{(1)} y_n \\ dy_2/dx &= \beta_1^{(2)} y_1 + \beta_2^{(2)} y_2 + \beta_3^{(2)} y_3 + \cdots + \beta_n^{(2)} y_n \\ dy_3/dx &= \beta_1^{(3)} y_1 + \beta_2^{(3)} y_2 + \beta_3^{(3)} y_3 + \cdots + \beta_n^{(3)} y_n \\ \vdots &= \vdots \\ dy_n/dx &= \beta_1^{(n)} y_1 + \beta_2^{(n)} y_2 + \beta_3^{(n)} y_3 + \cdots + \beta_n^{(n)} y_n \end{cases}$$

---

<sup>15</sup>L.H.S. =  $\frac{d}{dx}(ce^{\beta x}) = \beta ce^{\beta x}$  = R.H.S.

where  $\beta_j^{(i)}$  are the constant coefficients, or more compactly

$$\frac{dy_i}{dx} = \beta_1^{(i)}y_1 + \beta_2^{(i)}y_2 + \beta_3^{(i)}y_3 + \cdots + \beta_n^{(i)}y_n = \sum_{j=1}^n \beta_j^{(i)}y_j$$

or in matrix notation

$$\mathbf{y}' = A\mathbf{y}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{y}' = \begin{bmatrix} dy_1/dx \\ dy_2/dx \\ dy_3/dx \\ \vdots \\ dy_n/dx \end{bmatrix}$$

the superscript ' represents differentiation, and

$$A = \begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \beta_3^{(1)} & \cdots & \beta_n^{(1)} \\ \beta_1^{(2)} & \beta_2^{(2)} & \beta_3^{(2)} & & \beta_n^{(2)} \\ \beta_1^{(3)} & \beta_2^{(3)} & \beta_3^{(3)} & & \beta_n^{(3)} \\ \vdots & & & \ddots & \vdots \\ \beta_1^{(n)} & \beta_2^{(n)} & \beta_3^{(n)} & \cdots & \beta_n^{(n)} \end{bmatrix}$$

is an  $n \times n$  matrix holding the coefficients,  $A_{ij} = \beta_j^{(i)}$ .

An example is the system

$$\begin{cases} dy_1/dx = 3y_1 - y_2 \\ dy_2/dx = 2y_1 \end{cases}$$

which can be rewritten into

$$\mathbf{y}' = \begin{bmatrix} 3 & -1 \\ 2 & 0 \end{bmatrix} \mathbf{y}$$

Since each ODE concerning  $dy_i/dx$  for a fixed  $i$  now involves multiple dependent variables  $y_j$  at R.H.S., where  $j = 1, 2, 3, \dots$ , we cannot directly use the result from Properties 9.3.2 to derive the solution to the system, unless we can find a way to transform the system so that each equation involves a single dependent variable only. Notice that if we make a change of variables (Section 7.2.1) such that  $\mathbf{y} = P\mathbf{z}$  and hence  $\mathbf{y}' = P\mathbf{z}'$ , then the system  $\mathbf{y}' = A\mathbf{y}$  becomes

$$\begin{aligned} P\mathbf{z}' &= AP\mathbf{z} \\ \mathbf{z}' &= (P^{-1}AP)\mathbf{z} \end{aligned}$$

assuming  $P$  is invertible. The term  $P^{-1}AP$ , which is a matrix similar to  $A$ , immediately tells us a hint as it resembles a diagonalized form as described in Properties 9.2.2,  $P^{-1}AP = D$ . If diagonalization is indeed possible, the system will then become  $\mathbf{z}' = D\mathbf{z}$ . When written out, the transformed ODEs are

$$\begin{cases} dz_1/dx &= D_{11}z_1 \\ dz_2/dx &= D_{22}z_2 \\ dz_3/dx &= D_{33}z_3 \\ \vdots &= \vdots \\ dz_n/dx &= D_{nn}z_n \end{cases}$$

which are all solvable in their own by Properties 9.3.2 as each of them only involves a single dependent variable. Subsequently, we have the following conclusion.

**Properties 9.3.4.** For a system of first-order ODEs in the form of

$$\mathbf{y}' = A\mathbf{y}$$

where  $A$  is an  $n \times n$  square matrix with constant entries, if  $A$  is diagonalizable, then it can be solved by making the change of variables

$$\mathbf{y} = P\mathbf{z}$$

where  $P$  is the invertible matrix formed by combining all the  $n$  eigenvectors of  $A$  in columns as suggested by Properties 9.2.2. In this way, the system becomes

diagonalized

$$\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z}$$

and each of the  $z_i$  in  $\mathbf{z}$ ,  $i = 1, 2, 3, \dots, n$  can be solved separately. After working out  $\mathbf{z}$  as a whole, we can recover the required solution in terms of the original variables by computing  $\mathbf{y} = P\mathbf{z}$ .

**Example 9.3.1.** Solve the following system of first-order ODEs.

$$\begin{cases} dy_1/dx = 8y_1 - 12y_2 + 14y_3 \\ dy_2/dx = -2y_1 + 4y_2 - 4y_3 \\ dy_3/dx = -3y_1 + 6y_2 - 5y_3 \end{cases}$$

where the initial conditions at  $x = 0$  are  $y_1(0) = 3$ ,  $y_2(0) = 1$ ,  $y_3(0) = 5$ .

*Solution.* The system written in matrix notation is

$$\begin{bmatrix} dy_1/dx \\ dy_2/dx \\ dy_3/dx \end{bmatrix} = \begin{bmatrix} 8 & -12 & 14 \\ -2 & 4 & -4 \\ -3 & 6 & -5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

We leave to the readers to check that the eigenvectors for the coefficient matrix are  $(-2, 0, 1)^T$ ,  $(-5, 1, 3)^T$ ,  $(-4, 1, 2)^T$  corresponding to the eigenvalues of  $\lambda = 1, 2, 4$  respectively. Hence by Properties 9.3.4, we can make the change of variables

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

where the transformation matrix is formed by the three eigenvectors arranged in columns. The system then becomes

$$\begin{bmatrix} dz_1/dx \\ dz_2/dx \\ dz_3/dx \end{bmatrix} = \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 8 & -12 & 14 \\ -2 & 4 & -4 \\ -3 & 6 & -5 \end{bmatrix} \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

The solution for each of the components is  $z_1 = c_1 e^x$ ,  $z_2 = c_2 e^{2x}$ ,  $z_3 = c_3 e^{4x}$  according to Properties 9.3.2. The integration constants  $c_i$  can be determined by the initial conditions provided. Substituting them into the relation for change of variables gives

$$\begin{aligned} \begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} &= \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} z_1(0) \\ z_2(0) \\ z_3(0) \end{bmatrix} \\ \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} &= \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \end{aligned}$$

It is not hard to obtain  $c_1 = -10$ ,  $c_2 = 13$ ,  $c_3 = -12$ , and hence  $z_1 = -10e^x$ ,  $z_2 = 13e^{2x}$ ,  $z_3 = -12e^{4x}$ . So the full solution for  $y_i$  is

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} -10e^x \\ 13e^{2x} \\ -12e^{4x} \end{bmatrix} \\ \implies \begin{cases} y_1 = 20e^x - 65e^{2x} + 48e^{4x} \\ y_2 = 13e^{2x} - 12e^{4x} \\ y_3 = -10e^x + 39e^{2x} - 24e^{4x} \end{cases} \end{aligned}$$

or in vector notation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = e^x \begin{bmatrix} 20 \\ 0 \\ -10 \end{bmatrix} + e^{2x} \begin{bmatrix} -65 \\ -12 \\ 39 \end{bmatrix} + e^{4x} \begin{bmatrix} 48 \\ -12 \\ -24 \end{bmatrix}$$

□

**Short Exercise:** Derive another solution if the initial condition is  $y_1(x = 0) = 0$ ,  $y_2(x = 0) = 2$ ,  $y_3(x = 0) = -3$  instead.<sup>16</sup>

---

<sup>16</sup>The corresponding integration constants are  $c_1 = -1$ ,  $c_2 = -6$ ,  $c_3 = 8$ . So  $z_1 = -e^x$ ,  $z_2 = -6e^{2x}$ ,  $z_3 = 8e^{4x}$ , and  $y_1 = 2e^x + 30e^{2x} - 32e^{4x}$ ,  $y_2 = -6e^{2x} + 8e^{4x}$ ,  $y_3 = -e^x - 18e^{2x} + 16e^{4x}$ .

**Example 9.3.2.** Solve the system  $\mathbf{y}' = A\mathbf{y}$ , where

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

is the matrix as given in Example 9.2.3.

*Solution.* From the previous work we know that the eigenvectors are  $(1, 0, 0)^T$ ,  $(\iota, \iota, \sqrt{2})^T$  and  $(-\iota, -\iota, \sqrt{2})^T$  for  $\lambda = 1, 1 - \sqrt{2}\iota, 1 + \sqrt{2}\iota$  respectively. In this situation it is actually advantageous to work with complex numbers as we will see soon. Similar to the example above, letting  $\mathbf{y} = P\mathbf{z}$ , where

$$P = \begin{bmatrix} 1 & \iota & -\iota \\ 0 & \iota & -\iota \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}$$

then we have

$$\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z}$$

with

$$\begin{aligned} D &= P^{-1}AP = \begin{bmatrix} 1 & \iota & -\iota \\ 0 & \iota & -\iota \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & \iota & -\iota \\ 0 & \iota & -\iota \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - \sqrt{2}\iota & 0 \\ 0 & 0 & 1 + \sqrt{2}\iota \end{bmatrix} \end{aligned}$$

The solution in Properties 9.3.2 is valid even when  $\beta$  is complex. Hence the solution for  $\mathbf{z}$  will be

$$\begin{cases} z_1 = c_1 e^x \\ z_2 = c_2 e^{(1-\sqrt{2}\iota)x} = c_2 e^x e^{-\sqrt{2}\iota x} = c_2 e^x (\cos(\sqrt{2}x) - \iota \sin(\sqrt{2}x)) \\ z_3 = c_3 e^{(1+\sqrt{2}\iota)x} = c_3 e^x e^{\sqrt{2}\iota x} = c_3 e^x (\cos(\sqrt{2}x) + \iota \sin(\sqrt{2}x)) \end{cases}$$

where Euler's formula (Definition 8.1.6) is applied. Hence

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} 1 & i & -i \\ 0 & i & -i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} c_1 e^x \\ c_2 e^x (\cos(\sqrt{2}x) - i \sin(\sqrt{2}x)) \\ c_3 e^x (\cos(\sqrt{2}x) + i \sin(\sqrt{2}x)) \end{bmatrix} \\ &= \begin{bmatrix} c_1 e^x + (c_2 + c_3) e^x \sin(\sqrt{2}x) + i(c_2 - c_3) e^x \cos(\sqrt{2}x) \\ (c_2 + c_3) e^x \sin(\sqrt{2}x) + i(c_2 - c_3) e^x \cos(\sqrt{2}x) \\ \sqrt{2}(c_2 + c_3) e^x \cos(\sqrt{2}x) - i\sqrt{2}(c_2 - c_3) e^x \sin(\sqrt{2}x) \end{bmatrix} \\ &= \begin{bmatrix} c_1 e^x + C_2 e^x \sin(\sqrt{2}x) + iC_3 e^x \cos(\sqrt{2}x) \\ C_2 e^x \sin(\sqrt{2}x) + iC_3 e^x \cos(\sqrt{2}x) \\ \sqrt{2}C_2 e^x \cos(\sqrt{2}x) - i\sqrt{2}C_3 e^x \sin(\sqrt{2}x) \end{bmatrix} \end{aligned}$$

where we set  $C_2 = c_2 + c_3$ ,  $C_3 = c_2 - c_3$ . Both the real and imaginary part of  $\mathbf{y}$  will satisfy the system, and their linear combination forms the general solution. To see this, note that  $A$  is a real matrix, hence we can rewrite the ODE as

$$\begin{aligned} \mathbf{y}' &= A\mathbf{y} \\ \operatorname{Re}\{\mathbf{y}'\} + i \operatorname{Im}\{\mathbf{y}'\} &= A(\operatorname{Re}\{\mathbf{y}\} + i \operatorname{Im}\{\mathbf{y}\}) \\ \operatorname{Re}\{\mathbf{y}'\} + i \operatorname{Im}\{\mathbf{y}'\} &= \operatorname{Re}\{A\mathbf{y}\} + i \operatorname{Im}\{A\mathbf{y}\} \end{aligned}$$

Equating the real and imaginary parts then gives

$$\begin{aligned} \mathbf{y}'_{\operatorname{Re}} &= A\mathbf{y}_{\operatorname{Re}} \\ \mathbf{y}'_{\operatorname{Im}} &= A\mathbf{y}_{\operatorname{Im}} \end{aligned}$$

So the final answer, expressed in real values, is

$$\begin{cases} y_1 &= c_1 e^x + C_2 e^x \sin(\sqrt{2}x) + C_3 e^x \cos(\sqrt{2}x) \\ y_2 &= C_2 e^x \sin(\sqrt{2}x) + C_3 e^x \cos(\sqrt{2}x) \\ y_3 &= \sqrt{2}C_2 e^x \cos(\sqrt{2}x) - \sqrt{2}C_3 e^x \sin(\sqrt{2}x) \end{cases}$$

where  $c_1, C_2, C_3$  are to be decided by initial condition. The same solution basis of

$$\begin{bmatrix} e^x \sin(\sqrt{2}x) \\ e^x \cos(\sqrt{2}x) \\ \sqrt{2}e^x \cos(\sqrt{2}x) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} e^x \cos(\sqrt{2}x) \\ e^x \sin(\sqrt{2}x) \\ -\sqrt{2}e^x \sin(\sqrt{2}x) \end{bmatrix}$$

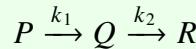
associated to  $C_2$  and  $C_3$ , will also be obtained if we consider only one eigenvalue out of the conjugate pair (either  $\lambda_- = 1 - \sqrt{2}i$  or  $\lambda_+ = 1 + \sqrt{2}i$ ) and compute either  $\mathbf{y} = \mathbf{v}_{\lambda_-} e^{\lambda_- x}$  or  $\mathbf{v}_{\lambda_+} e^{\lambda_+ x}$  where  $\mathbf{v}_{\lambda_-}$  and  $\mathbf{v}_{\lambda_+}$  are the corresponding eigenvectors.  $\square$

### Remarks

If the coefficient matrix for a system of first-order ODEs is not diagonalizable, then we need to employ alternative methods to solve the system. The most common approach, generalized from diagonalization, is to use Jordan Normal Form. This is an advanced topic and interested readers can go to the Appendix for reference.

## 9.4 Earth Science Applications

**Example 9.4.1.** A chemical tracer  $P$  decays into another tracer  $Q$ , which in turn decays to produce yet another tracer  $R$ . Both reactions are first-order



and hence have rate laws in form of

$$\begin{aligned} \frac{d[P]}{dt} &= -k_1[P] \\ \frac{d[Q]}{dt} &= k_1[P] - k_2[Q] \end{aligned}$$

$$\frac{d[R]}{dt} = k_2[Q]$$

Initially,  $[P] = 1$  mol per unit volume and  $[Q] = [R] = 0$ . ( $k_1, k_2$  have the unit of  $s^{-1}$ ). Express the time evolution of concentrations for all the three tracers  $[P]$ ,  $[Q]$ ,  $[R]$ , and find the time  $t_Q$  where  $[Q]$  reaches maximum in terms of  $k_1, k_2$ .

*Solution.* The three ODEs can be written as a matrix system of

$$\mathbf{y}' = A\mathbf{y} = \begin{bmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{bmatrix} \mathbf{y}$$

where  $\mathbf{y} = ([P], [Q], [R])^T$ . We will use Properties 9.3.4 and 9.2.2 to diagonalize the system. It is easy to see that the eigenvalues of the lower-triangular coefficient matrix  $A$  are  $0, -k_1$  and  $-k_2$ . We further assume that  $k_1 \neq k_2$ .<sup>17</sup> The eigenvector for  $\lambda = 0$  is clearly  $(0, 0, 1)^T$ , while that for  $\lambda = -k_1$  can be derived from solving

$$\left[ \begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ k_1 & k_1 - k_2 & 0 & 0 \\ 0 & k_2 & k_1 & 0 \end{array} \right]$$

which leads to an eigenvector of  $(k_1 - k_2, -k_1, k_2)^T$ . Similarly, the eigenvector for  $\lambda = -k_2$  can be checked to be  $(0, -1, 1)^T$ . The change of variables matrix will then be

$$P = \begin{bmatrix} k_1 - k_2 & 0 & 0 \\ -k_1 & -1 & 0 \\ k_2 & 1 & 1 \end{bmatrix}$$

and with  $\mathbf{y} = P\mathbf{z}$  we have

$$\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z}$$

---

<sup>17</sup>If  $k_1 = k_2$ , it can be found that the geometric multiplicity of  $\lambda = -k_1 = -k_2$  is only 1 while its algebraic multiplicity becomes 2 so diagonalization is not possible.

$$= \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{z}$$

the solution of which is  $z_1 = c_1 e^{-k_1 t}$ ,  $z_2 = c_2 e^{-k_2 t}$ ,  $z_3 = c_3$ . The initial conditions of  $[P](0) = 1$ ,  $[Q](0) = [R](0) = 0$  mean that

$$\begin{aligned} \mathbf{y}(0) &= P\mathbf{z}(0) \\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} k_1 - k_2 & 0 & 0 \\ -k_1 & -1 & 0 \\ k_2 & 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \end{aligned}$$

solving for it yields  $c_1 = \frac{1}{k_1 - k_2}$ ,  $c_2 = -\frac{k_1}{k_1 - k_2}$ ,  $c_3 = 1$ . Therefore,  $z_1 = \frac{1}{k_1 - k_2} e^{-k_1 t}$ ,  $z_2 = -\frac{k_1}{k_1 - k_2} e^{-k_2 t}$ ,  $z_3 = 1$ , and

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix} k_1 - k_2 & 0 & 0 \\ -k_1 & -1 & 0 \\ k_2 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{k_1 - k_2} e^{-k_1 t} \\ -\frac{k_1}{k_1 - k_2} e^{-k_2 t} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} e^{-k_1 t} \\ -\frac{k_1}{k_1 - k_2} e^{-k_1 t} + \frac{k_1}{k_1 - k_2} e^{-k_2 t} \\ \frac{k_2}{k_1 - k_2} e^{-k_1 t} - \frac{k_1}{k_1 - k_2} e^{-k_2 t} + 1 \end{bmatrix} \end{aligned}$$

Hence

$$\begin{cases} [P] &= e^{-k_1 t} \\ [Q] &= -\frac{k_1}{k_1 - k_2} e^{-k_1 t} + \frac{k_1}{k_1 - k_2} e^{-k_2 t} \\ [R] &= \frac{k_2}{k_1 - k_2} e^{-k_1 t} - \frac{k_1}{k_1 - k_2} e^{-k_2 t} + 1 \end{cases}$$

The time  $t_Q$  where  $[Q]$  reaches maximum occurs when  $\frac{d[Q]}{dt} = 0$ , which leads to

$$\begin{aligned} \frac{d[Q]}{dt} &= k_1[P] - k_2[Q] = 0 \\ k_1 e^{-k_1 t} - k_2 \left( -\frac{k_1}{(k_1 - k_2)} e^{-k_1 t} + \frac{k_1}{(k_1 - k_2)} e^{-k_2 t} \right) &= 0 \end{aligned}$$

$$\begin{aligned}
 k_1 e^{-k_1 t} + \frac{k_1 k_2}{(k_1 - k_2)} e^{-k_1 t} - \frac{k_1 k_2}{(k_1 - k_2)} e^{-k_2 t} &= 0 \\
 \frac{k_1^2}{(k_1 - k_2)} e^{-k_1 t} - \frac{k_1 k_2}{(k_1 - k_2)} e^{-k_2 t} &= 0 \\
 \frac{k_1^2}{(k_1 - k_2)} e^{-k_1 t} &= \frac{k_1 k_2}{(k_1 - k_2)} e^{-k_2 t} \\
 e^{-(k_1 - k_2)t} &= \frac{k_2}{k_1} \\
 -(k_1 - k_2)t &= \ln\left(\frac{k_2}{k_1}\right) \\
 \therefore t_Q &= -\frac{\ln\left(\frac{k_2}{k_1}\right)}{k_1 - k_2} = \frac{\ln k_1 - \ln k_2}{k_1 - k_2}
 \end{aligned}$$

□

**Example 9.4.2.** From Example 4.3.1, we know that the horizontal acceleration of a mass due to Coriolis Force is

$$\begin{aligned}
 \frac{du}{dt} &= fv \\
 \frac{dv}{dt} &= -fu
 \end{aligned}$$

where  $f = 2\Omega \sin \varphi$ , with  $\varphi$  being the latitude. Assume that some sea current in an inland sea (let's say, the Baltic Sea,  $\varphi \approx 58^\circ\text{N}$ ) is only subjected to Coriolis Force without any significant frictional dissipation or wind forcing, and hence the ODE system above is applicable to the change in the horizontal velocities of the current. Subsequently, solve this system and describe the nature of the sea current motion.

*Solution.* The two equations of motion can be written in a matrix system as

$$\frac{d\vec{v}}{dt} = A\vec{v} = \begin{bmatrix} 0 & f \\ -f & 0 \end{bmatrix} \vec{v}$$

where  $\vec{v} = (u, v)^T$ . The eigenvalues of the coefficient matrix are

$$\det(A - \lambda I) = \begin{vmatrix} -\lambda & f \\ -f & -\lambda \end{vmatrix} = 0$$

$$\lambda^2 + f^2 = 0 \implies \lambda = \pm fi$$

As noted in the end of Example 9.3.2, we can simply take either one of the conjugate eigenvalues, let's say  $\lambda_+ = fi$ , and compute the associated eigenvector:

$$\left[ \begin{array}{cc|c} -fi & f & 0 \\ -f & -fi & 0 \end{array} \right] \rightarrow \left[ \begin{array}{cc|c} -fi & f & 0 \\ 0 & 0 & 0 \end{array} \right] \quad R_2 + iR_1 \rightarrow R_2$$

which gives  $\vec{v}_{\lambda_+} = (-i, 1)^T$ . As a result,

$$\begin{aligned} \vec{v} &= \vec{v}_{\lambda_+} e^{\lambda_+ t} \\ &= \begin{bmatrix} -i \\ 1 \end{bmatrix} e^{ift} \\ &= \begin{bmatrix} -i \\ 1 \end{bmatrix} (\cos(ift) + i \sin(ift)) \quad (\text{Definition 8.1.6}) \\ &= \begin{bmatrix} -i(\cos(ift) + i \sin(ift)) \\ \cos(ift) + i \sin(ift) \end{bmatrix} \\ &= \begin{bmatrix} -i \cos(ift) + \sin(ift) \\ \cos(ift) + i \sin(ift) \end{bmatrix} = \begin{bmatrix} \sin(ift) \\ \cos(ift) \end{bmatrix} + i \begin{bmatrix} -\cos(ift) \\ \sin(ift) \end{bmatrix} \end{aligned}$$

Consider the real and imaginary parts as two linearly independent solutions, we have  $\vec{v} = c_1 \begin{bmatrix} \sin(ift) \\ \cos(ift) \end{bmatrix} + c_2 \begin{bmatrix} -\cos(ift) \\ \sin(ift) \end{bmatrix}$ , which translates to  $u = c_1 \sin(ift) - c_2 \cos(ift)$ ,  $v = c_1 \cos(ift) + c_2 \sin(ift)$ , a (counter-)clockwise elliptic motion in the Northern (Southern) Hemisphere where the sign of inertial frequency  $f$  is positive (negative). This is known as an *inertial oscillation*, which has a period of  $T = \frac{2\pi}{f} = \frac{2\pi}{2\Omega \sin \varphi} = \frac{1}{2 \sin \varphi} \left( \frac{2\pi}{\Omega} \right) = \frac{1}{2 \sin \varphi} (24 \text{ h})$ . In the case of the Baltic Sea, the oscillation period will then be about  $T = \frac{1}{2 \sin(58^\circ)} (24 \text{ h}) = 14.1 \text{ h}$ .  $\square$

## 9.5 Python Programming

Eigenvalues and eigenvectors of a square matrix can be computed through calling the `eig` function in the `scipy.linalg` library. Let's try it on our previous examples. First, for Example 9.1.2:

```
import numpy as np
from scipy import linalg

A = np.array([[1., 3., 1.],
              [0., 1., 0.],
              [1., 0., 2.]])
eigvals, eigvecs = linalg.eig(A)
```

`print(eigvals)` yields the eigenvalues of

```
[0.3820+0.j 2.618+0.j 1.+0.j]
```

$(\frac{3}{2} - \frac{\sqrt{5}}{2}, \frac{3}{2} + \frac{\sqrt{5}}{2}, 1)$  as expected, and `print(eigvecs)` outputs the eigenvectors in columns

```
[[ -0.85065081 -0.52573111  0.6882472 ]
 [ 0.          0.          0.22941573]
 [ 0.52573111 -0.85065081 -0.6882472 ]]
```

(normalized to unit length) corresponding to the eigenvalues as arranged in `eigvals`. Now if we test it on Example 9.1.1:

```
B = np.array([[1, -1],
              [0, 1]])
eigvals, eigvecs = linalg.eig(B)
```

then

```
print(eigvals)
print(eigvecs)
```

produces

```
[1.+0.j 1.+0.j]
[[1.000e+00 1.000e+00]
 [0.000e+00 2.220e-16]]
```

which indicates two eigenvectors for the eigenvalue of  $\lambda = 1$ . However, we already know that the geometric multiplicity of  $\lambda = 1$  is 1 where the only eigenvector is  $(1, 0)^T$ , and the second eigenvector given by `eigvecs` is spurious (notice that it is essentially the same as the first eigenvector with a very small perturbed value in the second component). This is because `scipy` is guaranteed to produce the same amount of eigenvectors as the extent of the matrix, furthered by the fact that it uses numerical approximation which contains round-off errors. To avoid this, we can use `sympy` instead for a more analytical output.

```
import sympy

B_sympy = sympy.Matrix(B)
print(B_sympy.eigenvals())
```

The `eigenvals` method gives  $\{1: 2\}$  which means that the eigenvalue of  $\lambda = 1$  has an algebraic multiplicity of 2. To compute the eigenvector(s), we similarly call the `eigenvects` method:

```
print(B_sympy.eigenvects())
```

which returns

```
[(1, 2, [Matrix([[1], [0]])])]
```

the last item in the list containing a single eigenvector of  $(1, 0)^T$  correctly. Diagonalization can be easily performed by the `diagonalize` method in `sympy`:

```
A_sympy = sympy.Matrix(A)
print(A_sympy.diagonalize())
```

which gives

```
(Matrix([[ 0.8506, -0.7265,    0.2482],
        [-7.5964e-65, -0.2421, -6.208e-66],
        [-0.52574,   0.7265,    0.4017]]), Matrix([
[0.3819,   0,      0],
[ 0,     1.0,    0],
[ 0,      0,  2.6180]]))
```

as the  $P$  (with small round-off errors for the zeros) and  $D$  matrix in  $D = P^{-1}AP$ .

## 9.6 Exercises

**Exercise 9.1** Argue that if the eigenvalues of a matrix  $A$  all have an algebraic multiplicity of 1, or in other words, no repeated root for the characteristic equation, then  $A$  is diagonalizable.

**Exercise 9.2** Find the eigenvalues of the following matrices.

$$A = \begin{bmatrix} 1 & 3 & 3 \\ 4 & 0 & 4 \\ 0 & 1 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

as well as their transpose.

**Exercise 9.3** Find the eigenvalues and corresponding eigenvectors for

$$A = \begin{bmatrix} 3 & 1 & 4 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Repeat the calculation for its inverse.

**Exercise 9.4** Find the eigenvalues and corresponding eigenvectors for

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

and perform Gram-Schmidt orthogonalization on the eigenvectors found.

**Exercise 9.5** Find all the eigenvalues and eigenvectors for the matrix

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

as well as its transposes  $A^T$ . Use your answer to support the fact that the transpose of a matrix always has the same eigenvalues, but may or may not lead to the same eigenvectors.

**Exercise 9.6** For the matrix

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

compute its characteristic polynomial. By applying Cayley-Hamilton Theorem,

(a) Find  $A^3$ ,

(b) Prove

$$A^n = \begin{bmatrix} 1 & -(2^n - 1) & 2^n - 1 \\ 1 & 1 & -1 \\ 1 & -(2^n - 1) & 2^n - 1 \end{bmatrix}$$

for  $n \geq 3$  by Mathematical Induction (assume this holds for  $n = k$ , then prove for  $n = k + 1$ ).

**Exercise 9.7** Apply diagonalization for the following matrix, and show that the characteristic polynomial remains unchanged under diagonalization, if possible.

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 2 \\ 0 & 0 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 5 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

**Exercise 9.8** Solve the following system of ordinary differential equations.

$$\begin{cases} y'_1 &= -y_1 + y_2 + 3y_3 \\ y'_2 &= -2y_1 + 3y_2 + 2y_3 \\ y'_3 &= -2y_1 + y_2 + 4y_3 \end{cases}$$

with the initial condition  $y_1(0) = 3, y_2(0) = 2, y_3(0) = 9$ .

**Exercise 9.9** Given an idealized situation where there are three chemical gases  $P, Q, R$ . Denote their concentrations by  $[P], [Q], [R]$  respectively. If they undergo reactions in a closed pathway  $P \rightarrow Q \rightarrow R \rightarrow P$  that can be regarded as first-order, so that the governing equations about their concentrations are

$$\begin{cases} d[P]/dt &= k_{31}[R] - k_{12}[P] \\ d[Q]/dt &= k_{12}[P] - k_{23}[Q] \\ d[R]/dt &= k_{23}[Q] - k_{31}[R] \end{cases}$$

where  $k_{mn}$  are all constants. Derive the time evolution of the concentrations of the three gases. What happens when  $t \rightarrow \infty$ ?

## Chapter 10

# Orthogonal and Normal Matrices

---

We have discussed about orthonormal bases in Section 7.2.2 and it is logical to go one step further and make a coordinate transformation matrix with these orthonormal basis vectors just as if they are ordinary basis vectors. It is not surprising that such a matrix is called an *orthonormal(-gonal) matrix*, but it also turns out that they carry some important properties which will be explored in this chapter. Particularly, we may want to ask if there exists an orthogonal change of coordinates matrix, which represents rotation and reflection geometrically, for making a given square matrix become diagonal, a.k.a. *orthogonal diagonalization* which is a stronger version of diagonalization introduced in the last chapter. Eventually this will lead to a major result known as the *Spectral Theorem*. These concepts will also be promoted to complex vectors and matrices, where the complex counterpart of orthogonal(-normal) is now known as *unitary*, accompanied by a less stringent condition known as *normal*.

## 10.1 Orthogonal Matrices

### 10.1.1 Definition of Orthogonal Matrices

In Section 4.2.1, we have talked about how two  $\mathbb{R}^n$  vectors are orthogonal to each other when their dot product is zero. We can extend the concept of orthogonality

to a matrix. An ***orthogonal matrix*** (sometimes called ***orthonormal***) is a matrix, each column of which is an  $\mathbb{R}^n$  vector of unit length and orthogonal to all other columns. We further require the number of columns is  $n$  too so that they actually form an ***orthonormal basis*** by Properties 6.3.12 and (b) of Properties 6.2.7. Hence it will be a square matrix.

**Definition 10.1.1** (Orthogonal Matrix). An  $n \times n$  square matrix

$$P = [\vec{v}^{(1)} | \dots | \vec{v}^{(j)} | \dots | \vec{v}^{(n)}]$$

is said to be an orthogonal matrix, if all of its columns  $\vec{v}^{(j)} \in \mathbb{R}^n$  satisfy the condition:

$$\vec{v}^{(p)} \cdot \vec{v}^{(q)} = \begin{cases} 1 & \text{if } p = q \quad \text{i.e. } \|\vec{v}^{(p)}\| = 1 \\ 0 & \text{if } p \neq q \end{cases}$$

so that  $\vec{v}^{(j)}, j = 1, 2, \dots, n$  forms an orthonormal basis.

The orthonormal basis vectors in an orthogonal matrix can be generated through the procedure of Gram-Schmidt orthogonalization with normalization, introduced in Section 7.2.2.

Short Exercise: Verify that

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}$$

is an orthogonal matrix.<sup>1</sup>

---

<sup>1</sup>There are 3  $\mathbb{R}^3$  column vectors in the matrix, where  $\vec{v}^{(1)} \cdot \vec{v}^{(2)} = (\frac{1}{\sqrt{2}})(\frac{1}{\sqrt{3}}) + (\frac{1}{\sqrt{2}})(-\frac{1}{\sqrt{3}}) + (0)(\frac{1}{\sqrt{3}}) = 0$ ,  $\vec{v}^{(1)} \cdot \vec{v}^{(3)} = (\frac{1}{\sqrt{2}})(-\frac{1}{\sqrt{6}}) + (\frac{1}{\sqrt{2}})(\frac{1}{\sqrt{6}}) + (0)(\frac{\sqrt{2}}{\sqrt{3}}) = 0$ . We leave to the readers to check  $\vec{v}^{(2)} \cdot \vec{v}^{(3)} = 0$  as well. Also,  $\|\vec{v}^{(1)}\| = \sqrt{(\frac{1}{\sqrt{2}})^2 + (\frac{1}{\sqrt{2}})^2 + (0)^2} = 1$  and  $\|\vec{v}^{(2)}\| = \sqrt{(\frac{1}{\sqrt{3}})^2 + (-\frac{1}{\sqrt{3}})^2 + (\frac{1}{\sqrt{3}})^2} = 1$ . Again we let the readers to check  $\|\vec{v}^{(3)}\| = 1$  too.

Due to its definition, an orthogonal matrix  $P$  has the property  $P^T P = I$ , since the resulting entries of this matrix product are basically dot products of the column vectors of  $P$  (see the explanation below Definition 4.2.1), explicitly

$$\begin{aligned}
 P^T P &= [\vec{v}^{(1)} | \dots | \vec{v}^{(j)} | \dots | \vec{v}^{(n)}]^T [\vec{v}^{(1)} | \dots | \vec{v}^{(j)} | \dots | \vec{v}^{(n)}] \\
 &= \begin{bmatrix} \vec{v}^{(1)T} \\ \vdots \\ \vec{v}^{(j)T} \\ \vdots \\ \vec{v}^{(n)T} \end{bmatrix} [\vec{v}^{(1)} | \dots | \vec{v}^{(j)} | \dots | \vec{v}^{(n)}] \\
 &= \begin{bmatrix} \vec{v}^{(1)} \cdot \vec{v}^{(1)} & \dots & \vec{v}^{(1)} \cdot \vec{v}^{(j)} & \dots & \vec{v}^{(1)} \cdot \vec{v}^{(n)} \\ \vdots & \ddots & \vdots & & \vdots \\ \vec{v}^{(j)} \cdot \vec{v}^{(1)} & \dots & \vec{v}^{(j)} \cdot \vec{v}^{(j)} & \dots & \vec{v}^{(j)} \cdot \vec{v}^{(n)} \\ \vdots & & \vdots & \ddots & \vdots \\ \vec{v}^{(n)} \cdot \vec{v}^{(1)} & \dots & \vec{v}^{(n)} \cdot \vec{v}^{(j)} & \dots & \vec{v}^{(n)} \cdot \vec{v}^{(n)} \end{bmatrix} \\
 &= \begin{bmatrix} 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 \end{bmatrix} = I_n
 \end{aligned}$$

By Definition 2.2.1,  $P^T P = I$  further leads to  $PP^T = P^T P = I$  where the transpose of the orthogonal matrix  $P^T = P^{-1}$  is at the same time its inverse. The argument works in both direction such that  $PP^T = I$  also means that  $P$  is orthogonal.

**Properties 10.1.2.**  $P$  is an orthogonal matrix if and only if  $PP^T = P^T P = I$  and its inverse is simply its transpose  $P^{-1} = P^T$ .

The equivalence between  $P^T P = I$  and  $PP^T = I$  for an orthogonal matrix extends Definition 10.1.1 where the former equality indicates that the rows of  $P$  also has to form an orthonormal basis. So a parallel definition of orthogonal matrices is

**Definition 10.1.3.** An  $n \times n$  square matrix

$$P = \begin{bmatrix} \vec{w}^{(1)T} \\ \vdots \\ \vec{w}^{(i)T} \\ \vdots \\ \vec{w}^{(n)T} \end{bmatrix}$$

is an orthogonal matrix, also if all row vectors  $\vec{w}^{(i)} \in \mathbb{R}^n$  satisfy the similar condition:

$$\vec{w}^{(p)} \cdot \vec{w}^{(q)} = \begin{cases} 1 & \text{if } p = q \quad \text{i.e. } \|\vec{w}^{(p)}\| = 1 \\ 0 & \text{if } p \neq q \end{cases}$$

so that  $\vec{w}^{(i)}, i = 1, 2, \dots, n$  forms an orthonormal basis too.

Short Exercise: Confirm Properties 10.1.2 for the matrix in the last short exercise.<sup>2</sup>

## 10.1.2 Geometric Implications of Orthogonal Matrices

All invertible matrices represent some sorts of coordinate transformation as highlighted in Section 7.2.1. Orthogonal matrices actually further belong to

---

<sup>2</sup>It is a direct computation that shows

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and  $PP^T = I$  similarly.

a special class of them. Before going into details, we need to start with some simple observations.

**Properties 10.1.4.** An orthogonal matrix  $P$  has a determinant of either 1 or  $-1$ .

*Proof.* Starting from Properties 10.1.2 and taking determinant on both sides, we have

$$\begin{aligned}\det(P^T P) &= \det(I) \\ \det(P^T) \det(P) &= (\det(P))^2 = 1 \\ \det(P) &= \pm 1\end{aligned}$$

where Properties 2.3.9 is used.  $\square$

Non-zero determinant reaffirms that the  $n \mathbb{R}^n$  row/column vectors in an orthogonal matrix  $P$  are linearly independent ((b) to (e) of Theorem 6.1.10).

**Properties 10.1.5.** Coordinate transformation by an orthogonal matrix  $P$  on a vector is length-preserving.

*Proof.* Assume the coordinate transformation is the passive type as described in Section 7.2.1. Then the new coordinates of a vector  $\vec{v}$  will be  $P^{-1}\vec{v} = P^T\vec{v}$  (Properties 10.1.2). The length of the vector in the new coordinate frame is

$$\begin{aligned}\|P^T\vec{v}\| &= \sqrt{(P^T\vec{v}) \cdot (P^T\vec{v})} && \text{(Properties 4.2.2)} \\ &= \sqrt{(P^T\mathbf{v})^T (P^T\mathbf{v})} && \text{(Properties 4.2.3)} \\ &= \sqrt{(\mathbf{v}^T P P^T \mathbf{v})} \\ &= \sqrt{(\mathbf{v}^T I \mathbf{v})} && \text{(Properties 10.1.2)} \\ &= \sqrt{(\mathbf{v}^T \mathbf{v})} = \|\vec{v}\|\end{aligned}$$

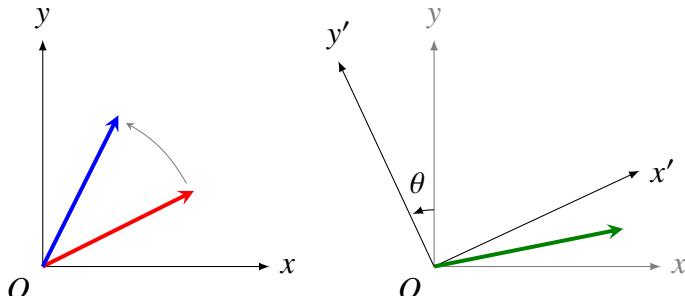
So the length of the vector remains the same when the coordinate system is changed. For active transformation that actually changes the vector instead of

the frame, the argument is very similar and the new vector will also have the same length as the old one. In fact, Properties 10.1.5 implies Properties 10.1.4 (the discussion about the geometric interpretation of coordinate transformation in Section 7.2.1 will be helpful).  $\square$

Now we are ready to see what kind of coordinate transformation an orthogonal matrix means.

**Theorem 10.1.6.** For an orthogonal matrix  $P$ , if  $P$  has a determinant of 1, then it is a rotation. On the other hand, if  $P$  has a determinant of  $-1$ , then it implies a reflection.

Just as in Section 7.2.1, depending on the situation, rotation and reflection can be viewed as (a) rotation and reflection of the vector while keeping the coordinate basis unchanged (active transformation), or (b) rotation and reflection of the coordinate system, while keeping the vector in the original place (passive transformation). By the previous length-preserving properties there will be no stretching/compression or shearing, and the rotation/reflection will be pure.



Left: case (a), active transformation, and Right: case (b), passive transformation, for 2D rotation.

The construction of a rotational or reflectional matrix requires the representation of the new unit axes in the old coordinate system as suggested in Theorem 7.1.12, but with the extra condition that they are orthogonal to each other. Once they are found, they can be combined column by column to form the transition

matrix. For an anti-clockwise (positive) rotation like the one to the right side of the figure above, we have

$$\begin{aligned}x' &= (\cos \theta)x + (\sin \theta)y \\y' &= (-\sin \theta)x + (\cos \theta)y\end{aligned}$$

The corresponding transformation matrix is then

$$\begin{aligned}P_B^S &= [[\hat{x}']_S | [\hat{y}']_S] \\&= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}\end{aligned}$$

where  $S$  denotes the original standard basis system, and  $B$  indicates the rotated coordinates. Alternatively,

$$\begin{aligned}P_S^B &= (P_B^S)^{-1} \\&= (P_B^S)^T \quad (\text{Properties 10.1.2}) \\&= \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}\end{aligned}$$

These two matrices can be compared to the one we see in the last chapter when we study the real variant of diagonalization for complex eigenvalues.

If we have an orthogonal matrix  $P$ , and a vector  $\vec{v}_0$  in the standard basis, then computing  $\vec{v}_n = P\vec{v}_0$  in a forward direction can be viewed as directly applying the corresponding rotation or reflection on the vector  $\vec{v}_0$  to make  $\vec{v}_n$  while staying in the same coordinate frame. This is the case (a) we have discussed above. For case (b), it is the exact opposite. Rotation or reflection of the coordinate system, where the concerned vector  $\vec{v}_0$  is fixed physically, is achieved by solving  $\vec{v}_0 = P\vec{v}_n$  in a backward manner, or equivalently  $\vec{v}_n = P^{-1}\vec{v}_0 = P^T\vec{v}_0$ . Taking transpose, it becomes  $\vec{v}_n^T = \vec{v}_0^T P$ . In such situation, the vector is still the same vector in a physical sense, but represented in the new coordinate system. Most situations belong to case (b), on which thereafter we will focus our discussion.

However, as a reminder, these two cases are much related. For example, an anti-clockwise/positive rotation of a vector in case (a), can be viewed as a

clockwise/negative rotation of the coordinate system in case (b), and vice versa. The two operations are only differed by a transpose. For successive rotations and reflections, the net transition matrix  $P_f$  is produced by taking the product of individual rotational and reflectional matrices each by each. One common convention has the order from right to left, where  $\vec{v}_n = (P_n^T \cdots P_3^T P_2^T P_1^T) \vec{v}_0$ , and  $P_f^T = P_n^T \cdots P_3^T P_2^T P_1^T$ . Another equivalent option is to do it from left to right by applying a transpose, as in  $\vec{v}_n^T = \vec{v}_0^T (P_1 P_2 P_3 \cdots P_n)$ ,  $P_f = P_1 P_2 P_3 \cdots P_n$ .

**Example 10.1.1.** Find the net transition matrix, if a rotation about  $y$ -axis of 40 degree in the positive direction is done first to produce an intermediate coordinate system  $x'$ ,  $y'$ ,  $z'$ , and then a reflection across the  $x'$ - $y'$  plane is made to generate the final coordinate system  $x''$ ,  $y''$ ,  $z''$  (so the new  $z''$  axis is the negative of  $z'$  axis).

*Solution.* The first transition matrix is

$$P_1 = \begin{bmatrix} \cos(40^\circ) & 0 & \sin(40^\circ) \\ 0 & 1 & 0 \\ -\sin(40^\circ) & 0 & \cos(40^\circ) \end{bmatrix}$$

The readers should verify this. The second transition matrix is simply

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

So the net transition matrix is

$$\begin{aligned} P_f &= P_1 P_2 \\ &= \begin{bmatrix} \cos(40^\circ) & 0 & \sin(40^\circ) \\ 0 & 1 & 0 \\ -\sin(40^\circ) & 0 & \cos(40^\circ) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \cos(40^\circ) & 0 & -\sin(40^\circ) \\ 0 & 1 & 0 \\ -\sin(40^\circ) & 0 & -\cos(40^\circ) \end{bmatrix} \end{aligned}$$

□

Short Exercise: What happens if we reverse the order of rotation/reflection?<sup>3</sup>

**Example 10.1.2.** For a vector  $\vec{v}_0$  in the three-dimensional standard basis  $\mathcal{S}$ , if the coordinate system undergoes a positive rotation about the straight line  $x = 0, y = z$  by a degree of  $\theta$ , find its representation  $\vec{v}_n$  in the new system  $\mathcal{B}$ .

*Solution.* The common way to construct transition matrix is to find the new axes in terms of the old coordinates. However, in this case it is harder because the axis about which the rotation occurs is not along any of the main axes. Here, we can do another rotation

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(45^\circ) & \sin(45^\circ) \\ 0 & -\sin(45^\circ) & \cos(45^\circ) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

about the  $x$ -axis first with a degree of  $45^\circ$  clockwise so that the intermediate  $z'$  axis is oriented along the desired line, while the  $y'$  axis points in the  $x = 0, y = -z$  direction relative to the original coordinate frame. Then we can apply the required rotation in the intermediate coordinate system about that  $z'$  axis, which has a simple transition matrix representation of

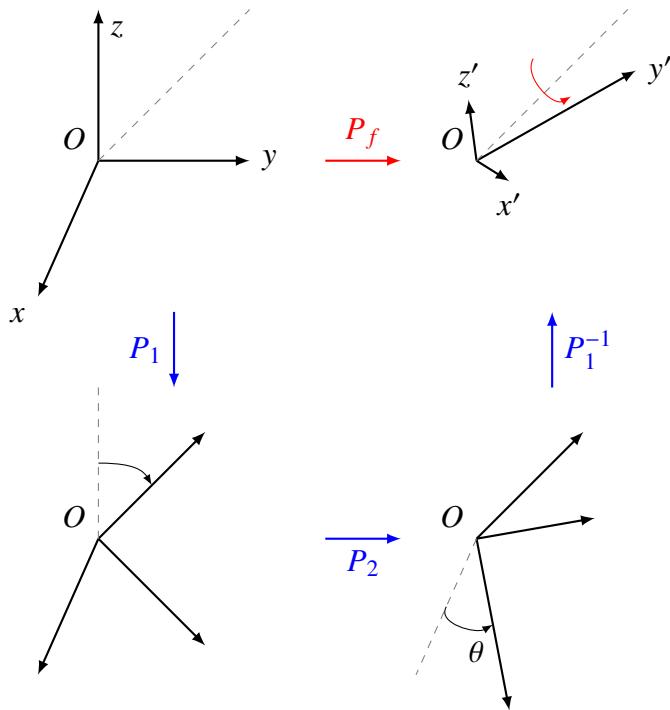
$$P_2 = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Finally, we undo the effect of the first rotation by multiplying by its inverse  $P_1^{-1}$  which is an orthogonal matrix and equal to  $P_1^T$ .

<sup>3</sup>The transition matrix becomes

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \cos(40^\circ) & 0 & \sin(40^\circ) \\ 0 & 1 & 0 \\ -\sin(40^\circ) & 0 & \cos(40^\circ) \end{bmatrix} = \begin{bmatrix} \cos(40^\circ) & 0 & \sin(40^\circ) \\ 0 & 1 & 0 \\ \sin(40^\circ) & 0 & -\cos(40^\circ) \end{bmatrix} \neq \begin{bmatrix} \cos(40^\circ) & 0 & -\sin(40^\circ) \\ 0 & 1 & 0 \\ -\sin(40^\circ) & 0 & -\cos(40^\circ) \end{bmatrix}$$

In general, finite rotations/reflections are not commutative and the order matters.



Hence the net transition matrix is

$$\begin{aligned}
 P_f &= P_1 P_2 P_1^{-1} = P_1 P_2 P_1^T \\
 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \\
 &= \begin{bmatrix} \cos \theta & -\frac{\sin \theta}{\sqrt{2}} & \frac{\sin \theta}{\sqrt{2}} \\ \frac{\sin \theta}{\sqrt{2}} & \frac{(\cos \theta)+1}{2} & \frac{-(\cos \theta)+1}{2} \\ -\frac{\sin \theta}{\sqrt{2}} & \frac{-(\cos \theta)+1}{2} & \frac{(\cos \theta)+1}{2} \end{bmatrix}
 \end{aligned}$$

For any vector  $\vec{v}_0 = (x_0, y_0, z_0)^T$  expressed in the standard coordinate basis, the new coordinates after rotation is

$$\vec{v}_0 = P_f \vec{v}_n$$

$$\begin{aligned}
 \vec{v}_n &= (P_f)^T \vec{v}_0 \\
 &= \begin{bmatrix} \cos \theta & \frac{\sin \theta}{\sqrt{2}} & -\frac{\sin \theta}{\sqrt{2}} \\ -\frac{\sin \theta}{\sqrt{2}} & \frac{(\cos \theta)+1}{2} & \frac{-(\cos \theta)+1}{2} \\ \frac{\sin \theta}{\sqrt{2}} & \frac{-(\cos \theta)+1}{2} & \frac{(\cos \theta)+1}{2} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \\
 &= \begin{bmatrix} (\cos \theta)x_0 + \left(\frac{\sin \theta}{\sqrt{2}}\right)y_0 + \left(-\frac{\sin \theta}{\sqrt{2}}\right)z_0 \\ \left(-\frac{\sin \theta}{\sqrt{2}}\right)x_0 + \left(\frac{(\cos \theta)+1}{2}\right)y_0 + \left(\frac{-(\cos \theta)+1}{2}\right)z_0 \\ \left(\frac{\sin \theta}{\sqrt{2}}\right)x_0 + \left(\frac{-(\cos \theta)+1}{2}\right)y_0 + \left(\frac{(\cos \theta)+1}{2}\right)z_0 \end{bmatrix}
 \end{aligned}$$

□

Short Exercise: For the example above, if  $(x_0, y_0, z_0)^T = (0, 1, 1)^T$  and  $\theta = \frac{\pi}{6}$ , find  $\vec{v}_n = (x_n, y_n, z_n)^T$ .<sup>4</sup>

## 10.2 Orthogonal Diagonalization

**Orthogonal diagonalization** is a special case of diagonalization, in which the transformation matrix  $P$  used to diagonalize the target matrix  $A$  is an orthogonal matrix. Since from Section 9.2.1 we know that diagonalization boils down to asking if there exists a basis, represented by the columns of  $P$ , such that  $D = P^{-1}AP$  is diagonal, orthogonal diagonalization is equivalent to demanding further that such a basis is orthonormal, i.e.  $P$  is an orthogonal matrix. We will discuss the case where  $A$  is a real matrix first.

---

<sup>4</sup>It is

$$\begin{bmatrix} \cos \frac{\pi}{6} & \frac{\sin \frac{\pi}{6}}{\sqrt{2}} & -\frac{\sin \frac{\pi}{6}}{\sqrt{2}} \\ -\frac{\sin \frac{\pi}{6}}{\sqrt{2}} & \frac{(\cos \frac{\pi}{6})+1}{2} & \frac{-(\cos \frac{\pi}{6})+1}{2} \\ \frac{\sin \frac{\pi}{6}}{\sqrt{2}} & \frac{-(\cos \frac{\pi}{6})+1}{2} & \frac{(\cos \frac{\pi}{6})+1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \left(\frac{\sqrt{3}}{2}\right)(0) + \left(\frac{1}{2\sqrt{2}}\right)(1) + \left(-\frac{1}{2\sqrt{2}}\right)(1) \\ \left(-\frac{1}{2\sqrt{2}}\right)(0) + \left(\frac{\frac{1}{2}+1}{2}\right)(1) + \left(\frac{-\frac{1}{2}+1}{2}\right)(1) \\ \left(\frac{1}{2\sqrt{2}}\right)(0) + \left(\frac{-\frac{1}{2}+1}{2}\right)(1) + \left(\frac{\frac{1}{2}+1}{2}\right)(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

In fact, all vector in form of  $s(0, 1, 1)^T$  where  $s$  is any number will keep the same coordinates after transformation, no matter what the value  $\theta$  takes, because the vector is oriented exactly along the rotational axis in question.

**Definition 10.2.1** (Orthogonal Diagonalization). Orthogonal diagonalization on a real  $n \times n$  square matrix  $A$  is finding an orthogonal matrix  $P$  that is consisted of columns being the  $n$  orthonormal  $\mathbb{R}^n$  eigenvectors of  $A$  (Definition 10.1.1) such that  $P^{-1}AP = P^TAP = D$  is a real diagonal matrix (Properties 9.2.2 and 10.1.2).

Note that the requirements of eigenvectors being  $\mathbb{R}^n$  and orthogonality defined by the real dot product implicitly restrain us to work over  $\mathbb{R}$  and only real eigenvalues are allowed. Not all real square matrices can be orthogonally diagonalized. Since orthogonal diagonalization is a stronger version of the ordinary diagonalization, the requirement is unsurprisingly stricter, however, it turns out that there is a really simple criterion: whether the matrix is symmetric.

**Theorem 10.2.2.** A real square matrix  $A$  can be orthogonally diagonalized if and only if  $A$  is symmetric.

The "only if" part is very easy to show.<sup>5</sup> As a corollary, all symmetric matrices have only real eigenvalues. On the other hand, the "if" part is much more difficult to derive and requires us to build some immediate results first, which will be done sequentially in the remaining of this section. Since the transformation matrix  $P$  used to diagonalize a given matrix  $A$  is consisted of the eigenvectors of  $A$  as pointed out by Properties 9.2.2, for  $P$  to be an orthogonal matrix, obviously we need to show that these eigenvectors are orthogonal to each other if  $A$  is symmetric. To this end, we have the following observation.

**Properties 10.2.3.** Any two eigenvectors of a real, symmetric matrix corresponding to two distinct eigenvalues are always orthogonal to each other.

*Proof.* Denote the two eigenvectors as  $\vec{v}_{\lambda_1}$  and  $\vec{v}_{\lambda_2}$  which have two different eigenvalues  $\lambda_1$  and  $\lambda_2$ . Then consider the quantity  $\vec{v}_{\lambda_1} \cdot (A\vec{v}_{\lambda_2})$  where  $A$  is the real,

---

<sup>5</sup>Note that if  $A$  is orthogonally diagonalizable then  $A = PDP^{-1} = PDPT$ . Then  $A^T = (PDP^T)^T = PD^TP^T = PDP^T = A$  since the transpose of any diagonal matrix  $D^T = D$  equals to itself.

symmetric matrix. By Definition 9.1.1, it is equal to  $\vec{v}_{\lambda_1} \cdot (\lambda_2 \vec{v}_{\lambda_2}) = \lambda_2 (\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2})$ . At the same time, using Properties 4.2.3 too, we have

$$\begin{aligned}\vec{v}_{\lambda_1} \cdot (A\vec{v}_{\lambda_2}) &= (A^T \vec{v}_{\lambda_1}) \cdot \vec{v}_{\lambda_2} \\ &= (A\vec{v}_{\lambda_1}) \cdot \vec{v}_{\lambda_2} && (A \text{ is symmetric}) \\ &= (\lambda_1 \vec{v}_{\lambda_1}) \cdot \vec{v}_{\lambda_2} = \lambda_1 (\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2}) && (\text{Definition 9.1.1 again})\end{aligned}$$

Therefore

$$\begin{aligned}\vec{v}_{\lambda_1} \cdot (A\vec{v}_{\lambda_2}) &= \lambda_1 (\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2}) = \lambda_2 (\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2}) \\ (\lambda_1 - \lambda_2) (\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2}) &= 0\end{aligned}$$

Since the two eigenvalues are required to be distinct,  $\lambda_1 \neq \lambda_2$ , and  $\vec{v}_{\lambda_1} \cdot \vec{v}_{\lambda_2}$  must be 0. By Properties 4.2.5, the two eigenvectors  $\vec{v}_{\lambda_1}$  and  $\vec{v}_{\lambda_2}$  are orthogonal to each other.  $\square$

Even when there are multiple eigenvectors (a geometric multiplicity/an eigenspace of dimension  $\geq 2$ ) associated to the same eigenvalue, we can mediate the problem by using the Gram-Schmidt Orthogonalization procedure introduced in Section 7.2.2 to produce an orthonormal basis to that eigenspace. Since all vectors in the eigenspace are subjected to the same eigenvalue, the choice of new eigenvectors will not disrupt the diagonalization process. So, with these, we have solved the "orthogonal" part of "orthogonal diagonalization", and the remaining half of the problem is to justify why symmetric matrices are always "diagonalizable", i.e. the geometric multiplicity of any eigenvalue is always equal to its algebraic multiplicity following Footnote 7 of Chapter 9, or in other words, there is no "*deficient*" eigenvalue.

**Properties 10.2.4.** The geometric multiplicity of an eigenvalue to a real symmetric matrix is always strictly equal to the algebraic multiplicity.

*Proof.* Let  $A$  be the  $n \times n$  real symmetric matrix. Denote the geometric multiplicity of some eigenvalue  $\lambda_{J_i}$  as  $n_{J_i}$  and its algebraic multiplicity by  $k_{J_i}$ . We will assume that  $n_{J_i} < k_{J_i}$  and derive a contradiction. The geometric

multiplicity of  $n_{J_i}$  implies that there are  $n_{J_i}$  (orthonormal, see the discussion above) eigenvectors  $\vec{v}_{\lambda_{J_i}}^{(1)}, \vec{v}_{\lambda_{J_i}}^{(2)}, \dots, \vec{v}_{\lambda_{J_i}}^{(n_{J_i})}$  (in  $\mathbb{R}^n$ ) corresponding to  $\lambda_{J_i}$  in the eigenspace  $\mathcal{E}_{J_i}$ . By (c) of Properties 6.2.7 and the Gram-Schmidt process (Definition 7.2.4), we can complete an orthonormal basis  $\mathcal{E}_{J_i} \oplus \mathcal{E}_{J_i^C}$  for  $\mathbb{R}^n$  (Properties 7.2.6) by filling them with some other  $n - n_{J_i}$  vectors. Let  $P$  be the orthogonal coordinate matrix that represents this basis in columns, such that

$$P = \left[ \vec{v}_{\lambda_{J_i}}^{(1)} | \vec{v}_{\lambda_{J_i}}^{(2)} | \cdots | \vec{v}_{\lambda_{J_i}}^{(n_{J_i})} | \text{other vectors used to fill the basis for } \mathbb{R}^n \right]$$

similar to the derivation of Theorem 9.1.3, and with the same logic we can apply a change of coordinates over  $A$  via  $P$ , such that

$$\begin{aligned} A' &= P^{-1}AP \\ &= \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix} \end{aligned}$$

However, this time, as  $P$  is designed to be an orthogonal matrix,  $A' = P^{-1}AP = P^TAP$  (Properties 10.1.2) can be easily shown to be symmetric like  $A$  as well.<sup>6</sup> Therefore, the upper-right block of  $A'$  also has to be a zero submatrix:

$$A' = \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & [\mathbf{0}]_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix}$$

which is now in the form of a matrix direct sum. This shows that we can treat it as  $A' = \lambda_{J_i} I_{n_{J_i}} \oplus A^C$  with respect to the  $\mathcal{E}_{J_i} \oplus \mathcal{E}_{J_i^C}$  vector direct sum basis following Definition 8.3.7 where  $A^C$  is an  $(n - n_{J_i}) \times (n - n_{J_i})$  matrix representing the bottom right asterisked block. From the perspective of a linear transformation, the restriction of  $A'$  to  $\mathcal{E}_{J_i^C}$  is simply  $A^C : \mathcal{E}_{J_i^C} \rightarrow \mathcal{E}_{J_i^C}$  which is a linear operator by itself. Now if the algebraic multiplicity of  $\lambda_{J_i}$  is  $k_{J_i}$ , then  $A$  will have a characteristic polynomial of  $p_A(\lambda) = (\lambda_{J_i} - \lambda)^{k_{J_i}} p^-(\lambda)$  where  $p^-(\lambda)$  is another polynomial. By Properties 9.2.3,  $A'$  will also have this same characteristic polynomial  $p_{A'}(\lambda) = (\lambda_{J_i} - \lambda)^{k_{J_i}} p^-(\lambda)$ , and from the structure of  $A'$  derived above, we know that  $p_{A'}(\lambda) = (\lambda_{J_i} - \lambda)^{n_{J_i}} p^C(\lambda)$  with  $p^C(\lambda)$

---

<sup>6</sup> $(P^TAP)^T = P^TA^TP = P^TAP$ .

denoting the characteristic polynomial for  $A^C$ , by repeated cofactor expansion along the first  $n_{J_i}$  columns. Thus comparing the two expressions we know that  $p^C(\lambda) = (\lambda_j - \lambda)^{k_{J_i} - n_{J_i}} p^-(\lambda)$ . Since we assume  $n_{J_i} < k_{J_i}$ ,  $k_{J_i} - n_{J_i} \geq 1$ ,  $p^C(\lambda)$  will contain some  $(\lambda_j - \lambda)$  factor. Hence  $A^C$ , as a square matrix by its own right, must have an eigenvector  $\vec{v}_{\lambda_{J_i}}^{(n_{J_i}+1)}$  in the  $\mathcal{E}_{J_i^c}$  basis corresponding to  $\lambda_{J_i}$  since its geometric multiplicity (in  $A^C$ ) must be at least 1. This  $\vec{v}_{\lambda_{J_i}}^{(n_{J_i}+1)}$  is also an eigenvector of the entire  $A$  matrix and will be linearly independent of  $\vec{v}_{\lambda_{J_i}}^{(1)}, \vec{v}_{\lambda_{J_i}}^{(2)}, \dots, \vec{v}_{\lambda_{J_i}}^{(n_{J_i})}$  because  $\mathcal{E}_{J_i} \oplus \mathcal{E}_{J_i^c}$  is a direct sum (see Definition 6.2.9). Therefore, this shows that the geometric multiplicity of  $\lambda_{J_i}$  in  $A$  is actually  $n_{J_i} + 1$  which contradicts our hypothesis that it is  $n_{J_i}$ , whenever  $n_{J_i} < k_{J_i}$ . (Or we can inductively use this argument until the geometric multiplicity adds up to  $k_{J_i}$ .) Hence the only reasonable conclusion is  $n_{J_i} = k_{J_i}$ .  $\square$

Properties 10.2.3 and 10.2.4 together imply the "if" part of Theorem 10.2.2 and the proof is completed. Recall that the essence of orthogonal diagonalization is to search for an orthonormal basis such that the linear operator or square matrix is diagonal with respect to it. These orthonormal basis vectors are at the same time the eigenvectors of the symmetric matrix following Properties 9.2.2. Hence as a corollary,

**Properties 10.2.5.** A real  $n \times n$  matrix has  $n$  linearly independent eigenvectors that form an orthonormal basis for  $\mathbb{R}^n$  if and only if it is symmetric if and only if it is orthogonally diagonalizable.

**Example 10.2.1.** Carry out orthogonal diagonalization on the matrix

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

*Solution.* First we observe that  $A$  is real and symmetric, and can be orthogonally diagonalized according to Theorem 10.2.2. It can be found that the eigenvectors,

after normalization, are

$$\vec{v}_\lambda = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \text{for } \lambda = 1$$

$$\vec{v}_\lambda = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \text{for } \lambda = 3$$

Hence we can construct

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

So that

$$\begin{aligned} P^T A P &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} = D \end{aligned}$$

□

Short Exercise: Confirm that  $P$  is orthogonal.<sup>7</sup>

---

<sup>7</sup>It is simply checking

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

**Remark**

From Properties 7.2.2 we know that orthogonal diagonalization  $D = P^{-1}AP = P^TAP$  is simply a special case of change of coordinates for a square matrix, where  $P$  is now orthogonal and only pure rotations and reflections are involved (Theorem 10.1.6).

## 10.3 Orthogonal Projections and Spectral Theorem

### 10.3.1 Projections onto a Subspace

Another important concept that also involves orthogonality is *orthogonal projections*. Back in Section 5.2.1 we have defined the projection of a vector  $\vec{v}$  onto another one  $\vec{u}$ , which is implicitly an orthogonal projection since the orthogonal component of  $\vec{v}$  normal to  $\vec{u}$  is removed while the parallel component of  $\vec{v}$  along  $\vec{u}$  is retained, i.e.  $\vec{v} = \overrightarrow{\text{proj}}_u v + (\vec{v} - \overrightarrow{\text{proj}}_u v)$  and  $\vec{u} \cdot (\vec{v} - \overrightarrow{\text{proj}}_u v) = 0$ . Here we can treat  $\vec{u}$  as the one-dimensional subspace generated by itself, and the projection of any vector  $\vec{v}$  onto  $\vec{u}$  is an operation to project the vector onto this subspace of  $\text{span}(\{\vec{u}\})$ . However, before digging deep into orthogonal projections, we have to generalize the notion of projections involving multi-dimensional subspaces first. Given two subspaces  $\mathcal{W}_1$  and  $\mathcal{W}_2$  of a vector space  $\mathcal{V}$  which forms a direct sum  $\mathcal{W}_1 \oplus \mathcal{W}_2 = \mathcal{V}$ , the projection of a vector onto  $\mathcal{W}_1$  along  $\mathcal{W}_2$  is a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  such that for any  $\vec{v} = \vec{w}_1 + \vec{w}_2$  where  $\vec{w}_1 \in \mathcal{W}_1$  and  $\vec{w}_2 \in \mathcal{W}_2$ ,  $T(\vec{v}) = \vec{w}_1$ , so that only the  $\mathcal{W}_1$  component is kept (projected onto) while the  $\mathcal{W}_2$  component is discarded. Clearly,  $\mathcal{R}(T) = \mathcal{W}_1$  and  $\mathcal{N}(T) = \mathcal{W}_2$  is the range and kernel of  $T$ , and  $\mathcal{R}(T) \oplus \mathcal{N}(T) = \mathcal{V}$ . Here we give another equivalent definition of a projection.

**Properties 10.3.1** (Projection). A linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  is a projection if and only if  $T^2 = T$ .

*Proof.* If  $T^2 = T$ , then for any  $\vec{v} \in \mathcal{V}$ , we have  $T^2(\vec{v}) = T(T(\vec{v})) = T(\vec{v})$ , therefore for any vector  $T(\vec{v}) = \vec{w}_1 \in \mathcal{W}_1 = \mathcal{R}(T)$  in the range of  $T$ ,  $T(\vec{w}_1) = \vec{w}_1$ , and the  $\mathcal{W}_1$  component remains unchanged. Now, we derive  $\mathcal{W}_2$  such that we can write  $\vec{v} = \vec{w}_1 + (\vec{v} - \vec{w}_1) = \vec{w}_1 + \vec{w}_2$ , where  $\vec{v} - \vec{w}_1 = \vec{w}_2 \in \mathcal{W}_2$  and  $\mathcal{W}_1 \oplus \mathcal{W}_2 = \mathcal{V}$ <sup>8</sup>, applying  $T$  on both sides gives

$$\begin{aligned} T(\vec{v}) &= T(\vec{w}_1) + T(\vec{v} - \vec{w}_1) && \text{(Linearity)} \\ \vec{w}_1 &= \vec{w}_1 + T(\vec{v} - \vec{w}_1) \\ \implies T(\vec{v} - \vec{w}_1) &= \mathbf{0} \end{aligned}$$

So  $T(\vec{v} - \vec{w}_1) = T(\vec{w}_2) = \mathbf{0}$  for any  $\vec{v} \in \mathcal{V}$  and  $\vec{w}_2 \in \mathcal{W}_2$  as well. This shows that  $\mathcal{W}_2 = \mathcal{N}(T)$  is the kernel of  $T$  and any  $\mathcal{W}_2$  component is annihilated. The "only if" part is simpler: given the effect of projection  $T$  as defined in the first paragraph, for any arbitrary  $\vec{v}$ ,  $T^2(\vec{v}) = T(T(\vec{v})) = T(\vec{w}_1) = \vec{w}_1 = T(\vec{v})$ , so it must be that  $T^2 = T$ .  $\square$

**Example 10.3.1.** Show that the linear operator  $T$  which has a matrix representation of

$$[T] = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

is a projection. Onto/along which subspace this projection is?

*Solution.* By Properties 10.3.1, we have to check if  $[T]^2 = [T]$ . A simple calculation yields

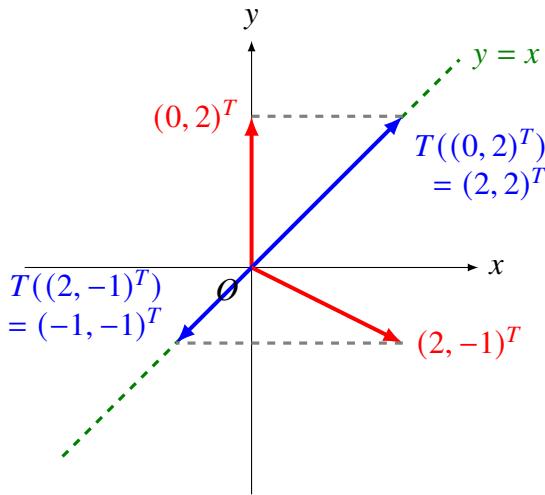
$$\begin{aligned} [T]^2 &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} (0)(0) + (1)(0) & (0)(1) + (1)(1) \\ (0)(0) + (1)(0) & (0)(1) + (1)(1) \end{bmatrix} \end{aligned}$$

---

<sup>8</sup> $\mathcal{W}_2$  can be shown to be a subspace and is linearly independent of  $\mathcal{W}_1$ . The actual projection matrix  $T$  is not only determined by  $\mathcal{W}_1$  but also  $\mathcal{W}_2$ , since the choice of  $\mathcal{W}_2$  will dictate  $\vec{v} - \vec{w}_1 \in \mathcal{W}_2$  and hence  $\vec{w}_1$  too.

$$= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} = [T]$$

So it is indeed a projection. The subspace onto which the projection is carried out simply equals to its range, or equivalently the column space of  $[T]$ , which can be immediately identified as  $\text{span}(\{(1, 1)^T\})$ , that is, the straight line  $y = x$ . Similarly, the projection is along its kernel/null space, which is also easily seen to be  $\text{span}(\{(1, 0)^T\})$ , which is just the  $x$ -axis.



□

### 10.3.2 Orthogonal Projections

After knowing how projections onto a subspace look like, we can now properly generalize the ***orthogonal projection*** (not to be confused with an orthogonal matrix) of a vector onto another vector, to a (possibly multi-dimensional) subspace. For a projection onto a subspace  $\mathcal{W}_1$  along another subspace  $\mathcal{W}_2$  to be an orthogonal projection, the two subspaces have to be the orthogonal complement to each other, such that for any  $\vec{v} = \vec{w}_1 + \vec{w}_2$ ,  $\vec{w}_1 \in \mathcal{W}_1$  and  $\vec{w}_2 \in \mathcal{W}_2$ ,  $\vec{w}_1$  and  $\vec{w}_2$  are orthogonal (i.e.  $\vec{w}_1 \cdot \vec{w}_2 = 0$ ,  $\mathcal{W}_1^\perp = \mathcal{W}_2$  and  $\mathcal{W}_2^\perp = \mathcal{W}_1$ ), and  $T(\vec{v}) = \vec{w}_1$  means that the orthogonal component  $\vec{w}_2 \in \mathcal{W}_2$  normal to  $\mathcal{W}_1$  is removed. As there is only one orthogonal complement for any (finite-dimensional) subspace, the range  $\mathcal{R}(T) = \mathcal{W}_1$  uniquely determines

$\mathcal{N}(T) = \mathcal{W}_2$  and hence  $T$ , where  $\mathcal{R}(T)^\perp = \mathcal{N}(T)$  and  $\mathcal{N}(T)^\perp = \mathcal{R}(T)$ . An equivalent condition of an orthogonal projection is that  $[T] = [T]^T$  is symmetric given we are working over reals.<sup>9</sup>

**Properties 10.3.2** (Orthogonal Projection). A *real, finite-dimensional* linear projection operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  is an orthogonal projection (with respect to the usual dot product) if and only if  $[T] = [T]^T$  in terms of its matrix representation.

*Proof.* The "if" part: We need to show that  $[T] = [T]^T$  implies  $\mathcal{R}(T)^\perp = \mathcal{N}(T)$  (and  $\mathcal{N}(T)^\perp = \mathcal{R}(T)$ )<sup>10</sup>. Let  $\vec{w}_1 \in \mathcal{R}(T)$  and  $\vec{w}_2 \in \mathcal{N}(T)$ , then

$$\begin{aligned}\vec{w}_1 \cdot \vec{w}_2 &= ([T]\vec{w}_1) \cdot \vec{w}_2 && (T(\vec{w}_1) = \vec{w}_1 \text{ by the definition of a projection}) \\ &= ([T]^T\vec{w}_1) \cdot \vec{w}_2 \\ &= \vec{w}_1 \cdot ([T]\vec{w}_2) && (\text{Properties 4.2.3}) \\ &= \vec{w}_1 \cdot \mathbf{0} = 0 && (T(\vec{w}_2) = \mathbf{0} \text{ by the definition of a projection})\end{aligned}$$

So any  $\vec{w}_2 \in \mathcal{N}(T)$  will be orthogonal to all  $\vec{w}_1 \in \mathcal{R}(T)$  and  $\mathcal{N}(T) \subseteq \mathcal{R}(T)^\perp$ . Now let  $\vec{w}_3 \in \mathcal{R}(T)^\perp$  and we want to show  $\vec{w}_3 \in \mathcal{N}(T)$ , i.e.  $T(\vec{w}_3) = \mathbf{0}$ , such that  $\mathcal{R}(T)^\perp \subseteq \mathcal{N}(T)$  and thus  $\mathcal{R}(T)^\perp = \mathcal{N}(T)$ . Consider  $\vec{w}_3 \cdot T^2(\vec{w}_3)$  where  $T^2(\vec{w}_3) \in \mathcal{R}(T)$  since  $T^2 = T$  represents the action of projection (Properties 10.3.1), and therefore  $\vec{w}_3 \cdot T^2(\vec{w}_3) = 0$  as  $\vec{w}_3 \in \mathcal{R}(T)^\perp$ , but also

$$\begin{aligned}\vec{w}_3 \cdot T^2(\vec{w}_3) &= \vec{w}_3 \cdot ([T]^2\vec{w}_3) = \vec{w}_3 \cdot ([T]^T[T]\vec{w}_3) && (\text{By assumption}) \\ &= ([T]\vec{w}_3) \cdot ([T]\vec{w}_3) && (\text{Properties 4.2.3}) \\ &= \| [T]\vec{w}_3 \|^2 = \| T(\vec{w}_3) \|^2\end{aligned}$$

Since this quantity is shown to be zero,  $\| T(\vec{w}_3) \| = 0$  and  $T(\vec{w}_3) = \mathbf{0}$  (see the remark below Properties 4.2.2), and we are done.

The "only if" part: Let  $T$  be an orthogonal projection,  $\vec{u} = \vec{w}_1^{(1)} + \vec{w}_2^{(1)}$  and

<sup>9</sup>Some may wonder why in Properties 10.3.1 we use  $T$  directly but here we circumvent by employing the matrix representation  $[T]$  instead. It is because we haven't defined the "transpose" or "symmetric" equivalent for a linear operator, which will be introduced as the adjoint in Chapter ??.

<sup>10</sup>This two are equivalent as  $\mathcal{V}$  is finite-dimensional.

$\vec{v} = \vec{w}_1^{(2)} + \vec{w}_2^{(2)}$  where  $\vec{w}_1^{(1)}, \vec{w}_1^{(2)} \in \mathcal{W}_1 = \mathcal{R}(T)$  and  $\vec{w}_2^{(1)}, \vec{w}_2^{(2)} \in \mathcal{W}_2 = \mathcal{N}(T)$ . Consider

$$\begin{aligned}\vec{u} \cdot T(\vec{v}) &= (\vec{w}_1^{(1)} + \vec{w}_2^{(1)}) \cdot (\vec{w}_1^{(2)}) = \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} + \vec{w}_2^{(1)} \cdot \vec{w}_1^{(2)} \\ &= \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)}\end{aligned}$$

where  $\vec{w}_2^{(1)} \cdot \vec{w}_1^{(2)} = 0$  because  $\mathcal{N}(T) = \mathcal{R}(T)^\perp$ . Similarly we have  $[T]\vec{u} \cdot \vec{v} = T(\vec{u}) \cdot \vec{v} = \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} + \vec{w}_1^{(1)} \cdot \vec{w}_2^{(2)} = \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} = \vec{u} \cdot T(\vec{v})$ . At the same time

$$\begin{aligned}\vec{u} \cdot T(\vec{v}) &= \vec{u} \cdot ([T]\vec{v}) \\ &= ([T]^T \vec{u}) \cdot \vec{v} \quad (\text{Properties 4.2.3})\end{aligned}$$

This shows that  $([T]\vec{u}) \cdot \vec{v} = ([T]^T \vec{u}) \cdot \vec{v}$  for any  $\vec{u}, \vec{v} \in \mathcal{V}$ . Particularly, since this equality holds for any  $\vec{v} \in \mathcal{V}$ , it is always true that  $[T]\vec{u} = [T]^T \vec{u}$ , which further holds for any  $\vec{u} \in \mathcal{V}$  and we conclude that  $[T] = [T]^T$ .  $\square$

**Example 10.3.2.** Example 10.3.1 has illustrated a projection in  $\mathbb{R}^2$  onto the straight line  $y = x$  along the  $x$ -axis. Continuing from this example, find the unique orthogonal projection onto the same line of  $y = x$  correspondingly.

*Solution.* The matrix representation of  $T$  is derived following Definition 7.1.2, during which the standard coordinates are used:

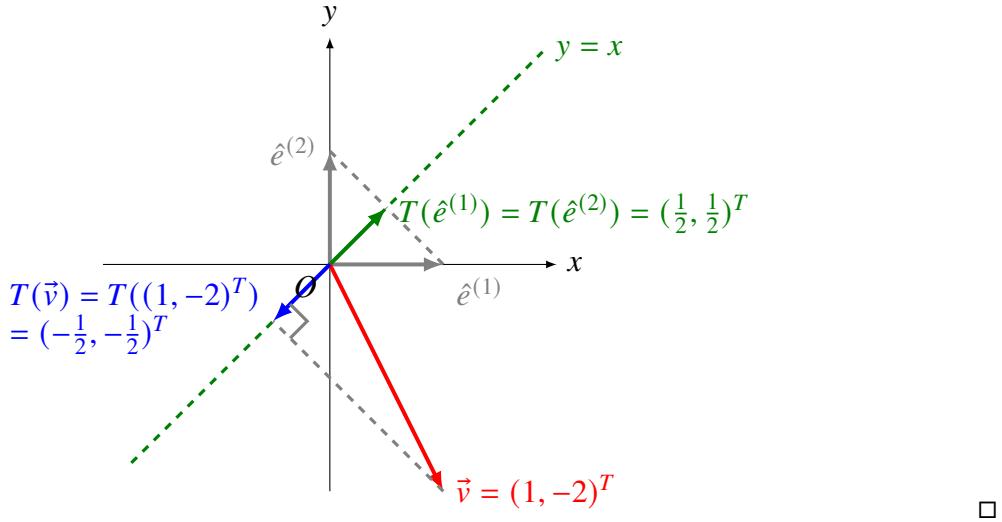
$$[T] = [T(\hat{e}^{(1)}) | T(\hat{e}^{(2)})]$$

whose columns are the outputs of applying the orthogonal projection on the standard unit vectors. Geometrically (see the figure below), it can be easily deduced that  $T(\hat{e}^{(1)}) = T(\hat{e}^{(2)}) = (\frac{1}{2}, \frac{1}{2})^T$ , and hence

$$[T] = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

As a small example, given  $\vec{v} = (1, -2)^T$ , then the orthogonal projection of  $\vec{v}$  onto the straight  $y = x$  is computed as

$$T(\vec{v}) = [T](1, -2)^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$



**Example 10.3.3.** Find the matrix for the orthogonal projection in  $\mathbb{R}^3$  onto the subspace described by the plane  $x + 2y - 4z = 0$ .

*Solution.* In a higher-dimensional scenario, it is beneficial to derive an orthonormal basis for both the range  $\mathcal{R}(T)$  and kernel  $\mathcal{N}(T)$  of the orthogonal projection  $T$  first, and work out the projection transformation in the coordinate system suggested by this basis. We can follow a similar approach as in Example 5.1.1, and one possible basis for the plane is simply  $\{(4, 0, 1)^T, (0, 2, 1)^T\}$ . Subsequently we can apply Gram-Schmidt orthogonalization (Definition 7.2.4) to obtain an orthonormal basis for the plane. We leave to the readers to check that it is  $\{(\frac{4}{\sqrt{17}}, 0, \frac{1}{\sqrt{17}})^T, (-\frac{2}{\sqrt{357}}, \frac{17}{\sqrt{357}}, \frac{8}{\sqrt{357}})^T\}$ . By Properties 7.2.6, we will derive the third vector to complete an orthonormal basis for  $\mathbb{R}^3$  and this third vector will represent the kernel  $\mathcal{N}(T) = \mathcal{R}(T)^\perp$  as the orthogonal complement to the plane. A quick way is to utilize cross product which yields  $(\frac{4}{\sqrt{17}}, 0, \frac{1}{\sqrt{17}})^T \times (-\frac{2}{\sqrt{357}}, \frac{17}{\sqrt{357}}, \frac{8}{\sqrt{357}})^T = (-\frac{1}{\sqrt{21}}, -\frac{2}{\sqrt{21}}, \frac{4}{\sqrt{21}})^T$ . The projection in this basis  $\mathcal{B} = \{(\frac{4}{\sqrt{17}}, 0, \frac{1}{\sqrt{17}})^T, (-\frac{2}{\sqrt{357}}, \frac{17}{\sqrt{357}}, \frac{8}{\sqrt{357}})^T, (-\frac{1}{\sqrt{21}}, -\frac{2}{\sqrt{21}}, \frac{4}{\sqrt{21}})^T\}$  will

have a matrix representation of

$$[T]_B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

as the first two basis vectors are in the range  $\mathcal{R}(T)$  and will be projected into itself, while the last basis vector is in the kernel  $\mathcal{N}(T)$  and hence annihilated. Now we can transform the projection matrix back into the standard basis system according to Properties 7.2.2, where

$$P_B^S = \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix}$$

and

$$\begin{aligned} [T]_S &= (P_S^B)^{-1} [T]_B P_S^B \\ &= P_B^S [T]_B (P_B^S)^T \quad (\text{Properties 10.1.2 for the orthogonal } P \text{ matrix}) \\ &= \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix}^T \\ &= \begin{bmatrix} \frac{20}{21} & -\frac{2}{21} & \frac{4}{21} \\ -\frac{2}{21} & \frac{17}{21} & \frac{8}{21} \\ \frac{4}{21} & \frac{8}{21} & \frac{5}{21} \end{bmatrix} \end{aligned}$$

To cross-check that  $[T]_S$  indeed represents an orthogonal projection, it is obvious that  $[T]_S$  is symmetric and the readers can verify that  $[T]_S^2 = [T]_S$ . Another cross-checking method is to pick a vector in the range (kernel), e.g.  $\vec{v}_R = (6, 1, 2)^T$  and confirm that the projection  $[T]_S \vec{v}_R = [T]_S(6, 1, 2)^T = (6, 1, 2)^T = \vec{v}_R$  leaves it unchanged (vanished).  $\square$

Now we can revisit the projection of a vector  $\vec{v}$  onto another vector  $\vec{u}$  as a special case of orthogonal projection in this section. In Definition 5.2.1, the projection

acting on  $\vec{v}$  is given as  $\overrightarrow{\text{proj}}_u v = \frac{\vec{v} \cdot \vec{u}}{\|\vec{u}\|^2} \vec{u}$ , which can be rewritten as

$$\begin{aligned} T(\vec{v}^T) &= \frac{1}{\|\mathbf{u}\|^2} (\mathbf{v}^T \mathbf{u}) \mathbf{u}^T \\ \implies T(\vec{v}) &= \frac{1}{\|\mathbf{u}\|^2} (\mathbf{v}^T \mathbf{u} \mathbf{u}^T)^T \\ &= \left( \frac{1}{\|\mathbf{u}\|^2} \mathbf{u} \mathbf{u}^T \right) \mathbf{v} \\ &= \left( \frac{\mathbf{u}}{\|\mathbf{u}\|} \left( \frac{\mathbf{u}}{\|\mathbf{u}\|} \right)^T \right) \mathbf{v} = [T] \mathbf{v} \end{aligned}$$

where the orthogonal projection matrix is  $[T] = \frac{\mathbf{u}}{\|\mathbf{u}\|} \left( \frac{\mathbf{u}}{\|\mathbf{u}\|} \right)^T = \hat{\mathbf{u}} \hat{\mathbf{u}}^T$  which is clearly symmetric and can be shown that  $[T]^2 = [T]$ .<sup>11</sup> For an orthogonal projection  $T$  in  $\mathbb{R}^n$  onto some  $r$ -dimensional subspace, which becomes its range  $\mathcal{R}(T)$  with an orthonormal basis  $\{\vec{w}^{(1)}, \vec{w}^{(2)}, \dots, \vec{w}^{(r)}\}$ , complete this orthonormal basis for  $\mathbb{R}^n$  by appending another orthonormal basis  $\{\vec{w}^{(r+1)}, \dots, \vec{w}^{(n)}\}$  of its kernel  $\mathcal{N}(T) = \mathcal{R}(T)^\perp$  (possible due to Properties 7.2.6) just like Example 10.3.3 above. By the same argument in that example, we know that

$$\begin{aligned} [T] &= [\vec{w}^{(1)} | \dots | \vec{w}^{(r)} | \vec{w}^{(r+1)} | \dots | \vec{w}^{(n)}] \begin{bmatrix} I_r & [\mathbf{0}]_{r \times (n-r)} \\ [\mathbf{0}]_{(n-r) \times r} & [\mathbf{0}]_{(n-r) \times (n-r)} \end{bmatrix} \begin{bmatrix} \vec{w}^{(1)T} \\ \vdots \\ \vec{w}^{(r)T} \\ \vec{w}^{(r+1)T} \\ \vdots \\ \vec{w}^{(n)T} \end{bmatrix} \\ &= \hat{\mathbf{W}}^{(1)} \hat{\mathbf{W}}^{(1)T} + \hat{\mathbf{W}}^{(2)} \hat{\mathbf{W}}^{(2)T} + \dots + \hat{\mathbf{W}}^{(r)} \hat{\mathbf{W}}^{(r)T} \end{aligned} \tag{10.1}$$

which generalizes the above form of orthogonal projection matrix in the case of a multi-dimensional range. Furthermore, geometrically, the shortest distance of a point to a subspace is found by the orthogonal projection of that point onto the subspace.

---

<sup>11</sup>  $(\hat{\mathbf{u}} \hat{\mathbf{u}}^T)(\hat{\mathbf{u}} \hat{\mathbf{u}}^T) = \hat{\mathbf{u}}(\hat{\mathbf{u}}^T \hat{\mathbf{u}})\hat{\mathbf{u}}^T = \hat{\mathbf{u}}(1)\hat{\mathbf{u}}^T = \hat{\mathbf{u}} \hat{\mathbf{u}}^T$  as  $\hat{\mathbf{u}}$  is a unit vector and the dot product  $\hat{\mathbf{u}}^T \hat{\mathbf{u}} = \vec{u} \cdot \vec{u} = \|\vec{u}\|^2 = 1$  is its square length.

**Properties 10.3.3.** The shortest distance of a subspace  $\mathcal{W} \subset \mathcal{V}$  to some point  $\vec{v} \in \mathcal{V}$  is achieved by  $\vec{v} - T(\vec{v})$  where  $T$  is the orthogonal projection operator onto  $\mathcal{W}$ .

*Proof.* For any other point  $\vec{w} \in \mathcal{W}$ , we have

$$\begin{aligned}\|\vec{v} - \vec{w}\|^2 &= \|(\vec{v} - T(\vec{v})) + (T(\vec{v}) - \vec{w})\|^2 \\ &= \|\vec{v} - T(\vec{v})\|^2 + \|T(\vec{v}) - \vec{w}\|^2 + 2(\vec{v} - T(\vec{v})) \cdot (T(\vec{v}) - \vec{w})\end{aligned}$$

Note that  $\vec{v} - T(\vec{v}) \in \mathcal{N}(T) = \mathcal{R}(T)^\perp$  and  $T(\vec{v}) - \vec{w} \in \mathcal{R}(T)$  as  $T$  is an orthogonal projection. Therefore,  $(\vec{v} - T(\vec{v})) \cdot (T(\vec{v}) - \vec{w}) = 0$ , and

$$\begin{aligned}\|\vec{v} - \vec{w}\|^2 &= \|\vec{v} - T(\vec{v})\|^2 + \|T(\vec{v}) - \vec{w}\|^2 \\ &> \|\vec{v} - T(\vec{v})\|^2\end{aligned}$$

as  $\vec{w} \neq T(\vec{v})$ ,  $\|T(\vec{v}) - \vec{w}\|^2 > 0$ . □

### 10.3.3 Spectral Theorem

With orthonormal matrices and projections defined, we have come to the most important theorem in this chapter, the **Spectral Theorem**.

**Theorem 10.3.4** (Spectral Theorem). For a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  where  $\mathcal{V}$  is a real, finite( $n$ )-dimensional vector space and the matrix representation of  $T$  is symmetric, i.e.  $[T] = [T]^T$ , denote its eigenvalues by  $\lambda_j$ ,  $j = 1, 2, \dots, n$  (counting repeated ones). Assume there are  $k$  distinct eigenvalues out of them. For every such a distinct eigenvalue, collect all the indices  $j$  where  $\lambda_j$  points to that same eigenvalue, let's say,  $\lambda_{J_i}$ , and put them in a set  $J_i$ ,  $i = 1, 2, \dots, k$ . Let the associated eigenspace be  $\mathcal{E}_{J_i}$  for the  $k$  sets of  $J_i$  and denote the orthogonal projection onto  $\mathcal{E}_{J_i}$  as  $T_{J_i}$ , then we have:

- (a)  $\mathcal{V} = \mathcal{E}_{J_1} \oplus \mathcal{E}_{J_2} \oplus \dots \oplus \mathcal{E}_{J_k} = \bigoplus_{i=1}^k \mathcal{E}_{J_i}$ ;
- (b)  $(\bigoplus_{i \in I} \mathcal{E}_{J_i})^\perp = \bigoplus_{i \notin I} \mathcal{E}_{J_i}$  where  $I$  is a set containing some indices

between 1 and  $k$ ;

- (c)  $T_{J_i}T_{J_{i'}} = \begin{cases} T_{J_i} & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases};$
- (d)  $I = T_{J_1} + T_{J_2} + \cdots + T_{J_k} = \sum_{i=1}^k T_{J_i}$ ; and
- (e)  $T = \lambda_{J_1}T_{J_1} + \lambda_{J_2}T_{J_2} + \cdots + \lambda_{J_k}T_{J_k} = \sum_{i=1}^k \lambda_{J_i}T_{J_i}.$

*Proof.* (a) By Theorem 10.2.2,  $[T]$  is (orthogonally) diagonalizable. The prior discussion in Section 9.2.3 has shown that

$$\bigoplus_{i=1}^k \mathcal{E}_{J_i} = \bigoplus_{j=1}^n \mathcal{E}_j = \mathcal{V}$$

where  $\mathcal{E}_{J_i} = \bigoplus_{j \in J_i} \mathcal{E}_j$  is the direct sum of one-dimensional subspaces generated by each of the linearly independent eigenvector corresponding to the same eigenvalue  $\lambda_{J_i}$ .

- (b) This is a consequence of applying Properties 7.2.6 on (a) where  $\bigoplus_{i \in \{I\}} \mathcal{E}_{J_i}$  and  $\bigoplus_{i \notin \{I\}} \mathcal{E}_{J_i}$  are derived from the two mutually exclusive portions of the orthonormal basis for  $\mathcal{V}$  generated by  $n$  orthonormal eigenvectors of the symmetric  $[T]$ , available due to the implication of Theorem 10.2.2.
- (c)  $T_{J_i}^2 = T_{J_i}$  by the definition of a projection (Properties 10.3.1). Also, we can express any  $\vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k}$  by (a), then  $T_{J_{i'}}(\vec{v}) = \vec{v}_{J_{i'}}$  due to the definition of an orthogonal projection and (b), as  $\mathcal{E}_{J_{i'}}^\perp = \bigoplus_{i \neq i'} \mathcal{E}_{J_i}$ . The same logic means that  $T_{J_i}(T_{J_{i'}}(\vec{v})) = T_{J_i}(\vec{v}_{J_{i'}}) = \mathbf{0}$  if  $i \neq i'$ . Since this holds for any  $\vec{v}$ ,  $T_{J_i}T_{J_{i'}} = 0$ .
- (d) Again for any  $\vec{v}$  we can rewrite it as  $\vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k}$  where  $\vec{v}_{J_i} \in \mathcal{E}_{J_i}$ . Since  $T_{J_i}$  is an orthogonal projection onto  $\mathcal{E}_{J_i}$  and also  $\mathcal{R}(T_{J_i})^\perp = \mathcal{E}_{J_{i'}}^\perp = \bigoplus_{i \neq i'} \mathcal{E}_{J_i} = \mathcal{N}(T_{J_i})$  as derived in (c),  $T_{J_i}(\vec{v}) = \vec{v}_{J_i}$ . Thus  $I\vec{v} = \vec{v} = T_{J_1}(\vec{v}) + T_{J_2}(\vec{v}) + \cdots + T_{J_k}(\vec{v}) = (T_{J_1} + T_{J_2} + \cdots + T_{J_k})(\vec{v})$  and it must be that  $I = T_{J_1} + T_{J_2} + \cdots + T_{J_k}$ .

(e) As usual,  $\vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k}$  and

$$\begin{aligned} T(\vec{v}) &= T(\vec{v}_{J_1}) + T\vec{v}_{J_2} + \cdots + T(\vec{v}_{J_k}) \\ &= \lambda_{J_1}\vec{v}_{J_1} + \lambda_{J_2}\vec{v}_{J_2} + \cdots + \lambda_{J_k}\vec{v}_{J_k} \quad (\text{Definition 9.1.1 for eigenvectors}) \\ &= \lambda_{J_1}T_{J_1}(\vec{v}) + \lambda_{J_2}T_{J_2}(\vec{v}) + \cdots + \lambda_{J_k}T_{J_k}(\vec{v}) \quad (\text{Just as (d)}) \\ &= (\lambda_{J_1}T_{J_1} + \lambda_{J_2}T_{J_2} + \cdots + \lambda_{J_k}T_{J_k})(\vec{v}) \end{aligned}$$

and thus the desired relation follows.

□

The set of eigenvalues  $\{\lambda_{J_i}\}$ ,  $i = 1, 2, \dots, k$  is known as the *spectrum* of  $T$ , hence comes the name of the theorem. The expression in (d) is called the *resolution of the identity* induced by  $T$ , and that in (e) is referred to as the *spectral decomposition* of  $T$ . The implications of these two parts of the Spectral Theorem are as follows: The resolution of the identity in (d) means that any vector  $\vec{v}$  can be decomposed into components in the orthogonal coordinate system derived from the orthonormal eigenvectors of  $T$ , where the corresponding coordinates are given by simply projecting onto each of these eigenvectors; The spectral decomposition in (e) means that any linear operator  $T$  that happens to have a symmetric matrix representation, when applied on a vector  $\vec{v}$ , can be regarded to be first decomposing  $\vec{v}$  into components in the orthogonal coordinate system as suggested by (d), and then scaling each of these components according to the spectrum (eigenvalues), which represents the "intensity" of  $T$  in the respective eigenspace. Since only projections and scalings are involved, it implies that  $T$  constitutes no rotation. Finally, notice that an orthogonal projection is a special case of (e) where the spectrum is consisted of 1 (range) and 0 (kernel) only.

Short Exercise: Verify statements (d) and (e) in the Spectral Theorem for the symmetric matrix in Example 10.2.1.<sup>12</sup>

---

<sup>12</sup>We will just show (d) here and leave (e) to the readers. By Equation 10.1, R.H.S. of (d) reads

$$T_{J_1} + T_{J_2} = (\mathbf{v}_{J_1}^{(1)T} \mathbf{v}_{J_1}^{(1)T} + \mathbf{v}_{J_1}^{(2)T} \mathbf{v}_{J_1}^{(2)T}) + \mathbf{v}_{J_2}^{(1)T} \mathbf{v}_{J_2}^{(1)T}$$

## 10.4 Normal Matrices and Unitary Diagonalization

### 10.4.1 Unitary and Normal Matrices

We now generalize the ideas of orthogonal matrices/diagonalization to complex vector space. The complex counterpart of orthogonal matrices are known as ***unitary matrices***.

**Definition 10.4.1.** A complex matrix  $A$  is said to be unitary if

$$A^*A = AA^* = I$$

where the superscript  $*$  denotes conjugate transpose (Definition 8.2.4).

Along the same line, the complex equivalent of orthogonal diagonalization is ***unitary diagonalization***.

**Definition 10.4.2.** A complex square matrix  $A$  is said to be unitarily diagonalizable if there is a unitary matrix  $U$  such that  $U^*AU = D$  yields a complex diagonal matrix whose non-zero entries are the eigenvalues of  $A$ .

This can be compared to Definition 10.2.1. In addition, since we know that the availability of orthogonal diagonalization depends on whether the matrix is symmetric, we may want to know if there is a comparable criterion for unitary diagonalization. A reasonable guess is the matrix  $A$  being *Hermitian* (Definition 8.2.5) such that  $A^* = A$  since it is the direct complex analog of a symmetric

$$\begin{aligned} &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

where the set indices  $J_1$  and  $J_2$  indicate the eigenvectors for the eigenvalues of  $\lambda = 1$  and  $3$  respectively.

matrix. However, the criterion is actually looser where the matrix is required to be just *normal*.

**Definition 10.4.3** (Normal Matrix). A matrix  $A$  is known as normal whenever

$$A^*A = AA^*$$

which may or may not be equal to the identity.

So normality is a less strict condition than being unitary. It is also very easy to see that a Hermitian matrix is always normal. In the next part, we are going to derive how unitary diagonalization manifests and why it is enabled by normality.

## 10.4.2 Unitary Diagonalization

As in Properties 9.2.2, the unitary matrix  $U$  that diagonalizes some  $n \times n$  complex, normal matrix  $A$  is formed by the  $n$  linearly independent eigenvectors of  $A$  arranged in columns, which also like in Definition 10.2.1 these eigenvectors have to be orthonormal. Note that orthogonality is now defined with respect to the complex dot product which has a complex conjugate applied on the second input vector. To show the equivalence between a matrix being unitary diagonalizable and its normality, we can of course follow the similar steps in the proof of Theorem 10.2.2 that uses Properties 10.2.3 and 10.2.4. However, an alternative approach of utilizing the **Schur's Triangularization Theorem** will be adopted here.

**Theorem 10.4.4** (Schur's Triangularization Theorem). For any square complex (real) matrix  $A$ , there exists a unitary (orthogonal) matrix  $U$  ( $P$ ) such that  $S = U^*AU$  ( $P^TAP$ ) is upper-triangular (where the eigenvalues of  $A$ , counting algebraic multiplicity, are located along the main diagonal of  $S$ ) if the characteristic polynomial of  $A$  splits over  $\mathbb{C}$  ( $\mathbb{R}$ )).

Notice the requirement in the end of the theorem (see Footnote 6 of Chapter 9). Recall that every polynomial splits over  $\mathbb{C}$ . So in other words, every square

matrix is *unitarily similar* to an upper-triangular matrix. (But not always orthogonally similar to a real diagonal matrix as the characteristic polynomial may not split over  $\mathbb{R}$ <sup>13</sup>.)

*Proof.* We will use mathematical induction on the size of  $A$  to show the theorem for the complex case. If  $A$  is  $1 \times 1$  then the result is automatic. Then assume  $A$  is  $n \times n$  and the theorem is true for all  $r \times r$  matrices where  $r < n$ , particularly for  $r = n - 1$ . As the characteristic polynomial splits, there is always an available (complex) root to the polynomial as one of the eigenvalues of  $A$ , let's say  $\lambda_1$ , and a corresponding eigenvector  $\vec{v}_\lambda^{(1)}$ . Use the Gram-Schmidt process (with respect to the complex dot product now) to generate a other set of  $n - 1$  vectors  $\vec{w}^{(2)}, \dots, \vec{w}^{(n)}$  to complete an orthonormal basis for  $\mathbb{C}^n$ . The matrix formed by arranging these orthonormal basis vectors (normalized into unit length) in columns,

$$Q = \left[ \vec{v}_\lambda^{(1)} | \vec{w}^{(2)} | \dots | \vec{w}^{(n)} \right] = \left[ \vec{v}_\lambda^{(1)} | W \right]$$

is Hermitian and unitary,  $Q^{-1} = Q^* = Q$ . A change of coordinates on  $A$  by  $Q$  as in Properties 7.2.2 is

$$\begin{aligned} Q^{-1}AQ &= Q^*AQ = Q^*A \left[ \vec{v}_\lambda^{(1)} | W \right] \\ &= Q^* \left[ A\vec{v}_\lambda^{(1)} | AW \right] \\ &= Q^* \left[ \lambda_1 \vec{v}_\lambda^{(1)} | AW \right] \quad (\text{Definition 9.1.1} \\ &\quad \text{for an eigenvalue/eigenvector}) \\ &= \left[ \frac{\vec{v}_\lambda^{(1)*}}{W^*} \right] \left[ \lambda_1 \vec{v}_\lambda^{(1)} | AW \right] \quad (Q \text{ is Hermitian}) \\ &= \left[ \left[ \frac{\vec{v}_\lambda^{(1)*}}{W^*} \right] \lambda_1 \vec{v}_\lambda^{(1)} | \left[ \frac{\vec{v}_\lambda^{(1)*}}{W^*} \right] AW \right] \\ &= \left[ \begin{array}{cc} \lambda_1 & \vec{v}_\lambda^{(1)*} AW \\ \mathbf{0} & W^* AW \end{array} \right] = \left[ \begin{array}{cc} \lambda_1 & \vec{v}_\lambda^{(1)*} AW \\ \mathbf{0} & W' \end{array} \right] \end{aligned}$$

---

<sup>13</sup>which is equivalent to all eigenvalues being real.

which is expressed in the form of a  $2 \times 2$  block matrix. The leftmost column takes its form because  $\vec{v}_\lambda^{(1)}$  is now a unit vector and hence  $\mathbf{v}_\lambda^{(1)*} \mathbf{v}_\lambda^{(1)} = 1$  and it is orthogonal to other vectors in  $W$  by construct. Now we can use the induction hypothesis on the bottom right block  $W' = W^* AW$  which is an  $(n-1) \times (n-1)$  matrix, such that there exists a unitary matrix  $R$  so that  $R^* W' R$  is upper-triangular. Consider

$$U = Q \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix}$$

This matrix is unitary (Definition 10.4.1), as

$$\begin{aligned} U^* U &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} Q^* Q \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} (I) \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \quad (\text{Q is unitary}) \\ &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* R \end{bmatrix} = I \quad (\text{R is unitary}) \end{aligned}$$

Subsequently,

$$\begin{aligned} S = U^* A U &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} Q^* A Q \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} A W \\ \mathbf{0} & W' \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \quad (\text{from above}) \\ &= \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} A W \\ \mathbf{0} & R^* W' \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} A W R \\ \mathbf{0} & R^* W' R \end{bmatrix} \end{aligned}$$

which is an upper-triangular matrix as desired since the induction hypothesis demands the bottom-right block  $R^* W' R$  to be upper-triangular too. And therefore, the diagonal elements of  $S$  are exactly its eigenvalues, which will be the same as those of  $A$  since they are unitarily similar.  $\square$

With Schur's Triangularization Theorem, the fact that normal matrices are unitarily diagonalizable is attainable.

**Theorem 10.4.5.** A square matrix  $A$  can be unitarily diagonalized if and only if  $A$  is normal.

We will prove the "only if" part first. This is a bit harder than that in Theorem 10.2.2:

$$A^*A = (UDU^*)^*(UDU^*) = UD^*(U^*U)DU^* = UD^*DU^*$$

and

$$AA^* = (UDU^*)(UDU^*)^* = UD(U^*U)DU^*U^* = UDD^*U^*$$

But  $D^*D = DD^*$  since they are diagonal and commute, so  $A^*A = AA^*$ . The "if" part is also a bit more tricky. By Theorem 10.4.4, we can always write  $S = U^*AU$  where  $S$  is upper-triangular (and  $U$  is unitary), and hence  $S^* = U^*A^*U$  is lower-triangular. Notice that

$$\begin{aligned} S^*S &= U^*A^*UU^*AU = U^*A^*AU && (U \text{ is unitary}) \\ &= U^*AA^*U && (A \text{ is normal}) \\ &= (U^*AU)(U^*A^*U) = SS^* \end{aligned}$$

With

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ 0 & s_{22} & s_{23} & \cdots & s_{2n} \\ 0 & 0 & s_{33} & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_{nn} \end{bmatrix} \quad \text{and } S^* = \begin{bmatrix} \overline{s_{11}} & 0 & 0 & \cdots & 0 \\ \overline{s_{12}} & \overline{s_{22}} & 0 & \cdots & 0 \\ \overline{s_{13}} & \overline{s_{23}} & \overline{s_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{s_{1n}} & \overline{s_{2n}} & \overline{s_{3n}} & \cdots & \overline{s_{nn}} \end{bmatrix}$$

then consider

$$\begin{aligned} (S^*S)_{11} &= (SS^*)_{11} \\ \overline{s_{11}}s_{11} &= s_{11}\overline{s_{11}} + s_{12}\overline{s_{12}} + \cdots + s_{1n}\overline{s_{1n}} \\ 0 &= |s_{12}|^2 + \cdots + |s_{1n}|^2 \geq 0 \end{aligned}$$

we must have  $s_{12} = \cdots = s_{1n} = 0$ . Similarly, consider  $(S^*S)_{22} = (SS^*)_{22}, \dots, (S^*S)_{nn} = (SS^*)_{nn}$  in order, we conclude that all off-diagonal elements  $s_{ij} = 0$  when  $i < j$  are zero and thus  $S$  is diagonal. Hence  $A$  is actually unitarily diagonalized into  $D = S = U^*AU$  by the matrix  $U$ .

**Example 10.4.1.** Unitarily diagonalize the following matrix.

$$A = \begin{bmatrix} 1 & 1+i \\ 1-i & 2 \end{bmatrix}$$

*Solution.* This can be seen as a Hermitian matrix, and therefore unitary diagonalization is possible by Theorem 10.4.5. The characteristic equation is

$$\begin{aligned} (1 - \lambda)(2 - \lambda) - (1 + i)(1 - i) &= 0 \\ (2 - 3\lambda + \lambda^2) - 2 &= 0 \\ \lambda^2 - 3\lambda &= 0 \\ \lambda &= 0 \text{ or } 3 \end{aligned}$$

The eigenvalues being real for an Hermitian matrix is not a mere coincidence and will be proved afterwards. In fact, it is an extension of the fact that the eigenvalues of any symmetric matrix are all reals. The eigenvectors can be computed to be  $(-1 - i, 1)^T$  for  $\lambda = 0$  and  $(1 + i, 2)^T$  for  $\lambda = 3$ . After normalization, they are  $\frac{1}{\sqrt{3}}(-1 - i, 1)^T$  and  $\frac{1}{\sqrt{6}}(1 + i, 2)^T$  respectively. Now define

$$U = \begin{bmatrix} \frac{-1-i}{\sqrt{3}} & \frac{1+i}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}$$

Then the unitary diagonalization has the form of

$$U^*AU = D$$

$$\begin{bmatrix} \frac{-1+i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1-i}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 1 & 1+i \\ 1-i & 2 \end{bmatrix} \begin{bmatrix} \frac{-1-i}{\sqrt{3}} & \frac{1+i}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix}$$

□

**Properties 10.4.6.** The eigenvalues of any Hermitian matrix  $A = A^*$  must be real.

*Proof.* Consider  $\vec{v} \cdot (A\vec{v})$  where  $\vec{v}$  is a complex eigenvector of  $A$  and the complex dot product is used in place. Then

$$\begin{aligned}\vec{v} \cdot (A\vec{v}) &= \vec{v} \cdot (\lambda\vec{v}) && \text{(Definition 9.1.1)} \\ &= \bar{\lambda}(\vec{v} \cdot \vec{v}) && \text{(Properties 8.2.3)} \\ &= \bar{\lambda}\|\vec{v}\|^2\end{aligned}$$

But also

$$\begin{aligned}\vec{v} \cdot (A\vec{v}) &= (A^*\vec{v}) \cdot \vec{v} && \text{(Properties 8.2.7)} \\ &= (A\vec{v}) \cdot \vec{v} && (A \text{ is Hermitian}) \\ &= (\lambda\vec{v}) \cdot \vec{v} && \text{(Definition 9.1.1)} \\ &= \lambda(\vec{v} \cdot \vec{v}) && \text{(Properties 8.2.3)} \\ &= \lambda\|\vec{v}\|^2\end{aligned}$$

Therefore,  $\bar{\lambda}\|\vec{v}\|^2 = \lambda\|\vec{v}\|^2$ , and since  $\vec{v} \neq \mathbf{0}$ ,  $\|\vec{v}\| \neq 0$ , the eigenvalue  $\bar{\lambda} = \lambda$  has to be real.  $\square$

**Example 10.4.2.** Check

$$A = \begin{bmatrix} \frac{5}{3} - \frac{1}{3}\iota & \frac{1}{3} + \iota \\ -1 + \frac{1}{3}\iota & \frac{4}{3} + \frac{1}{3}\iota \end{bmatrix}$$

is normal and carry out unitary diagonalization on it.

*Solution.* First,

$$A^*A = \begin{bmatrix} \frac{5}{3} + \frac{1}{3}\iota & -1 - \frac{1}{3}\iota \\ \frac{1}{3} - \iota & \frac{4}{3} - \frac{1}{3}\iota \end{bmatrix} \begin{bmatrix} \frac{5}{3} - \frac{1}{3}\iota & \frac{1}{3} + \iota \\ -1 + \frac{1}{3}\iota & \frac{4}{3} + \frac{1}{3}\iota \end{bmatrix} = \begin{bmatrix} 4 & -1 + \iota \\ -1 - \iota & 3 \end{bmatrix}$$

and

$$AA^* = \begin{bmatrix} \frac{5}{3} - \frac{1}{3}\iota & \frac{1}{3} + \iota \\ -1 + \frac{1}{3}\iota & \frac{4}{3} + \frac{1}{3}\iota \end{bmatrix} \begin{bmatrix} \frac{5}{3} + \frac{1}{3}\iota & -1 - \frac{1}{3}\iota \\ \frac{1}{3} - \iota & \frac{4}{3} - \frac{1}{3}\iota \end{bmatrix} = \begin{bmatrix} 4 & -1 + \iota \\ -1 - \iota & 3 \end{bmatrix} = A^*A$$

So normality holds. Now the characteristic equation is

$$\begin{aligned} \left(\frac{5}{3} - \frac{1}{3}\iota - \lambda\right)\left(\frac{4}{3} + \frac{1}{3}\iota - \lambda\right) - \left(-1 + \frac{1}{3}\iota\right)\left(\frac{1}{3} + \iota\right) &= 0 \\ \lambda^2 - 3\lambda + (3 + \iota) &= 0 \end{aligned}$$

which can be checked to have the complex roots of  $\lambda_1 = 1 + \iota$  and  $\lambda_2 = 2 - \iota$  as the two eigenvalues. The corresponding orthonormal eigenvectors are  $\vec{v}_{\lambda_1} = (\frac{1}{\sqrt{3}}, \frac{1+\iota}{\sqrt{3}})^T$  and  $\vec{v}_{\lambda_2} = (\frac{-1+\iota}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$ , and thus the unitary diagonalization reads

$$\begin{aligned} D &= U^*AU \\ &= \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1-\iota}{\sqrt{3}} \\ \frac{-1-\iota}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{5}{3} - \frac{1}{3}\iota & \frac{1}{3} + \iota \\ -1 + \frac{1}{3}\iota & \frac{4}{3} + \frac{1}{3}\iota \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-1+\iota}{\sqrt{3}} \\ \frac{1+\iota}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \\ &= \begin{bmatrix} 1 + \iota & 0 \\ 0 & 2 - \iota \end{bmatrix} \end{aligned}$$

where

$$U = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-1+\iota}{\sqrt{3}} \\ \frac{1+\iota}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

□

## 10.5 Python Programming

Orthogonal diagonalization is also performed by the `diagonalize` method in `sympy` as long as the matrix is symmetric. Let's try this with Example 10.2.1:

```
import numpy as np
import sympy

A = np.array([[1., 0., 0.],
              [0., 2., 1.],
              [0., 1., 2.]])
```

```
A_sympy = sympy.Matrix(A)
print(A_sympy.is_symmetric())
P, D = A_sympy.diagonalize()
print(P, D)
```

which returns `true` and

```
Matrix([[1.0000, 0, 0],
       [0, 0.7071, -0.7071],
       [0, -0.7071, -0.7071]])
Matrix([[1.0000, 0, 0],
       [0, 1.0000, 0],
       [0, 0, 3.0000]])
```

as expected. We can confirm the  $P$  matrix is orthogonal by

```
print(np.allclose(np.array(P.T @ P, dtype=float), np.identity(3))) # True
```

which is slight complicated since there will be numerical errors when computing  $P$  and  $P^T P$  so we need to first convert the matrix product to `np.array` and use `np.allclose` which does not check exact but close equality against  $I$  over all entries. The same ideas apply for unitary diagonalization and let's use Example 10.4.2 to demonstrate. First, we check normality by

```
A = np.array([[5/3 - (1/3)*1j, 1/3 + 1j],
             [-1 + (1/3)*1j, 4/3 + (1/3)*1j]])

print(np.allclose(np.conjugate(A).T @ A, A @ np.conjugate(A).T))
```

which outputs `true`. Next,

```
A_sympy = sympy.Matrix(A)

U, D = A_sympy.diagonalize()
print(U, D)
print(np.allclose(np.array(U.H @ U, dtype=complex), np.identity(2)))
```

gives

```
Matrix([[-0.4898 - 0.6531*I, 0.0816 + 0.571*I],
       [-0.0816 + 0.5715*I, -0.4898 + 0.6531*I]])
```

```
Matrix([[2.0 - 1.0*I, 0],
       [0, 1.0 + 1.0*I]])
```

and `true` too.

## 10.6 Exercises

**Exercise 10.1** Determine if the following matrices are orthogonal. If so, also determine if they represent rotation or reflection.

$$(a) \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{\sqrt{3}}{2\sqrt{6}} & \frac{1}{2} & -\frac{\sqrt{3}}{2\sqrt{6}} \\ -\frac{1}{2\sqrt{2}} & \frac{\sqrt{3}}{2} & \frac{1}{2\sqrt{2}} \end{bmatrix}$$

$$(b) \begin{bmatrix} -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{\sqrt{3}}{4} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{4} & \frac{3}{4} & \frac{1}{2} \end{bmatrix}$$

$$(c) \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ 1 & -1 & -\frac{1}{2} \\ 1 & 1 & 2 \end{bmatrix}$$

$$(d) \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix}$$

**Exercise 10.2** Represent the following operations on a three-dimensional  $xyz$  coordinate system by a transitional matrix  $P$ . State clearly the relationship between the vectors in old and new coordinate system ( $\vec{v}_0$  and  $\vec{v}_n$ ), as well as the matrix  $P$ .

- (a) Rotation about the z-axis (x-axis and y-axis revolving around the z-axis) counter-clockwise by 45 degrees,

- (b) Reflection of the x-axis across the y-z plane and subsequently rotation about the intermediate y-axis counter-clockwise by 30 degrees,
- (c) Rotation about the y-axis clockwise by 45 degrees, and then rotation about the new z-axis counter-clockwise by 60 degrees.

It is emphasized that the order of operations is important.

**Exercise 10.3** Argue that

$$P = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}$$

is a transition matrix that represents a reflection about the infinitely long straight line with an angle of  $\frac{\theta}{2}$  passing through the origin on the  $xy$  plane.

**Exercise 10.4** For the symmetric matrix

$$\begin{bmatrix} 1 & a & a \\ a & 5 & 3 \\ a & 3 & 5 \end{bmatrix}$$

It is given that one of the eigenvalues is zero and the product of other two eigenvalues is 18. Using the knowledge that the trace and characteristic polynomial are invariants, i.e. remains unchanged after diagonalization. Find

- (a) The two remaining eigenvalues,
- (b) The possible values of  $a$ ,
- (c) For every possible case, find the eigenvector corresponding to the eigenvalue of zero, then carry out orthogonal diagonalization.

**Exercise 10.5** Orthogonally diagonalize the following matrix, if possible.

$$(a) \begin{bmatrix} 7 & 8 & 10 \\ 1 & 6 & 3 \\ 5 & 1 & 2 \end{bmatrix}$$

(b)  $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(c)  $\begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}$

**Exercise 10.6** Find the eigenvalues and corresponding eigenspaces for the matrix  $A = I_n - \mathbf{u}\mathbf{u}^T$  where  $\vec{u} \in \mathbb{R}^n$  is a unit vector such that  $\|\mathbf{u}\| = 1$ . Hint: use the Spectral Theorem and note that  $\vec{u}$  alone is linearly independent and can be completed to an orthonormal basis.

**Exercise 10.7** Find the orthogonal projection operator in  $\mathbb{R}^4$  on the subspace intersected by the two three-dimensional hyperplanes  $x + 2y - 3z + 4w = 0$  and  $3x - y + z - 2w = 0$ .

**Exercise 10.8** Check the Spectral Theorem (Theorem 10.3.4) for the symmetric matrix below:

$$A = \begin{bmatrix} \frac{5}{3} & 0 & -\frac{\sqrt{2}}{3} \\ 0 & 2 & 0 \\ -\frac{\sqrt{2}}{3} & 0 & \frac{4}{3} \end{bmatrix}$$

**Exercise 10.9** Show that the matrix below is Hermitian and unitarily diagonalize it.

$$A = \begin{bmatrix} 1 & -i & 0 \\ i & 2 & 1+i \\ 0 & 1-i & 3 \end{bmatrix}$$

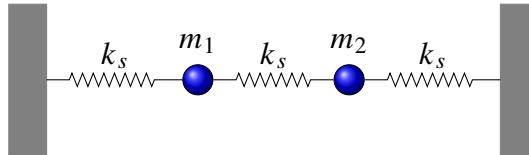
**Exercise 10.10** Hooke's law states that the force acting on a mass by spring is given by  $F = -kx$  where  $k$  is the spring constant and  $x$  is the displacement (extension or compression) from equilibrium position. By considering Newton's second law,  $F = ma$ , we have

$$ma = m \frac{d^2x}{dt^2} = -kx \text{ and hence } x'' = \frac{d^2x}{dt^2} = -\frac{k}{m}x$$

The general solution is

$$x = C \cos(\omega t - \theta)$$

where  $\omega = \sqrt{\frac{k}{m}}$  and  $C, \theta$  are some arbitrary constants to be determined from the initial condition. Consider the situation shown in the figure below. Find the two equations describing motions of the two masses  $m_1, m_2$  respectively, in terms of the displacements  $\mathbf{x} = (x_1, x_2)^T$ , and re-written them into matrix form. Carry out orthogonal diagonalization to simplify the equations and find the general solutions for the motions.



## Chapter 11

# Quadratic Forms

---

Symmetric matrices often see their appearance in the areas of Geometry, Statistics, Physics, and more, taking the form of *quadratic forms*. They can be used to describe a family of geometric shapes known as *conic sections*, including ellipses and hyperbola. A more common application of quadratic forms in Earth Sciences will be obtaining the *covariance matrix* between different variables, eventually leading to *Principal Component Analysis (PCA)* which breaks down the variables into isolated modes that explain the spread of their distributions. In Atmospheric Science, it is more commonly known as *Empirical Orthogonal Functions (EOF)* and has been widely used to analyze prominent, large-scale climate patterns like *El Niño–Southern Oscillation (ENSO)*.

## 11.1 Mathematical and Geometric Ideas of Quadratic Forms

### 11.1.1 Definition of Quadratic Forms

The word *quadratic* is commonly associated to *quadratic equations* in the form of  $y = ax^2 + bx + c$ . **Quadratic forms** are the generalization of quadratic equations when there are multiple variables  $x_1, x_2, x_3, \dots$ : the possible quadratic forms will be made up of the usual quadratic terms  $x_1^2, x_2^2, x_3^2, \dots$ , as well as

the *cross-product terms*  $x_p x_q$ ,  $p \neq q$ . The usual quadratic terms then can be seen as just another kind of cross-product terms when  $p = q$ . We will limit our discussion to real vector spaces.

**Definition 11.1.1.** Real quadratic forms of multiple variables  $\vec{x} = (x_1, x_2, x_3, \dots)^T$  has a structure of

$$Q(\vec{x}) = \sum_{p=1}^n \sum_{q=1}^n \beta_{pq} x_p x_q$$

Note that it is a real scalar.

For examples, in a two variables situation,  $x^2 + 3xy + y^2$ ,  $3x^2 - 4xy$  are quadratic forms, while  $x^2 + y$ ,  $xy + xy^2$  are not. Notice that  $x_p x_q$  and  $x_q x_p$  are actually the same term, and we will replace  $\beta_{pq}$  or  $\beta_{qp}$  by  $\frac{1}{2}(\beta_{pq} + \beta_{qp})$  as a single coefficient for both of them.

**Properties 11.1.2.** All quadratic forms like those in Definition 11.1.1 for any real vector  $\vec{x} = (x_1, x_2, x_3, \dots)^T \in \mathcal{V}$  can be expressed as

$$Q : \mathcal{V} \rightarrow \mathbb{R}, Q(\vec{x}) = \vec{x}^T B \vec{x}$$

where  $B$  is symmetric and has the form of

$$\begin{bmatrix} \beta_{11} & \frac{1}{2}(\beta_{12} + \beta_{21}) & \frac{1}{2}(\beta_{13} + \beta_{31}) & \cdots \\ \frac{1}{2}(\beta_{12} + \beta_{21}) & \beta_{22} & \frac{1}{2}(\beta_{23} + \beta_{32}) & \\ \frac{1}{2}(\beta_{13} + \beta_{31}) & \frac{1}{2}(\beta_{23} + \beta_{32}) & \beta_{33} & \\ \vdots & & & \ddots \end{bmatrix}$$

The readers can check that the above matrix expression indeed leads to the

desired quadratic form by a direct expansion.<sup>1</sup> In the same essence, we can always express any quadratic form using the symmetric part of a matrix. For instance, the quadratic form  $x^2 - 2xy + 3y^2$  can be rewritten as

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Short Exercise: Verify the quadratic form by expanding it.<sup>2</sup>

Even if we are given  $\vec{x}^T A \vec{x}$  where  $A$  is not symmetric to begin with, we can extract the symmetric part of  $A$  (see Exercise 2.8), that is,  $B = \frac{1}{2}(A + A^T)$ .

<sup>1</sup>

$$\begin{aligned} & \begin{bmatrix} x_1 & x_2 & x_3 & \cdots \end{bmatrix} \begin{bmatrix} \beta_{11} & \frac{1}{2}(\beta_{12} + \beta_{21}) & \frac{1}{2}(\beta_{13} + \beta_{31}) & \cdots \\ \frac{1}{2}(\beta_{12} + \beta_{21}) & \beta_{22} & \frac{1}{2}(\beta_{23} + \beta_{32}) & \\ \frac{1}{2}(\beta_{13} + \beta_{31}) & \frac{1}{2}(\beta_{23} + \beta_{32}) & \beta_{33} & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 & x_3 & \cdots \end{bmatrix} \begin{bmatrix} \beta_{11}x_1 + \frac{1}{2}(\beta_{12} + \beta_{21})x_2 + \frac{1}{2}(\beta_{13} + \beta_{31})x_3 + \cdots \\ \frac{1}{2}(\beta_{12} + \beta_{21})x_1 + \beta_{22}x_2 + \frac{1}{2}(\beta_{23} + \beta_{32})x_3 + \cdots \\ \frac{1}{2}(\beta_{13} + \beta_{31})x_1 + \frac{1}{2}(\beta_{23} + \beta_{32})x_2 + \beta_{33}x_3 + \cdots \\ \vdots \end{bmatrix} \\ &= x_1(\beta_{11}x_1 + \frac{1}{2}(\beta_{12} + \beta_{21})x_2 + \frac{1}{2}(\beta_{13} + \beta_{31})x_3 + \cdots) \\ &\quad + x_2(\frac{1}{2}(\beta_{12} + \beta_{21})x_1 + \beta_{22}x_2 + \frac{1}{2}(\beta_{23} + \beta_{32})x_3 + \cdots) \\ &\quad + x_3(\frac{1}{2}(\beta_{13} + \beta_{31})x_1 + \frac{1}{2}(\beta_{23} + \beta_{32})x_2 + \beta_{33}x_3 + \cdots) \\ &= \beta_{11}x_1^2 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \cdots \\ &\quad + \beta_{21}x_2x_1 + \beta_{22}x_2^2 + \beta_{23}x_2x_3 + \cdots \\ &\quad + \beta_{31}x_3x_1 + \beta_{32}x_3x_2 + \beta_{33}x_3^2 + \cdots \end{aligned}$$

<sup>2</sup>

$$\begin{aligned} & \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} x - y \\ -x + 3y \end{bmatrix} \\ &= x(x - y) + y(-x + 3y) = x^2 - xy - xy + 3y^2 = x^2 - 2xy + 3y^2 \end{aligned}$$

Reconstruction of the quadratic form by  $\vec{x}^T B \vec{x}$  will still be equivalent to the original  $\vec{x}^T A \vec{x}$ :

$$\begin{aligned}
 \vec{x}^T B \vec{x} &= \frac{1}{2} \vec{x}^T (A + A^T) \vec{x} \\
 &= \frac{1}{2} \vec{x}^T A \vec{x} + \frac{1}{2} \vec{x}^T A^T \vec{x} \\
 &= \frac{1}{2} \vec{x}^T A \vec{x} + \frac{1}{2} (\vec{x}^T A^T \vec{x})^T && (\vec{x}^T A^T \vec{x} = (\vec{x}^T A^T \vec{x})^T \text{ since it is just} \\
 &&& \text{a scalar in a } 1 \times 1 \text{ singleton block}) \\
 &= \frac{1}{2} \vec{x}^T A \vec{x} + \frac{1}{2} \vec{x}^T A \vec{x} && (\text{Properties 2.1.4}) \\
 &= \vec{x}^T A \vec{x}
 \end{aligned}$$

Similarly the skew-symmetric part does not contribute anything to the quadratic form<sup>3</sup>, and we will characterize all quadratic forms using symmetric matrices henceforth. This  $B$  matrix is also sometimes considered to behave as a *symmetric bilinear form*.<sup>4</sup>

### 11.1.2 (Semi)Definiteness and Congruence

An important attribute of quadratic forms is their *(semi)definiteness*. If a quadratic form  $Q(\vec{x})$  is *positive/negative-definite*, it means that it always output positive/negative numbers no matter what the input vector  $\vec{x}$  is, as long as  $\vec{x} \neq \mathbf{0}$  is a non-zero vector. Semidefiniteness loosens the restriction such that the quadratic form can also produce zero for some non-zero  $\vec{x}$ , in other words, a positive(negative)-semidefinite quadratic form always gives non-negative (non-positive) numbers. Now we will show that definiteness is related to the eigenvalues of the symmetric matrix associated to the quadratic form.

<sup>3</sup>  $\frac{1}{2} \vec{x}^T (A - A^T) \vec{x} = \frac{1}{2} \vec{x}^T A \vec{x} - \frac{1}{2} \vec{x}^T A^T \vec{x} = \frac{1}{2} \vec{x}^T A \vec{x} - \frac{1}{2} (\vec{x}^T A^T \vec{x})^T = \frac{1}{2} \vec{x}^T A \vec{x} - \frac{1}{2} \vec{x}^T A \vec{x} = 0$

<sup>4</sup> A real bilinear form  $B(\vec{x}, \vec{y}) : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$  takes two real vectors  $\vec{x}, \vec{y} \in \mathcal{V}$  and returns a real scalar.

**Definition 11.1.3.** For any quadratic form  $Q(x) = \vec{x}^T B \vec{x}$ ,  $Q$  (or  $B$ ) is called

- (a) positive-definite, if for any  $\vec{x} \neq \vec{0}$ ,  $\vec{x}^T B \vec{x} > 0$  (positive-semidefinite if  $\vec{x}^T B \vec{x} \geq 0$ ),
- (b) negative-definite, if for any  $\vec{x} \neq \vec{0}$ ,  $\vec{x}^T B \vec{x} < 0$  (negative-semidefinite if  $\vec{x}^T B \vec{x} \leq 0$ ),
- (c) indefinite if  $\vec{x}^T B \vec{x}$  can take both positive and negative values,

**Theorem 11.1.4.** The quadratic form  $Q(x) = \vec{x}^T B \vec{x}$  where  $B$  is symmetric, is

- (a) positive definite, if and only if all eigenvalues of  $B$  are positive (positive semi-definite if and only if all eigenvalues of  $B$  are non-negative),
- (b) negative definite, if and only if all eigenvalues of  $B$  are negative (negative semi-definite if and only if all eigenvalues of  $B$  are non-positive),
- (c) indefinite when there are both positive and negative eigenvalues for  $B$ .

*Proof.* We will only show the positive definite part, but the other cases essentially follow the same logic. Since we are given a symmetric matrix, the Spectral Theorem (Theorem 10.3.4) naturally comes off as handy. Assume we are working in a  $n$ -dimensional real vector space  $\mathcal{V}$ . Part (d) of the Spectral Theorem shows that any vector  $\vec{x} \in \mathcal{V}$  can be rewritten into  $\vec{x} = \vec{x}_{J_1} + \vec{x}_{J_2} + \cdots + \vec{x}_{J_k}$  where each of  $\vec{x}_{J_i} \in \mathcal{E}_{J_i}$  belongs to the respective eigenspace of  $B$ . Then as in the derivation for part (e) of the Spectral Theorem

$$B\vec{x} = \lambda_{J_1}\vec{x}_{J_1} + \lambda_{J_2}\vec{x}_{J_2} + \cdots + \lambda_{J_k}\vec{x}_{J_k}$$

Subsequently

$$\begin{aligned} Q(x) &= \vec{x}^T B \vec{x} \\ &= (\vec{x}_{J_1} + \vec{x}_{J_2} + \cdots + \vec{x}_{J_k}) \cdot (\lambda_{J_1}\vec{x}_{J_1} + \lambda_{J_2}\vec{x}_{J_2} + \cdots + \lambda_{J_k}\vec{x}_{J_k}) \\ &= \lambda_{J_1}(\vec{x}_{J_1} \cdot \vec{x}_{J_1}) + \lambda_{J_2}(\vec{x}_{J_2} \cdot \vec{x}_{J_2}) + \cdots + \lambda_{J_k}(\vec{x}_{J_k} \cdot \vec{x}_{J_k}) \end{aligned}$$

$$= \lambda_{J_1} \|\vec{x}_{J_1}\|^2 + \lambda_{J_2} \|\vec{x}_{J_2}\|^2 + \cdots + \lambda_{J_k} \|\vec{x}_{J_k}\|^2$$

where similarly  $\vec{x}_{J_i} \cdot \vec{x}_{J_{i'}} = 0$  whenever  $i \neq i'$  (Properties 10.2.3) so we have the second to third line. Now, since for any  $\vec{x}$ , all the square length quantities  $\|\vec{x}_{J_i}\|^2 \geq 0$  are greater than or equal to 0, if all  $\lambda_{J_i} > 0$  are positive and  $\vec{x}$  is not a zero vector such that at least one of the  $\|\vec{x}_{J_i}\|^2 > 0$  is positive, then the quadratic form  $Q(x) > 0$  will always take a positive value. The converse can be established by considering the contrapositive and following the same line of reasoning.  $\square$

**Example 11.1.1.** Show that the symmetric matrix

$$B = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$$

is positive-definite.

*Solution.* By Theorem 11.1.4, we simply check whether its eigenvalues are positive. Its characteristic polynomial

$$\begin{aligned} \det(B - \lambda I) &= \begin{vmatrix} 4 - \lambda & 1 \\ 1 & 4 - \lambda \end{vmatrix} = (4 - \lambda)(4 - \lambda) - (1)(1) \\ &= (16 - 8\lambda + \lambda^2) - 1 \\ &= 15 - 8\lambda + \lambda^2 = (3 - \lambda)(5 - \lambda) \end{aligned}$$

has  $\lambda = 3, 5$  as its roots which are both positive. Hence  $B$  is positive-definite. We can double-check by the explicit method of completing the square:

$$\begin{aligned} [x \ y] \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= 4x^2 - 2xy + 4y^2 \\ &= (x^2 - 2xy + y^2) + 3x^2 + 3y^2 \\ &= (x - y)^2 + 3x^2 + 3y^2 > 0 \end{aligned}$$

as long as  $x$  and  $y$  are not both zeros.  $\square$

Since it has been shown that symmetric matrices can undergo orthogonal diagonalization, which is essentially a change of coordinates to make the matrix representation of a linear operator diagonal, we may ask in general how coordinate transformation works for a quadratic/symmetric bilinear form in general. However, bear in mind that the previous way of changing coordinates (Properties 7.2.2,  $A' = P^{-1}AP$ ) is based on regarding the matrix to be a linear transformation, and it is reasonable that the rule of coordinate transformation will be somehow different when the matrix actually acts as a quadratic form. Let's take a step back and consider the change of coordinates for a vector in Theorem 7.1.12:  $[\vec{x}]_B = P_{B'}^B [\vec{x}']_{B'}$ , hence

$$\begin{aligned} Q(\vec{x}) &\equiv [\vec{x}]_B^T B [\vec{x}]_B \\ &= (P_{B'}^B [\vec{x}']_{B'})^T B (P_{B'}^B [\vec{x}']_{B'}) \\ &= [\vec{x}']_{B'}^T ((P_{B'}^B)^T B P_{B'}^B) [\vec{x}']_{B'} = [\vec{x}']_{B'}^T B' [\vec{x}']_{B'} \end{aligned}$$

so we identify the coordinate transformation of a quadratic form by  $B' = P^T B P$  where  $P$  is some invertible basis matrix. In this case,  $B'$  and  $B$  are known as *congruent*.

**Definition 11.1.5.** The coordinate transformation of a symmetric matrix  $B$  as a quadratic form follows

$$B' = P^T B P$$

where  $P$  is invertible and made up of the column vectors in the new basis expressed relative to the old basis. Any pair of  $B'$  and  $B$  related in this way are referred to as *congruent*.

Fortunately, in the previous orthogonal diagonalization process, we have  $P^{-1} = P^T$  and hence the corresponding coordinate transformation of a symmetric matrix by either treating it as a linear operator or quadratic form coincides. Hence the quadratic form  $B$  can be transformed into (and congruent to) a diagonal matrix  $D = P^T B P$  where  $P$  is consisted of all the orthonormal eigenvectors of  $B$  (Properties 10.2.5)<sup>5</sup>. Furthermore, such a  $D$  is consisted of the eigenvalues of

---

<sup>5</sup>Notice that this is not the only way to make a quadratic form diagonal (unlike a linear

$B$  and let's say  $r$  of them are positive:  $\lambda_1^+, \lambda_2^+, \dots, \lambda_r^+$ ,  $s$  of them are negative  $\lambda_1^-, \lambda_2^-, \dots, \lambda_s^-$ , and the remaining eigenvalues are zeros. Arranging them in the order of positive-negative-zero, and via an extra real diagonal factor matrix  $F$ , where

$$F_{kk} = \begin{cases} \frac{1}{\sqrt{\lambda_k^+}} & 1 \leq k \leq r \\ \frac{1}{\sqrt{-\lambda_k^-}} & r+1 \leq k \leq r+s \\ 1 & k > r+s \end{cases}$$

it is easy to see that  $F$  is invertible and we further transform the quadratic form as

$$\begin{aligned} C &= F^T D F \\ &= \left[ \begin{array}{cccccc} \frac{1}{\sqrt{\lambda_1^+}} & 0 & & & & \\ & \ddots & & & & \\ 0 & \frac{1}{\sqrt{\lambda_r^+}} & 0 & & & \\ & 0 & \frac{1}{\sqrt{-\lambda_1^-}} & 0 & & \\ & & \ddots & & & \\ 0 & & 0 & \frac{1}{\sqrt{-\lambda_s^-}} & 0 & \\ & & & 0 & 1 & 0 \\ & & & & 0 & \ddots \end{array} \right]^T \left[ \begin{array}{cccccc} \lambda_1^+ & 0 & & & & \\ & \ddots & & & & \\ 0 & \lambda_r^+ & 0 & & & \\ & 0 & \lambda_1^- & 0 & & \\ & & \ddots & & & \\ 0 & & 0 & \lambda_s^- & 0 & \\ & & & 0 & 0 & 0 \\ & & & & 0 & \ddots \end{array} \right] \left[ \begin{array}{cccccc} \frac{1}{\sqrt{\lambda_1^+}} & 0 & & & & \\ & \ddots & & & & \\ 0 & \frac{1}{\sqrt{\lambda_r^+}} & 0 & & & \\ & 0 & \frac{1}{\sqrt{-\lambda_1^-}} & 0 & & \\ & & \ddots & & & \\ 0 & & 0 & \frac{1}{\sqrt{-\lambda_s^-}} & 0 & \\ & & & 0 & 1 & 0 \\ & & & & 0 & \ddots \end{array} \right] \end{aligned}$$

transformation being dictated by its eigenvectors) and there exist many other  $P$  that can do it. (However, an important observation about them is Theorem 11.1.7, to be introduced soon.)

$$= \begin{bmatrix} 1 & & 0 & & & \\ & \ddots & & & & \\ 0 & & 1 & 0 & & \\ & & 0 & -1 & & 0 \\ & & & & \ddots & \\ 0 & & & & -1 & 0 \\ & & & & 0 & 0 & 0 \\ & & & & 0 & \ddots & \end{bmatrix} = \begin{bmatrix} I_r & & & \\ & -I_s & & \\ & & [\mathbf{0}] & \\ & & & \end{bmatrix}$$

which is known as the **canonical quadratic form** for  $B$ . To summarize, the matrix product  $PF$  converts  $B$  into such a form by  $(PF)^T B (PF) = F^T (P^T B P) F = F^T D F = C$  and therefore  $B$  is congruent to its canonical quadratic form. The following theorem shows that two different canonical quadratic form cannot be congruent and hence the canonical quadratic form of any matrix is unique.

**Theorem 11.1.6.** Two canonical quadratic form of the same extent  $n$

$$C = \begin{bmatrix} I_r & & \\ & -I_s & \\ & & [\mathbf{0}] \end{bmatrix} \quad C' = \begin{bmatrix} I_{r'} & & \\ & -I_{s'} & \\ & & [\mathbf{0}] \end{bmatrix}$$

are congruent if and only if  $r = r'$  and  $s = s'$ . As a corollary, this shows the uniqueness of canonical quadratic form.

*Proof.* The "if" part is trivial. For the "only if" part, without loss of generality, assume  $r' > r$ . If the congruence relation  $C' = P^T C P$  has to hold, then consider

$$\mathbf{e}^{(j)T} C' \mathbf{e}^{(j)} = 1 > 0 \quad \text{where } 1 \leq j \leq r'$$

which is also equal to

$$\mathbf{e}^{(j)T} P^T C P \mathbf{e}^{(j)} = \mathbf{p}^{(j)T} C \mathbf{p}^{(j)}$$

where  $P = [\mathbf{p}^{(1)} | \dots | \mathbf{p}^{(n)}]$  is consisted of  $n$  column vectors  $\mathbf{p}^{(j)}$ . We claim that there exists a non-trivial linear combination of  $\mathbf{p}^{(j)}$ ,  $1 \leq j \leq r'$ , i.e.

$\mathbf{q} = \sum_{j=1}^{r'} c_j \mathbf{p}^{(j)}$ , such that  $\mathbf{q}_i = 0$  for  $1 \leq i \leq r$ .<sup>6</sup> Subsequently, consider  $\mathbf{x} = \sum_{j=1}^{r'} c_j \mathbf{e}^{(j)}$ , and

$$\begin{aligned}\mathbf{x}^T C' \mathbf{x} &= [c_1 \ \cdots \ c'_r \ \ 0 \ \ \cdots] \begin{bmatrix} I_{r'} & & \\ & -I_{s'} & \\ & & [\mathbf{0}] \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c'_r \\ 0 \\ \vdots \end{bmatrix} \\ &= c_1^2 + \cdots + c'^2_r = \sum_{j=1}^{r'} c_j^2 > 0\end{aligned}$$

but also  $P\mathbf{x} = \sum_{j=1}^{r'} c_j P\mathbf{e}^{(j)} = \sum_{j=1}^{r'} c_j \mathbf{p}^{(j)} = \mathbf{q}$ , thus similarly

$$\begin{aligned}\mathbf{x}^T P^T C P \mathbf{x} &= (P\mathbf{x})^T C P \mathbf{x} \\ &= \mathbf{q}^T C \mathbf{q} \\ &= [0 \ \cdots \ 0 \text{ (the } r\text{-th entry)} \ *] \begin{bmatrix} I_r & & \\ & -I_s & \\ & & [\mathbf{0}] \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \text{ (the } r\text{-th entry)} \\ * \end{bmatrix} \leq 0\end{aligned}$$

Hence  $0 < \mathbf{x}^T C' \mathbf{x} = \mathbf{x}^T P^T C P \mathbf{x} \leq 0$  which is a contradiction, and it must be that  $r' = r$ .<sup>7</sup> By the same logic, we have  $s = s'$  as well.  $\square$

<sup>6</sup>The corresponding system is

$$\begin{bmatrix} \mathbf{p}_1^{(1)} & \mathbf{p}_1^{(2)} & \cdots & \mathbf{p}_1^{(r')} \\ \vdots & \vdots & & \vdots \\ \mathbf{p}_r^{(1)} & \mathbf{p}_r^{(2)} & \cdots & \mathbf{p}_r^{(r')} \\ \vdots & \vdots & & \vdots \\ \mathbf{p}_n^{(1)} & \mathbf{p}_n^{(2)} & \cdots & \mathbf{p}_n^{(r')} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ \vdots \\ c_{r'} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \text{ (the } r\text{-th entry)} \\ * \end{bmatrix}$$

The part below the  $p$ -th row does not matter since the constraints are for the first  $p$  rows and so it is effective a  $p \times p'$  underdetermined homogeneous linear system. By the discussion in Section 3.2.1, we know that there will be non-trivial solutions for the  $c_j$ ,  $1 \leq j \leq r'$ .

<sup>7</sup>The same argument in opposite direction will show that it is also not possible to have  $r' < r$ .

An immediate result from this is the **Sylvester's Law of Inertia**.

**Theorem 11.1.7** (Sylvester's Law of Inertia). All diagonalized representations of any quadratic form have the same numbers of positive, negative, and zero diagonal entries. They are collectively known as the *signature* of the quadratic form. Furthermore, if two diagonalized quadratic forms have the same signature, they are congruent.

*Proof.* If two diagonalized representations of a quadratic form have different signatures, then they can be transformed into two canonical quadratic forms with those two sets of signatures using suitable factor matrices as introduced previously. However, this violates the uniqueness of canonical quadratic form in Theorem 11.1.6, and hence the two diagonalized representations of a quadratic form must have the same signature. The last statement follows as they will have the same canonical quadratic form.  $\square$

**Example 11.1.2.** Show that

$$B = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix} \quad \text{and} \quad B' = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$$

are congruent.

*Solution.* By the Sylvester's Law of Inertia above, we simply check if the two (symmetric) quadratic forms have the same numbers of positive/negative/zero eigenvalues as they will be the diagonal entries when converted via orthogonal diagonalization. The eigenvalues of  $B$  is found by

$$\det(B - \lambda I) = \begin{vmatrix} 2 - \lambda & 3 \\ 3 & -\lambda \end{vmatrix} = 0$$

$$(2 - \lambda)(-\lambda) - (3)(3) = -9 - 2\lambda + \lambda^2 = 0$$

whose solutions are

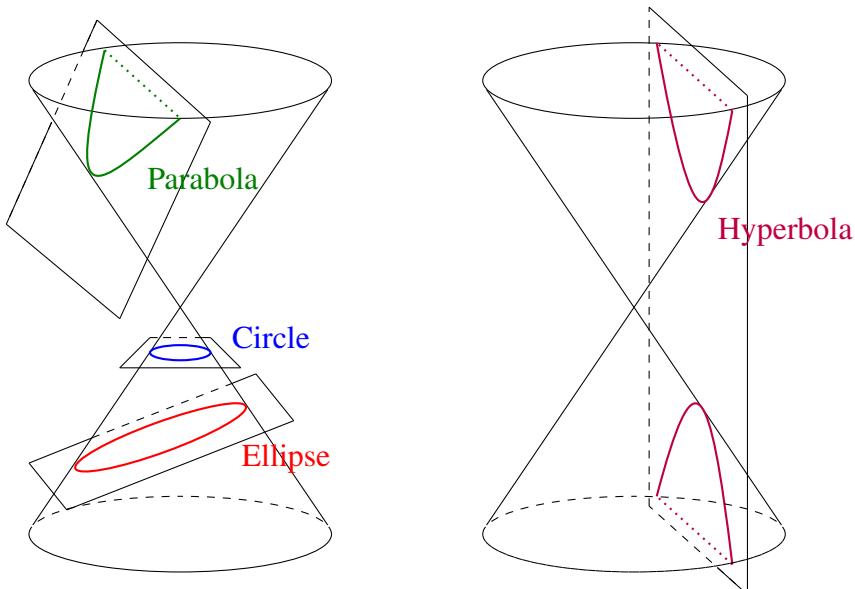
$$\lambda = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(-9)}}{2}$$

$$= 1 \pm \sqrt{10}$$

so there will be one positive and negative eigenvalue for  $B$ . It is obvious that the eigenvalues for  $B'$  are  $\lambda = 1, -2$  so one of them is positive and another is negative as well. Therefore, they are congruent.<sup>8</sup>  $\square$

### 11.1.3 Conic Sections

**Conic Sections** are the name given to three types of geometric curves in a two-dimensional space, *ellipses/circles*, *parabola* and *hyperbola*. The name originates from the fact that they can be obtained by intersecting a plane with a *double cone*, which come from the equation form below.



(Adapted from the code of Ridlo W. Wibowo)

---

<sup>8</sup>One possible choice of  $P$  as in  $B' = P^T B P$  is  $P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -1 \\ 0 & \frac{2}{3} \end{bmatrix}$ .

**Definition 11.1.8** (Conic Sections). Conic Sections (circles, ellipses, parabola, hyperbola) are the curves generated by a second-degree polynomial in two variables  $(x, y)$  that takes the general form of

$$ax^2 + bxy + cy^2 + mx + ny - h = 0$$

where  $a, b, c, m, n$  and  $h$  are all constants. It can be expressed as a quadratic form:

$$\mathbf{x}^T B \mathbf{x} = \begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} a & \frac{b}{2} & \frac{m}{2} \\ \frac{b}{2} & c & \frac{n}{2} \\ \frac{m}{2} & \frac{n}{2} & -h \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0$$

where  $\mathbf{x}^T = (x, y, 1)^T$ .

To see what types of conic sections the quadratic form represents, we can examine the determinants of  $B$  and its minor  $B_{33}$ . We simply state the results below.

**Properties 11.1.9.** The quadratic form constructed in Definition 11.1.8 represents a degenerate conic if  $\det(B) = 0$ . Otherwise, if  $\det(B) \neq 0$ , it indicates

- a hyperbola if  $\det(B_{33}) < 0$ ;
- a parabola if  $\det(B_{33}) = 0$ ;
- an ellipse if  $\det(B_{33}) > 0$ .

where

$$B_{33} = \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix}$$

In the case of an ellipse so that  $\det(B_{33}) > 0$ , if  $a = c$  and  $b = 0$ , then it is further reduced to a circle.

However, for simplicity, we will only discuss the **central conics** where the linear terms  $mx$  and  $ny$  do not appear. This excludes the case of parabola, only

keeping the ellipses and hyperbola. The quadratic form then can be simplified as follows.

**Properties 11.1.10.** Ellipses (plus circles) and hyperbola, centered at the origin, are called central conics and have the form of

$$ax^2 + bxy + cy^2 = h$$

or expressed as a quadratic form of

$$\mathbf{x}^T B_{33} \mathbf{x} = h$$

where now  $\mathbf{x} = (x, y)^T$  only and  $B_{33}$  is as defined in Properties 11.1.9.

By Properties 11.1.9, they can be classified by the discriminant  $\Delta = b^2 - 4ac$ , which is easily seen to be equal to  $-4 \det(B_{33})$ : The discriminant is positive (negative) when the graph is a hyperbola (ellipse) and  $\det(B_{33})$  is negative (positive). A zero discriminant actually represents a "parabola", but the removal of linear terms in the conic section equation reduces the parabola to degenerate straight lines.

Short Exercise: Identify the types of curve generated by  $x^2 - xy + 2y^2 = 3$  and  $x^2 + xy - y^2 = 1$ .<sup>9</sup>

Notice that sometimes an "ellipse" where  $\det(B_{33}) > 0$  may not produce real graph and is imaginary. (Take  $x^2 + 2y^2 = -3$  as an example.) To address this, we can link the definiteness property of quadratic forms to arrive at an equivalent classification:

**Theorem 11.1.11.** Given  $\mathbf{x}^T B_{33} \mathbf{x} = h$  as in Properties 11.1.10, where  $h$  is chosen to be 1 for scaling, then it represents

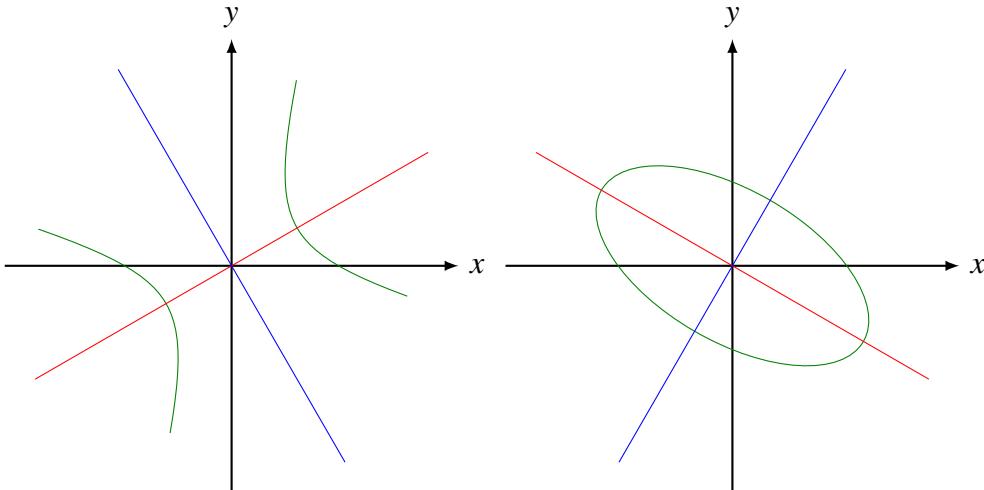
- an ellipse if  $B_{33}$  is positive definite,
- a hyperbola if  $B_{33}$  is indefinite,

---

<sup>9</sup>The first one is an ellipse ( $\Delta = (-1)^2 - 4(1)(2) = -7 < 0$ ) and the second one is a hyperbola ( $\Delta = (1)^2 - 4(1)(-1) = 5 > 0$ ).

- no real graph if  $B_{33}$  is negative definite.

The above works because if the central conic is a hyperbola and  $\det(B_{33})$  is negative, then from the view point of orthogonal diagonalization the  $2 \times 2 B_{33}$  matrix must have one positive and one negative eigenvalue, which by Theorem 11.1.4 is the same as being indefinite. It is similar for an ellipse where  $B_{33}$  being positive-definite means that its two eigenvalues are both positive and  $\det(B_{33})$  is positive as well. Meanwhile, when  $B_{33}$  is negative-definite, the two eigenvalues are both negative and  $\det(B_{33})$  is still positive. Nevertheless, as  $h$  is chosen to be 1, the negative-definiteness means that  $\mathbf{x}$  has no real solution.



Left: A hyperbola ( $x^2 + 2\sqrt{3}xy - y^2 = 1$ ), Right: An ellipse ( $\frac{7}{4}x^2 + \frac{3}{2}\sqrt{3}xy + \frac{13}{4}y^2 = 1$ ). Both of them are rotated from their standard position so that their major axis (red) and minor axis (blue) are not aligned with the  $x/y$  axes and make an angle of 30 degrees.

For example, the quadratic equation represented by the quadratic form  $\mathbf{x}^T B \mathbf{x} = 1$ , where

$$B = \begin{bmatrix} 1 & -2 \\ -2 & 3 \end{bmatrix}$$

is equivalent to  $x^2 - 4xy + 3y^2 = 1$ .  $B$  can be shown to have an eigenvalue of  $\lambda = 2 \pm \sqrt{5}$ . As  $\lambda_+ = 2 + \sqrt{5} > 0$  and  $\lambda_- = 2 - \sqrt{5} < 0$ ,  $B$  is indefinite and the curves are a pair of hyperbola by Theorem 11.1.11.

The figure in the last page shows that hyperbola and ellipses can be rotated from their standard position. The effect on the resulted quadratic equation by an orthogonal matrix is to produce cross-product terms ( $xy$  in two-dimensional cases), which can be eliminated by an inverse rotation to restore the curves so that the major and minor axes are again oriented along the  $x$  and  $y$  axes. If the graphs start with being tilted by an angle of  $\theta$ , we can make a rotation by the same angle  $\theta$  but in an opposite direction to recover the *standard position*. It is equivalent to rotate the coordinate system by an angle of  $\theta$  in the same sense of the initial tilting. The readers can be referred back to Section 10.1.2 and Definition 11.1.5 about the rotation of a coordinate system for a quadratic form.

**Example 11.1.3.** Rotate the quadratic equation  $x^2 - xy + y^2 = 1$  so that the major axis lies along the  $x$ -axis.

First, we cast the equation into the quadratic form  $\mathbf{x}^T B \mathbf{x} = 1$ , with

$$B = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

We first find the eigenvalues of  $B$ , the characteristic equation is

$$\begin{aligned} \begin{vmatrix} 1 - \lambda & -\frac{1}{2} \\ -\frac{1}{2} & 1 - \lambda \end{vmatrix} &= (1 - \lambda)^2 - (-\frac{1}{2})^2 = 0 \\ \lambda^2 - 2\lambda + \frac{3}{4} &= 0 \\ \lambda &= \frac{1}{2} \text{ or } \frac{3}{2} \end{aligned}$$

So by Theorem 11.1.11,  $A$  is positive definite and it is an ellipse. The smaller (larger) eigenvalue corresponds to the major (minor) axis. Now we consider an orthogonal matrix  $P$  to perform a rotation on the coordinate system, with the old

coordinates related to the new coordinates by  $\mathbf{x} = P\mathbf{x}'$ . So the quadratic form is transformed to

$$(P\mathbf{x}')^T B (P\mathbf{x}') = \mathbf{x}'^T (P^T B P) \mathbf{x}'$$

We immediately identify  $P^T B P$  as a rotation of the coordinate system for the matrix  $B$ , as noted by Definition 11.1.5. Section 10.2 tells us that we can deal with the cross-product terms by orthogonal diagonalization, which turns the off-diagonal entries in  $B$  into zeros. The normalized eigenvectors of  $B$  are found to be

$$\vec{v}_\lambda = \begin{cases} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} & \text{for } \lambda = \frac{1}{2} \\ \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} & \text{for } \lambda = \frac{3}{2} \end{cases}$$

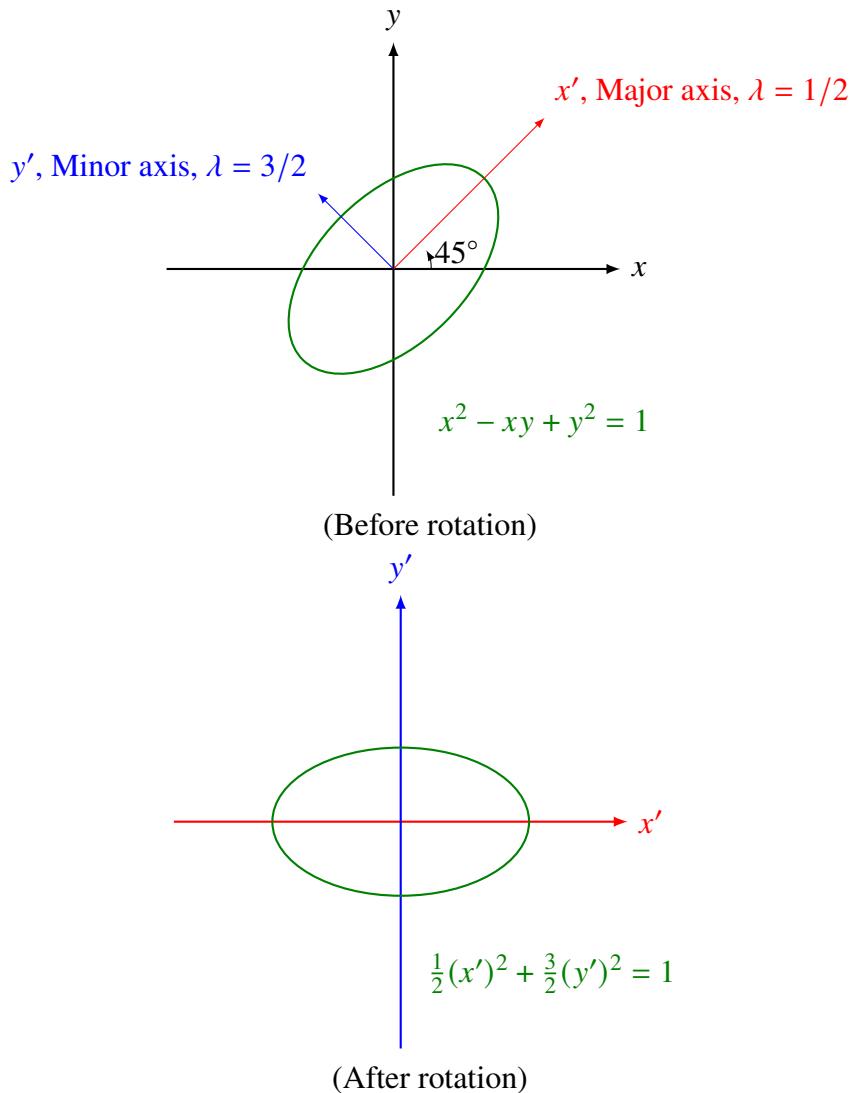
Hence we can set

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

So that

$$P^T B P = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} = D$$

The new equation is seen to be  $\mathbf{x}'^T D \mathbf{x}' = 1$ , or  $\frac{1}{2}(x')^2 + \frac{3}{2}(y')^2 = 1$ . Below are the diagrams before and after the rotation. The major (minor) axis now matches the  $x(y)$ -axis as we place the smaller (larger) eigenvalue in the first (second) diagonal entry in  $D$ .



The degree of tilting can be found to be exactly  $\pi/4 = 45^\circ$ , by comparing the general two-dimensional rotation matrix

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

against  $P$ .  $\cos \theta = \frac{1}{\sqrt{2}}$  and  $\sin \theta = \frac{1}{\sqrt{2}}$  implies that  $\tan \theta = 1$ , and  $\theta = \pi/4$ . The possibility of eliminating the cross-product terms in quadratic forms is formally known as the **Principal Axes Theorem**.

**Theorem 11.1.12** (Principal Axes Theorem). For a quadratic form  $\mathbf{x}^T B \mathbf{x}$ , where  $B$  is a symmetric matrix, we can always make an orthogonal change of variable  $\mathbf{x}' = P^T \mathbf{x}$  (or equivalently  $\mathbf{x} = P \mathbf{x}'$ ) such that it turns into  $\mathbf{x}'^T D \mathbf{x}' = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \dots$  which contains purely quadratic terms and no cross-product terms. The primed coordinates  $\mathbf{x}'$  then represent the *principal axes*.  $P$  is formed by the set of orthonormal column eigenvectors of  $B$  and  $D$  is a diagonal matrix with entries being the eigenvalues of  $B$ .

This is simply a rephrasing of Definition 10.2.1 and Properties 10.2.5. In general, for a two-dimensional quadratic form

$$\begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix}$$

It can undergo a rotation of the coordinate system by an angle  $\theta$  such that

$$\begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix}$$

the off-diagonal elements become zero. The required  $\theta$  is found by expanding the left hand side and equating both sides, which gives

$$\begin{aligned} -\sin \theta(a \cos \theta + \frac{b}{2} \sin \theta) + \cos \theta(\frac{b}{2} \cos \theta + c \sin \theta) &= 0 \\ \frac{c-a}{2} \sin(2\theta) + \frac{b}{2} \cos(2\theta) &= 0 \\ \cot(2\theta) &= \frac{a-c}{b} \end{aligned}$$

where we have applied the familiar double angle formulas from the first line to second line.

## Generalizing to the Three-dimensional Space

Since physically we are living in a three-dimensional world, it is normal to ask if we can extend the idea of geometrically quadratic shapes from two spatial axes to three. This is possible and we only need to modify the  $\mathbf{x}$  in the quadratic form to encompass the third axis so that  $\mathbf{x} = (x, y, z)^T$  and  $B$  in  $\mathbf{x}^T B \mathbf{x}$  is now a  $3 \times 3$  symmetric matrix. Now the quadratic shapes include *ellipsoids* and *hyperboloids*, and the change in coordinates to convert them into the standard position follows the exact same orthogonal diagonalization procedure. We ask the readers to try working with them in Exercise ??.

## 11.2 Statistics with Quadratic Form

### 11.2.1 Variance and Covariance

One important quantity in the world of Statistics is the **variance** of a *random variable* or *time-series*. Variance can be viewed as the spread of distribution behind the random variable. Larger the variance, the more disperse the data points are. In Earth Sciences, we often use it to quantify the variability of certain phenomena or patterns, e.g. the variance of spacetime-filtered winds can tell us how active the corresponding wave type is.

#### Single Distribution

We start with the simplest case, the definition of variance for the distribution of a single random variable first. Since in real-life we can only do a finite amount of samplings, the variance of a random variable is always approximated and inferred from the data points if we do not know the underlying statistical distribution.

**Definition 11.2.1.** For a distribution  $X$ , with  $n$  data  $x_1, x_2, x_3, \dots, x_n$ , its **population variance** is

$$\begin{aligned}\sigma^2 &= \text{Var}(X) = \frac{1}{n}((x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \dots + (x_n - \mu)^2) \\ &= \frac{1}{n} \sum_{k=1}^n (x_k - \mu)^2\end{aligned}$$

where  $\mu$  is the **mean**, or **expected value** of  $X$ , and is computed by

$$\mu = E(X) = \frac{1}{n}(x_1 + x_2 + x_3 + \dots + x_n)$$

that is, the average of all data. Hence, variance is the average of squares of difference between the data and their mean, equivalent to  $E((X - \mu)^2)$ .

A simpler formula for computing the population variance is

$$\begin{aligned}\sigma^2 &= E((X - \mu)^2) \\ &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &\quad \text{(Note that } E(X) \text{ is a constant} \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \quad \text{and } E(E(X)) = E(X), \\ &\quad E(XE(X)) = E(X)E(X).) \\ &= E(X^2) - (E(X))^2 = E(X^2) - \mu^2\end{aligned}$$

As said before we always have a finite sample, to account this, sometimes we have to use the sample variance  $s^2$ , which is the same as population variance but with the  $\frac{1}{n}$  factor replaced by  $\frac{1}{n-1}$ . (Hence  $s^2 = \frac{n}{n-1}\sigma^2$ .) As an example, given a dataset  $X$ , with 5 data  $\vec{x} = (1, 3, 6, 9, 11)^T$ , then their mean is

$$\mu = \frac{1}{5}(1 + 3 + 6 + 9 + 11) = 6$$

and the population variance is

$$\sigma^2 = \frac{1}{5}((1 - 6)^2 + (3 - 6)^2 + (6 - 6)^2 + (9 - 6)^2 + (11 - 6)^2) = 13.6$$

We can also use the short-cut formula.

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 \\ &= \frac{1}{5}(1^2 + 3^2 + 6^2 + 9^2 + 11^2) - 6^2 \\ &= 49.6 - 36 \\ &= 13.6\end{aligned}$$

Short Exercise: Find the sample variance of  $X$ .<sup>10</sup>

Note that the variance formula in Definition 11.2.1 can be written as a dot product shown below.

**Properties 11.2.2.** Given a distribution  $X$ , with  $n$  data  $\vec{x} = (x_1, x_2, x_3, \dots, x_n)^T$ , and a mean of  $\mu$ , the population variance can be written as

$$\frac{1}{n}(\vec{x}' \cdot \vec{x}') = \frac{1}{n}\mathbf{x}'^T \mathbf{x}'$$

where  $\mathbf{x}' = \vec{x}' = \vec{x} - \mu$  is the centered distribution with the mean  $\mu$  removed.

It is simply a matter of observing that  $\text{Var}(X) = \frac{1}{n}((x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2) = \frac{1}{n}(x_1 - \mu, x_2 - \mu, \dots, x_n - \mu)^T \cdot (x_1 - \mu, x_2 - \mu, \dots, x_n - \mu)^T$ . Notice that variance, as a sum of squares, can never be negative, and is a positive-semidefinite quantity.

Short Exercise: Discuss under what situation the variance will be zero.<sup>11</sup>

## Linear Combination of Multiple Distributions

Sometimes we may need to consider the "overall" distribution of the sum of multiple variables. More generally, given any linear combination of multiple

---

<sup>10</sup>It is  $s^2 = \frac{1}{5-1}((1-6)^2 + (3-6)^2 + (6-6)^2 + (9-6)^2 + (11-6)^2) = 17$ . (Or simply compute  $\frac{5}{5-1}\sigma^2$ .)

<sup>11</sup>When all data are equal (to the mean).

distributions, like  $Z = c_1X^{(1)} + c_2X^{(2)} + \dots + c_nX^{(n)}$ , we may want to know about how to compute its mean and variance. The mean will be simply  $\mu_Z = E(c_1+c_2X^{(2)}+\dots+c_nX^{(n)}) = c_1E(X^{(1)})+c_2E(X^{(2)})+\dots+c_nE(X^{(n)}) = c_1\mu_1 + c_2\mu_2 + \dots + c_n\mu_n$ , where  $E$  is linear and  $E(X^{(j)}) = \mu_j$  is the mean of  $X^{(j)}$ . The variance  $\text{Var}(Z)$  is a bit more complicated. First, we need to introduce the concept of **covariance** between any two variables, which indicates how they change together.

**Definition 11.2.3.** For two distributions  $X$  and  $Y$  consisted of  $n$  pairs of data, their population covariance is

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{1}{n}((x_1 - \mu_x)(y_1 - \mu_y) + (x_2 - \mu_x)(y_2 - \mu_y)) \\ &\quad + \dots + (x_n - \mu_x)(y_n - \mu_y)) \\ &= \frac{1}{n} \sum_{k=1}^n (x_k - \mu_x)(y_k - \mu_y)\end{aligned}$$

where  $\mu_x$  and  $\mu_y$  are the population means of  $X$  and  $Y$  respectively. It can be easily seen that  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$  so the order does not matter. If  $\vec{x}'$  and  $\vec{y}'$  are the centered data with their respective mean subtracted away, then their covariance can be denoted by a dot product as

$$\frac{1}{n} \vec{x}' \cdot \vec{y}' = \frac{1}{n} \mathbf{x}'^T \mathbf{y}'$$

For sample covariance, it is

$$q_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

where  $\bar{x}$  and  $\bar{y}$  are the *sample means* of  $X$  and  $Y$  which happen to have the same values as  $\mu_x$  and  $\mu_y$ .

There is also a short-cut formula very similar to that for variance:

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

$$\begin{aligned}
 &= E(XY - E(X)Y - XE(Y) + E(X)E(Y)) \\
 &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\
 &= E(XY) - E(X)E(Y)
 \end{aligned}$$

In general, if  $\text{Cov}(X, Y)$  is positive (negative), it means that when  $X$  increases,  $Y$  tends to increase (decrease) together. Finally, a direct comparison reveals that  $\text{Cov}(X, X) = \text{Var}(X)$  for any distribution  $X$ .

**Example 11.2.1.** Two time-series of measured zonal and meridional wind speed  $U$  and  $V$  at a weather station, are as shown in the table below.

(in $\text{m s}^{-1}$ )	$U$	$V$
1st Measurement	4.4	-3.5
2nd Measurement	3.8	-2.6
3rd Measurement	3.3	-2.7
4th Measurement	2.8	-1.4
5th Measurement	2.9	-1.2
6th Measurement	1.7	-0.8
7th Measurement	2.1	-1.1

Find the covariance of  $U$  and  $V$ .

*Solution.* It is not hard to get  $\mu_U = 3.0$  and  $\mu_V = -1.9$ . By Definition 11.2.3, we have

$$\begin{aligned}
 \text{Cov}(U, V) &= \frac{1}{7}[(4.4 - 3.0)((-3.5) - (-1.9)) + (3.8 - 3.0)((-2.6) - (-1.9)) \\
 &\quad + (3.3 - 3.0)((-2.7) - (-1.9)) + (2.8 - 3.0)((-1.4) - (-1.9)) \\
 &\quad + (2.9 - 3.0)((-1.2) - (-1.9)) + (1.7 - 3.0)((-0.8) - (-1.9)) \\
 &\quad + (2.1 - 3.0)((-1.1) - (-1.9))] \\
 &= \frac{-5.36}{7} = -0.77 \text{ m}^2 \text{ s}^{-2}
 \end{aligned}$$

Alternatively, the short-cut formula gives

$$\text{Cov}(U, V) = E(UV) - \mu_U \mu_V$$

$$\begin{aligned}
 &= \frac{1}{7}[(4.4)(-3.5) + (3.8)(-2.6) + (3.3)(-2.7) + (2.8)(-1.4) \\
 &\quad + (2.9)(-1.2) + (1.7)(-0.8) + (2.1)(-1.1)] - (3.0)(-1.9) \\
 &= (-6.466) - (-5.7) = -0.77 \text{ m}^2 \text{ s}^{-2}
 \end{aligned}$$

□

There are two take-away observations from the above example. First, if  $X$  and  $Y$  both have the same unit  $a$ , then the unit of their covariance, or the variance for each of them individually, will have a unit of  $a^2$ . If  $Y$  has a unit of  $b$  instead then their covariance will have a unit of  $ab$ . Also, covariance can take negative values, which is different from variance which is always non-negative.

Another useful measure related to variance and covariance is **correlation**. For two distribution  $X$  and  $Y$ , the correlation is defined by the following formula.

**Definition 11.2.4** (Correlation). The correlation of two distributions  $X$  and  $Y$  is

$$\begin{aligned}
 \rho_{xy} &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\
 &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Cov}(X, X)\text{Cov}(Y, Y)}}
 \end{aligned}$$

where  $\text{Var}$  and  $\text{Cov}$  are computed as given in Definitions 11.2.1 and 11.2.3.

Moreover,

**Properties 11.2.5.** The correlation between any two distributions  $X$  and  $Y$  falls in the range between  $-1$  and  $1$ , i.e.  $-1 \leq \rho_{xy} \leq 1$ .

*Proof.* We can rewrite the correlation using the vector notation for covariance in Definition 11.2.3, which gives

$$\rho_{xy} = \frac{(\vec{x}' \cdot \vec{y}')}{\sqrt{(\vec{x}' \cdot \vec{x}')( \vec{y}' \cdot \vec{y}')}}$$

$$= \frac{(\vec{x}' \cdot \vec{y}')}{\sqrt{\|\vec{x}'\| \|\vec{y}'\|}}$$

where  $\vec{x}' = \vec{x} - \mu_x$  and  $\vec{y}' = \vec{y} - \mu_y$  are centered by removing the mean from the original distributions. Observe that this quantity takes the same form as the one in the Cauchy-Schwarz Inequality (Theorem 4.2.6), and by that we promptly know that  $|\rho_{xy}| \leq 1$ .  $\square$

Correlation between two distributions  $X$  and  $Y$  indicates how their data varies together just like covariance, but normalized by their variances so that it is dimensionless and will not depend on the units used. Therefore, correlation can be considered as a standardized version of covariance that can be compared across different pairs of variables and is more interpretable. If the correlation is positive, then  $X$  and  $Y$  will generally increase or decrease together. On the other hand, if the correlation is negative, then when one of them increases, one of them will tend to decrease, and vice versa. Higher the magnitude of correlation, stronger the *linear* relationship. Notice the word "linear" here. If the correlation is close to zero, it simply means that there is no clear linear relationship between them, but this does not exclude the possibility of having other relationships, e.g. exponential or quadratic.

In the last example,  $\text{Cov}(U, V) = -0.77 \text{ m}^2 \text{ s}^{-2}$ ,  $\text{Var}(U) = 0.75 \text{ m}^2 \text{ s}^{-2}$ ,  $\text{Var}(V) = 0.90 \text{ m}^2 \text{ s}^{-2}$ , and  $\rho_{uv} = \frac{-0.77}{\sqrt{(0.75)(0.90)}} \approx -0.94$ . We have used the population variance and covariance for the computation, but they can be replaced by the sample counterparts. It may be tempting to claim that a strong negative relationship exists in this case, however, the sample size here is a bit small for this result to be meaningful.

Short Exercise: When will  $\rho_{xy}$  take the value of 1 (or  $-1$ )?<sup>12</sup>

We are now prepared to derive the variance formula for linear combinations of multiple variables.

---

<sup>12</sup>It will happen if  $X$  and  $Y$  have a perfect linear positive (negative) relationship so they appear as a straight line  $Y = aX + b$  on the  $xy$ -plane,  $a > 0$  ( $a < 0$ ).

**Properties 11.2.6.** For a distribution stemmed from a linear combination of multiple random variables, in the form of  $Z = c_1 X^{(1)} + c_2 X^{(2)} + \dots + c_n X^{(n)}$ , where the coefficients  $\vec{c} = (c_1, c_2, \dots, c_n)^T$  are all constants, the population variance  $\text{Var}(Z)$  can be expressed as a quadratic form  $\vec{c}^T Q \vec{c}$ , where

$$Q = \begin{bmatrix} \text{Cov}(X^{(1)}, X^{(1)}) & \text{Cov}(X^{(1)}, X^{(2)}) & \dots & \text{Cov}(X^{(1)}, X^{(n)}) \\ \text{Cov}(X^{(2)}, X^{(1)}) & \text{Cov}(X^{(2)}, X^{(2)}) & \dots & \text{Cov}(X^{(2)}, X^{(n)}) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X^{(n)}, X^{(1)}) & \text{Cov}(X^{(n)}, X^{(2)}) & \dots & \text{Cov}(X^{(n)}, X^{(n)}) \end{bmatrix}$$

$$= \begin{bmatrix} \text{Var}(X^{(1)}) & \text{Cov}(X^{(1)}, X^{(2)}) & \dots & \text{Cov}(X^{(1)}, X^{(n)}) \\ \text{Cov}(X^{(2)}, X^{(1)}) & \text{Var}(X^{(2)}) & \dots & \text{Cov}(X^{(2)}, X^{(n)}) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X^{(n)}, X^{(1)}) & \text{Cov}(X^{(n)}, X^{(2)}) & \dots & \text{Var}(X^{(n)}) \end{bmatrix}$$

is the so-called **covariance matrix** so that  $Q_{ij} = \text{Cov}(X^{(i)}, X^{(j)})$ . If  $[X'] = [X'^{(1)} | X'^{(2)} | \dots | X'^{(n)}]$  is the matrix consisted of the centered variables  $X'^{(j)} = X^{(j)} - E(X^{(j)})$  in columns, then we have  $Q = \frac{1}{n}[X']^T[X']$ .

*Proof.* Let's say we have  $m$  data for  $Z$ :  $z_1, z_2, \dots, z_m$ . Denote the mean of  $X^{(j)}$  by  $\mu_j$ . Starting from the expression in Definition 11.2.1, we have

$$\begin{aligned} \text{Var}(Z) &= \frac{1}{m} \sum_{k=1}^m (z_k - \mu_z)^2 \\ &= \frac{1}{m} \sum_{k=1}^m \left( \sum_{j=1}^n c_j x_k^{(j)} - \sum_{j=1}^n c_j \mu_j \right)^2 \\ &= \frac{1}{m} \sum_{k=1}^m \left( \sum_{j=1}^n (c_j x_k^{(j)} - c_j \mu_j) \right)^2 \\ &= \frac{1}{m} \sum_{k=1}^m \left[ \left( \sum_{i=1}^n c_i (x_k^{(i)} - \mu_i) \right) \left( \sum_{j=1}^n c_j (x_k^{(j)} - \mu_j) \right) \right] \end{aligned}$$

(Changing to a new dummy summation variable)

$$\begin{aligned}
 &= \frac{1}{m} \sum_{k=1}^m \left( \sum_{i=1}^n \sum_{j=1}^n c_i c_j (x_k^{(i)} - \mu_i)(x_k^{(j)} - \mu_j) \right) \\
 &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \left( \frac{1}{m} \sum_{k=1}^m (x_k^{(i)} - \mu_i)(x_k^{(j)} - \mu_j) \right) \\
 &\quad (\text{Switching the order of summation}) \\
 &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \text{Cov}(X^{(i)}, X^{(j)}) \quad (\text{Definition 11.2.3}) \\
 &= \vec{c}^T Q \vec{c}
 \end{aligned}$$

□

For two variables situation, it reduces to

$$\vec{c}^T Q \vec{c} = [c_1 \ c_2] \begin{bmatrix} \text{Cov}(X^{(1)}, X^{(1)}) & \text{Cov}(X^{(1)}, X^{(2)}) \\ \text{Cov}(X^{(2)}, X^{(1)}) & \text{Cov}(X^{(2)}, X^{(2)}) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

**Example 11.2.2.** From the previous wind speed observations example, if  $W = 0.8U - 0.6V$ , find  $\text{Var}(W)$ .

*Solution.* Earlier calculations show  $\text{Var}(U) = 0.75 \text{ m}^2 \text{ s}^{-2}$ ,  $\text{Var}(V) = 0.90 \text{ m}^2 \text{ s}^{-2}$ ,  $\text{Cov}(U, V) = \text{Cov}(V, U) = -0.77 \text{ m}^2 \text{ s}^{-2}$ . Inserting the values for the expression in Properties 11.2.6, we have

$$\begin{aligned}
 \text{Var}(W) &= [c_u \ c_v] \begin{bmatrix} \text{Var}(U) & \text{Cov}(U, V) \\ \text{Cov}(U, V) & \text{Var}(V) \end{bmatrix} \begin{bmatrix} c_u \\ c_v \end{bmatrix} \\
 &= [0.8 \ -0.6] \begin{bmatrix} 0.75 & -0.77 \\ -0.77 & 0.90 \end{bmatrix} \begin{bmatrix} 0.8 \\ -0.6 \end{bmatrix} \\
 &= 1.54 \text{ m}^2 \text{ s}^{-2}
 \end{aligned}$$

□

Finally, recall that variance is always a positive-semidefinite quantity, and since it can be calculated as a quadratic form made by the covariance matrix according to Properties 11.2.6, any covariance matrix will also be a positive-semidefinite quadratic form.

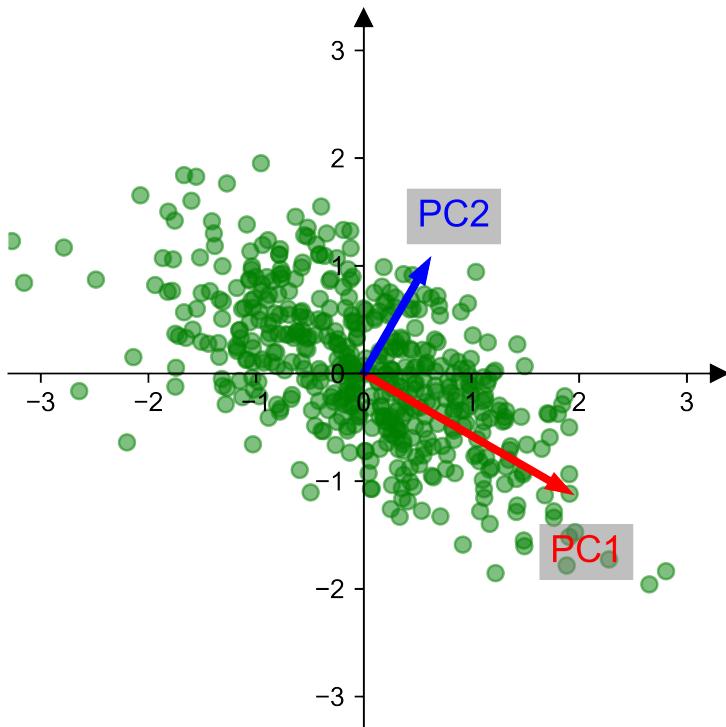
### 11.2.2 Principal Component Analysis (PCA)

A common practice in Earth Science, as well as the growing field of Machine Learning, is to compress the dimensions of a large dataset having many variables/features (*dimensionality reduction*). Given a high number of variables (for example, concentrations of various biochemical substances like blood cells or ions in the blood samples of some hospital patients) in measurements, we want to process them to extract and retain the most important patterns or signals. **Principal Component Analysis (PCA)**, which also known as **Empirical Orthogonal Functions (EOFs)** in Atmospheric Sciences, is the most common technique for this purpose, by finding the mode, or more precisely, the linear combination of features, which maximizes the variance of the data along that direction.

Consider the simplest case with two variables, or time-series  $X$  and  $Y$  first. Assume they have  $n$  pairs of data, from  $\mathbf{x}_1^T = (x_1, y_1)^T$  to  $\mathbf{x}_n^T = (x_n, y_n)^T$ . We can compute the covariance matrix

$$Q = \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{bmatrix}$$

introduced in the last section. Principal Component Analysis sets out to find a unit vector  $\mathbf{e}$ , so that the variance of data projected onto the direction indicated by  $\mathbf{e}$ , which is  $\mathbf{e}^T Q \mathbf{e}$  following from Properties 11.2.6, is maximized.



The two *principal axes/directions* (PCs) found by Principal Component Analysis for a set of data (green). The longer, red (shorter, blue) arrow represents the direction of largest (smallest) variance. The data points can be seen to spread more (less) along that direction.

Now the problem is to find, under what situation  $\mathbf{e}^T Q \mathbf{e}$  will assume its largest value. Here, we introduce a famous technique, called **Lagrange Multiplier**, coming from elementary Calculus. We will proceed with the case of two variables for brevity.

**Theorem 11.2.7** (Lagrange Multiplier). To find the extremal values attained by a function  $f(u, v, \dots)$ , under the constraint  $g(u, v, \dots) = 0$ , we consider the expression

$$h(u, v, \dots) = f(u, v, \dots) - \lambda g(u, v, \dots)$$

where  $\lambda$  is a constant so that the system below has a solution:

$$\begin{cases} \partial h / \partial u = 0 \\ \partial h / \partial v = 0 \\ \vdots = 0 \end{cases}$$

$\partial/\partial u$  ( $\partial/\partial v$ ) means differentiating with respect to  $u$  ( $v$ ) only while treating other variables as constants. The values of  $u$  and  $v$  (as well as other variables) required to attain the extrema for  $f$  are determined by solving the system of equations above.

We are now going to find the value of  $x'$  and  $y'$  so that  $\mathbf{e}^T Q \mathbf{e}$  obtains the maximum for  $\mathbf{e}^T = (x', y')$ . The constraint is that  $\mathbf{e}$  is a unit vector as a direction, and hence by the method of Lagrange Multiplier outlined above, we have

$$g(x', y') = x'^2 + y'^2 - 1 = 0$$

and with  $f(x', y') = \mathbf{e}^T Q \mathbf{e}$

$$\begin{aligned} h(x', y') &= (\mathbf{e}^T Q \mathbf{e}) - \lambda(x'^2 + y'^2 - 1) \\ &= [x' \ y'] \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} - \lambda(x'^2 + y'^2 - 1) \\ &= x'^2 \text{Cov}(X, X) + 2x'y' \text{Cov}(X, Y) + y'^2 \text{Cov}(Y, Y) - \lambda(x'^2 + y'^2 - 1) \end{aligned}$$

according to Properties 11.2.6. Carrying out the differentiation gives

$$\begin{cases} \partial h / \partial x' = 2x' \text{Cov}(X, X) + 2y' \text{Cov}(X, Y) - 2\lambda x' = 0 \\ \partial h / \partial y' = 2x' \text{Cov}(X, Y) + 2y' \text{Cov}(Y, Y) - 2\lambda y' = 0 \end{cases}$$

This system can be immediately be simplified and recognised as

$$\begin{aligned} 2 \begin{bmatrix} \text{Cov}(X, X) - \lambda & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) - \lambda \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} &= 0 \\ (Q - \lambda I) \mathbf{e} &= 0 \end{aligned}$$

which is an eigenvalue problem as introduced in Section 9.1. Hence we conclude that  $f(x', y') = \mathbf{e}^T Q \mathbf{e}$  can attain an extremal value when  $\mathbf{e}^T = (x', y')^T$  is a unit eigenvector of  $Q$ . Notice that since  $Q$  is a symmetric matrix, the eigenvectors of  $Q$  form an orthonormal basis and are orthogonal to each other by Properties 10.2.5. The corresponding magnitude of variance  $\mathbf{e}^{(j)T} Q \mathbf{e}^{(j)}$  for the  $j$ -th eigenvector is

$$\begin{aligned}\mathbf{e}^{(j)T} Q \mathbf{e}^{(j)} &= \mathbf{e}^{(j)T} (\lambda_j \mathbf{e}^{(j)}) \\ &= \lambda_j (\mathbf{e}^{(j)T} \mathbf{e}^{(j)}) \\ &= \lambda_j \|\mathbf{e}^{(j)}\| \\ &= \lambda_j\end{aligned}$$

where we have used the facts that  $Q \mathbf{e}^{(j)} = \lambda_j \mathbf{e}^{(j)}$  as per Definition 9.1.1 and the length of a unit vector is 1. This means that the variance along the direction of eigenvector is exactly equal to the corresponding eigenvalue. Note that orthogonal diagonalization (see the discussion below Definition 11.1.5) transforms the quadratic form to a diagonal matrix consisted of the eigenvalues  $\lambda_j$ , with respect to the coordinate system made up of the orthonormal eigenvectors  $\mathbf{e}^{(j)}$ . Therefore, from this perspective we can come to the same conclusion that the variance of a transformed variable along the direction indicated by each eigenvector is equal to its eigenvalue, and further, the covariance between two orthogonal directions represented by any pair of distinct eigenvectors are zero, and hence the transformed variables are made uncorrelated.

**Theorem 11.2.8** (Principal Component Analysis). For a covariance matrix  $Q$  (which happens to be symmetric), the variance  $\mathbf{e}^T Q \mathbf{e}$  achieves its maximum value  $\lambda_1$  along the direction  $\mathbf{e}^{(1)}$ , the largest eigenvalue of  $Q$  and the associated unit eigenvector. Generalizing, as  $Q$  has  $n$  orthonormal eigenvectors  $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \dots, \mathbf{e}^{(n)}$ , arranged by the magnitude of corresponding eigenvalues  $\lambda_1 > \lambda_2 > \dots > \lambda_n$  from largest to smallest, the largest variance will be  $\lambda_1$  when the direction is along  $\mathbf{e}^{(1)}$ , the second largest will be  $\lambda_2$  for  $\mathbf{e}^{(2)}$  and so on, with the smallest variance being  $\lambda_n$  for  $\mathbf{e}^{(n)}$ . This set of orthonormal eigenvectors are called the **principal axes/directions** in Principal Component Analysis.

As a side note, any quadratic form  $\mathbf{e}^T B \mathbf{e}$  will attain its maximum and minimum when  $\mathbf{e}$  is the eigenvector that represents the largest and smallest eigenvalue, which will also be the value that  $\mathbf{e}^T B \mathbf{e}$  takes when it happens. As before, this is under the constraint that  $\mathbf{e}$  is a unit vector, and this result bears the name of *Constrained Extremum Theorem*.

Going back to the problem of Principal Component Analysis, for each principal direction  $\mathbf{e}^{(j)}$  and the variance  $\lambda_j$ , we can compute the ratio of *explained variance*, which is the fraction of  $\lambda_j$  over the total variance, the sum of eigenvalues/variances from all eigenvectors for the covariance matrix. This quantity allows us to access how well the principal direction contributes to the total variance. In the coordinate system constructed by the orthonormal eigenvectors of  $Q$ ,  $[\mathbf{e}] = [\mathbf{e}^{(1)} | \mathbf{e}^{(2)} | \dots | \mathbf{e}^{(n)}]$ , the new coordinates for any data point are  $\vec{u}_i = [\mathbf{e}]^T \vec{x}_i$  (see Section 10.1.2). By the Spectral Theorem 10.3.4, this  $\vec{u}_i$  can be regarded to be the projection of  $\vec{x}_i^T = (x_i, y_i)^T$ , the  $i$ -th pair of data, onto the principal directions and are called the *principal components (PCs)*.

Usually, at the start of PCA we will detrend the data and remove the mean from each variable, such that  $x'_i = x_i - \bar{x}$  and  $y'_i = y_i - \bar{y}$  are used to replace  $x_i$  and  $y_i$ . This enables us to express the covariance matrix as  $Q = \frac{1}{n}[X'|Y']^T[X'|Y']$  following the end remark of Properties 11.2.6.

**Example 11.2.3.** The temperature data of two cities  $M$  and  $N$  are as follows.

(in °C)	$M$	$N$		$M$	$N$
1st Day	21.6	22.3	8th Day	22.1	22.4
2nd Day	21.8	21.6	9th Day	21.5	21.7
3rd Day	20.9	21.2	10th Day	22.8	22.5
4th Day	21.6	21.7	11th Day	22.2	21.6
5th Day	23.4	23.2	12th Day	23.0	23.3
6th Day	24.7	24.1	13th Day	24.2	24.7
7th Day	22.0	23.9	14th Day	23.8	23.1

Perform Principal Component Analysis over them and find the most important principal direction, and extract the time-series of the corresponding PC.

*Solution.* The means of  $M$  and  $N$  are  $22.5^{\circ}\text{C}$  and  $22.6^{\circ}\text{C}$  respectively. After detrending by subtracting the respective means, the new data are

(in $^{\circ}\text{C}$ )	$M'$	$N'$		$M'$	$N'$
1st Day	-1.5	-0.3	8th Day	-0.4	-0.2
2nd Day	-0.7	-1.0	9th Day	-1.0	-0.9
3rd Day	-1.6	-1.4	10th Day	0.3	-0.1
4th Day	-0.9	-0.9	11th Day	-0.3	-1.0
5th Day	0.9	0.6	12th Day	0.5	0.7
6th Day	2.2	1.5	13th Day	1.7	2.1
7th Day	-0.5	0.4	14th Day	1.3	0.5

From Properties 11.2.6, the sample covariance matrix is

$$\begin{aligned} Q &= \frac{1}{14-1} \begin{bmatrix} M' \cdot M' & M' \cdot N' \\ N' \cdot M' & N' \cdot N' \end{bmatrix} \\ &= \frac{1}{13} \begin{bmatrix} -1.5 & -0.7 & -1.6 & \dots \\ -0.3 & -1.0 & -1.4 & \end{bmatrix} \begin{bmatrix} -1.5 & -0.3 \\ -0.7 & -1.0 \\ -1.6 & -1.4 \\ \vdots \end{bmatrix} \\ &= \frac{1}{13} \begin{bmatrix} 18.18 & 13.66 \\ 13.66 & 13.64 \end{bmatrix} = \begin{bmatrix} 1.398 & 1.051 \\ 1.051 & 1.049 \end{bmatrix} \end{aligned}$$

The unit eigenvectors for  $Q$  and thus the principal directions can be found to be  $\mathbf{e}^{(1)} = (0.763, 0.647)^T$  with a larger variance  $\lambda_1 = 2.289 (^{\circ}\text{C})^2$ , and  $\mathbf{e}^{(2)} = (-0.647, 0.763)^T$  of a smaller variance  $\lambda_2 = 0.159 (^{\circ}\text{C})^2$ . The first principal component accounts for  $\frac{2.289}{2.289+0.159} \approx 93.5\%$  of the total variance.

We can project every pair of data  $\vec{x}'_i^T = (m'_i, n'_i)^T$  onto the principal directions by computing  $\vec{u}_i = [\mathbf{e}]^T \vec{x}'_i$ , where  $[\mathbf{e}] = [\mathbf{e}^{(1)} | \mathbf{e}^{(2)}]$ . The resulting principal components time-series are

	D-1	D-2	D-3	D-4	D-5	D-6	D-7
$u^{(1)}$	-1.338	-1.181	-2.126	-1.268	1.075	2.648	-0.123
$u^{(2)}$	0.741	-0.310	-0.034	-0.105	-0.124	-0.278	0.628
	D-8	D-9	D-10	D-11	D-12	D-13	D-14
$u^{(1)}$	-0.434	-1.345	0.164	-0.875	0.834	2.655	1.315
$u^{(2)}$	0.106	-0.040	-0.270	-0.569	0.211	0.503	-0.459

In details, the two principal components for the first day is computed by

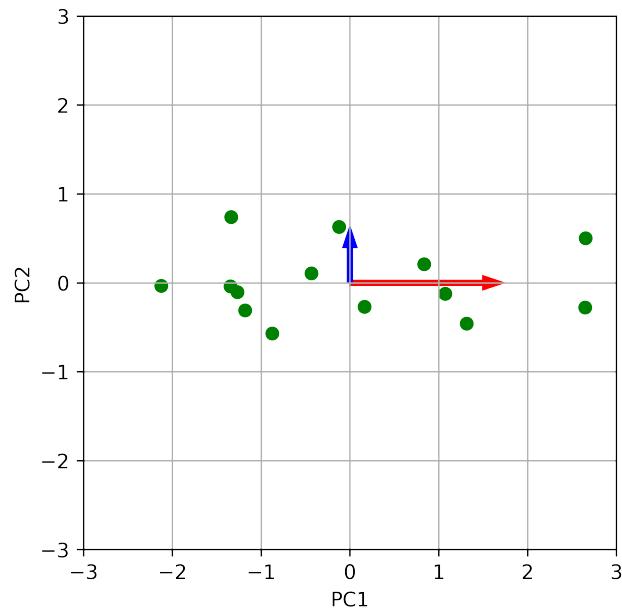
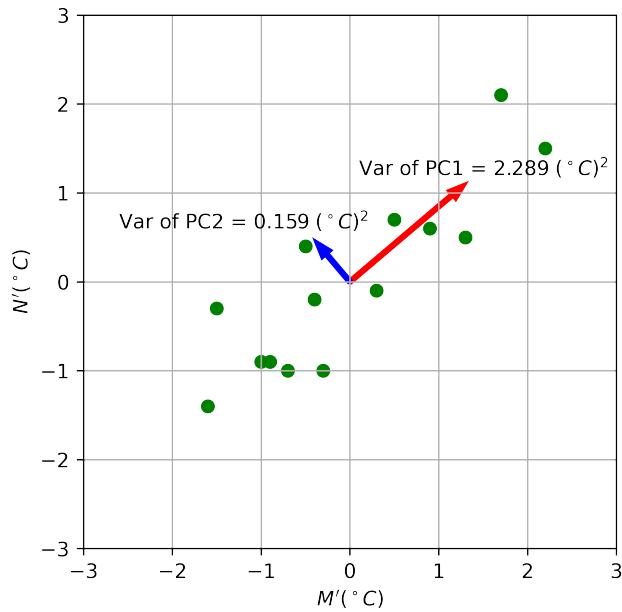
$$\begin{bmatrix} u_1^{(1)} \\ u_1^{(2)} \end{bmatrix} = \begin{bmatrix} 0.763 & 0.647 \\ -0.647 & 0.763 \end{bmatrix} \begin{bmatrix} -1.5 \\ -0.3 \end{bmatrix} = \begin{bmatrix} -1.338 \\ 0.741 \end{bmatrix}$$

The original (detrended) data can be recovered by  $\vec{x}'_i = [\mathbf{e}] \vec{u}'_i$ . If we want to extract the signal originated from the first principal direction only, we can simply remove other column eigenvector(s) in  $[\mathbf{e}]$  as well as discard the other principal value(s) in  $\vec{u}'_i$ . The time-series reconstructed by the first PC mode is hence computed by  $\vec{x}'_i = \mathbf{e}^{(1)} u_i^{(1)}$ :

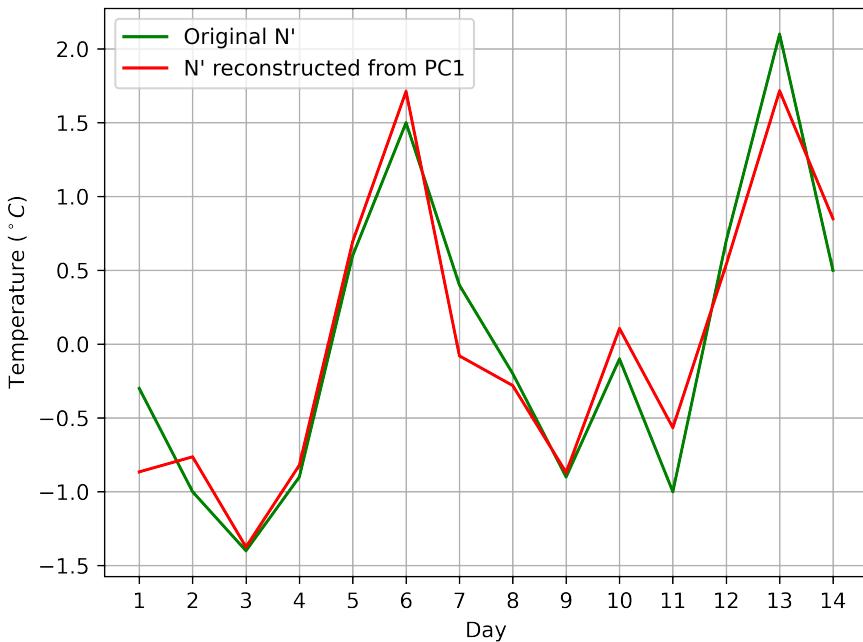
(in °C)	D-1	D-2	D-3	D-4	D-5	D-6	D-7
M'	-1.021	-0.901	-1.622	-0.968	0.820	2.020	-0.094
N'	-0.865	-0.763	-1.374	-0.820	0.695	1.712	-0.079
	D-8	D-9	D-10	D-11	D-12	D-13	D-14
M'	-0.331	-1.026	0.125	-0.668	0.636	2.025	1.003
N'	-0.281	-0.869	0.106	-0.566	0.539	1.716	0.850

For example, for the third day PC1 contributes

$$\vec{x}'_3 = \begin{bmatrix} m'_3 \\ n'_3 \end{bmatrix} = \mathbf{e}^{(1)} u_3^{(1)} = \begin{bmatrix} 0.763 \\ 0.647 \end{bmatrix} \begin{bmatrix} -2.126 \\ -1.374 \end{bmatrix} = \begin{bmatrix} -1.622 \\ -1.374 \end{bmatrix}$$



The data (green) before and after rotation to the principal axes, with the one with a larger/smaller variance shown as a red/blue arrow.



Comparison between the original time-series and the one reconstructed using only first PC for  $N'$ .

□

## 11.3 Python Programming

Since doing EOFs as an Earth Science application is essentially a PCA which has to rely on a computer when the dataset is huge, we will get into the Python programming part first in this chapter. We will need the scikit-learn package (`sklearn`) for this, and let's use Example 11.2.3 for demonstration.

```
import numpy as np
from sklearn.decomposition import PCA

X = np.array([[21.0, 22.3],
              [21.8, 21.6],
```

```
[20.9, 21.2],  
[21.6, 21.7],  
[23.4, 23.2],  
[24.7, 24.1],  
[22.0, 23.0],  
[22.1, 22.4],  
[21.5, 21.7],  
[22.8, 22.5],  
[22.2, 21.6],  
[23.0, 23.3],  
[24.2, 24.7],  
[23.8, 23.1]])
```

We have to prepare the data where each column represents the time-series of one variable so the shape of  $X$  is (Number of samples, Number of features). Now define a PCA object and fit it with the data. We can choose how many PCs to be used, and for now we will keep all of them so that `n_components = 2`.

```
pca = PCA(n_components=2)  
pca.fit(X)
```

We can retrieve the principal directions and variances by

```
print(pca.components_ )  
print(pca.explained_variance_ )
```

which gives

```
[[ 0.76286648  0.64655605]  
 [-0.64655605  0.76286648]]  
 [2.28902525  0.15866706]]
```

Notice that the principal directions are arranged in rows so that to get the first one we write `pca.components_[0, : ]` that returns `[0.763 0.647]`. The percentage of explained variances are simply computed by

```
print(pca.explained_variance_ /np.sum(pca.explained_variance_ ))
```

that returns `[0.9352 0.0648]`. To obtain the time-series of transformed PCs, we simply use the `transform` method:

```
Z = pca.transform(X)  
print(Z)
```

yielding the expected output of

```
[[ -1.33826654  0.74097414]
 [ -1.18056259 -0.31027725]
 [ -2.12576485 -0.03352339]
 ...
 [ 1.31500445 -0.45908963]]
```

To reconstruct the time-series using only some (the first) PC, we can do the followings.

```
Z_trimmed = np.copy(Z)
Z_trimmed[:,1:] = 0
X_inv = pca.inverse_transform(Z_trimmed)
print(X_inv)
```

These generate

```
[[21.47908131 21.73473567]
 [21.59938837 21.83670011]
 [20.87832525 21.22557387]
 ...
 [23.50317282 23.45022409]]
```

equivalent to the last table in Example 11.2.3 but with the original means included. The following code paragraphs produce the main part of the three plots shown in the example.

```
Y = X - np.mean(X, axis=0)
lambda_1 = pca.explained_variance_[0]
lambda_2 = pca.explained_variance_[1]

plt.scatter(Y[:,0], Y[:,1], color="g")
plt.xlim([-3,3])
plt.ylim([-3,3])
plt.arrow(0,0,lambda_1**0.5*pca.components_[0,0],lambda_1
          **0.5*pca.components_[0,1], color="r", width=0.05)
plt.arrow(0,0,lambda_2**0.5*pca.components_[1,0],lambda_2
          **0.5*pca.components_[1,1], color="b", width=0.05)
plt.grid()
plt.gca().set_aspect("equal")
plt.show()

plt.scatter(Z[:,0], Z[:,1], color="g")
```

```
plt.xlim([-3,3])
plt.ylim([-3,3])
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.grid()
plt.gca().set_aspect("equal")
plt.show()

plt.plot(np.arange(1,14+1), Y[:,1], color="g", label="Original
N'")
plt.plot(np.arange(1,14+1), X_inv[:,1] - np.mean(X, axis=0)
[1], color="r", label="N' reconstructed from PC1")
plt.legend()
plt.xticks(np.arange(1,14+1))
plt.show()
```

## 11.4 Earth Science Applications: Empirical Orthogonal Functions (EOFs)

Here we will read the ERA5 dataset for sea surface temperature (SST) and find the dominant patterns of global SST variability. The data can be retrieved from <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download> and selecting the options "Type: Reanalysis, Variable: Sea surface temperature, Time: 00:00, Data format: NetCDF4". We will choose the time period from 1991 to 2020, every month/day, and a spatial domain from 45 °N to 45 °S and 110 °E to 80 °W. We will also need the land-sea mask that can be downloaded via [https://confluence.ecmwf.int/download/attachments/140385202/lsm\\_1279l4\\_0.1x0.1.grb\\_v4\\_unpack.nc?version=1&modificationDate=1591983422208&api=v2](https://confluence.ecmwf.int/download/attachments/140385202/lsm_1279l4_0.1x0.1.grb_v4_unpack.nc?version=1&modificationDate=1591983422208&api=v2). Now, import the required packages.

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import xarray as xr
from sklearn.decomposition import PCA
import cartopy.crs as ccrs
```

## 11.4 Earth Science Applications: Empirical Orthogonal Functions (EOFs)

Load the SST data, work on coarse grids with a  $2.5^\circ \times 2.5^\circ$  spatial resolution (optional to the reduce computational cost), and exclude leap days.

```
SST_nc = xr.open_dataset("ERA5_SST.nc")
coarse_lat = np.array(SST_nc["latitude"][:,::10]) # 0.25 deg *
    10 = 2.5 deg
coarse_lon = np.array(SST_nc["longitude"][:,::10])
time = pd.Series(SST_nc["valid_time"])
no_leapdays = ~((time.dt.month == 2) & (time.dt.day == 29))
```

Also, prepare the land-sea mask to exclude land grid points and flatten it for indexing later.

```
land_sea_mask_nc = xr.open_dataset("lsm_127914_0.1x0.1.
    grb_v4_unpack.nc")
land_sea_mask = np.array(land_sea_mask_nc["lsm"].sel({
    "latitude": coarse_lat, "longitude": coarse_lon}, method="nearest")[:, ...])
land_sea_mask = np.where(land_sea_mask > 0, 1, 0).astype(bool)
land_sea_mask_flatten = land_sea_mask.flatten()
```

Preprocessing the SST data by subtracting the yearly climatology to acquire the anomaly fields, scale them by the square roots of cosines of the latitudes to account for area weighting and keep only the valid overwater grid points using the land-sea mask:

```
SST_data_arr = np.array(SST_nc["sst"].sel({"latitude":
    coarse_lat, "longitude": coarse_lon}))
SST_data_arr_no_leap = SST_data_arr[no_leapdays, ...].reshape
    (30, 365, len(coarse_lat), len(coarse_lon))
# The SST array now has the shape of 30 years * 365 days *
    nlat * nlon
SST_clim = np.mean(SST_data_arr_no_leap, axis=0)
SST_anomaly = SST_data_arr_no_leap - SST_clim

cos_factor_root = np.cos(np.deg2rad(coarse_lat))**0.5
SST_anomaly_weighted = SST_anomaly * cos_factor_root[None, None
    , :, None]

SST_flatten = SST_anomaly_weighted.reshape(30*365, len(
    coarse_lat)*len(coarse_lon))
SST_valid = SST_flatten[:, ~land_sea_mask_flatten]
```

Call the PCA and fit it with the prepared, flattened SST data:

```
SST_PCA = PCA(n_components=3) # any number will be fine
SST_PCA.fit(SST_valid)
```

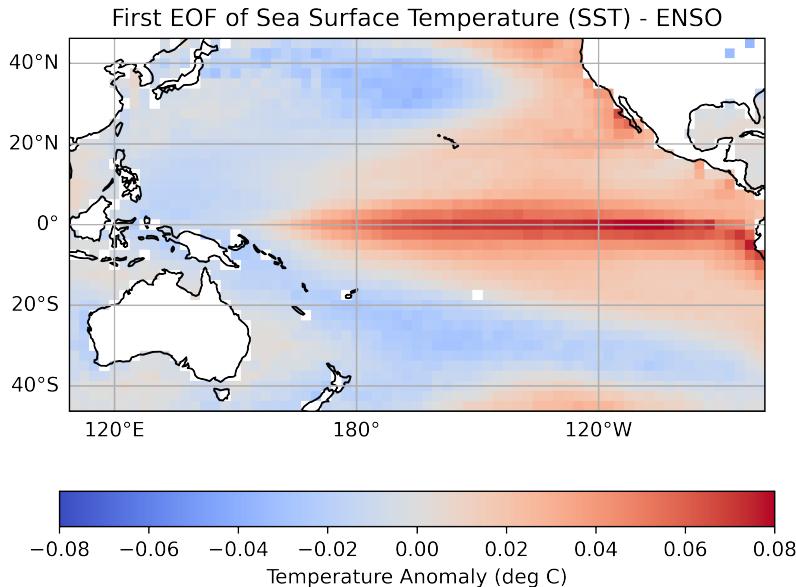
Then, recover the latitude-longitude structure of the first PC:

```
PC1 = np.full(len(coarse_lat)*len(coarse_lon), np.nan)
PC1[~land_sea_mask_flatten] = SST_PCA.components_[0,:]
    # Fill PC1 at appropriate overwater entries
PC1_2D = PC1.reshape(len(coarse_lat), len(coarse_lon)) /
    cos_factor_root[:,None]
    # Invert the cosine factor
```

and plot it on the map.

```
plt.figure()
plt.subplot(111, projection=ccrs.PlateCarree(central_longitude
    =180)) # the central_longitude option is needed to plot
    across the International Date Line
plt.pcolormesh(coarse_lon, coarse_lat, PC1_2D, cmap="coolwarm"
    , vmin=-0.08, vmax=0.08, transform=ccrs.PlateCarree())
plt.gca().coastlines()
gl = plt.gca().gridlines(draw_labels=True)
gl.top_labels = False
plt.title("First EOF of Sea Surface Temperature (SST) - ENSO")
plt.colorbar(orientation="horizontal", label="Temperature
    Anomaly (deg C)")
plt.savefig("SST_EOF")
```

You should be able to get the following figure.



This EOF pattern represents the ***El Niño–Southern Oscillation (ENSO)*** phenomenon where the western Pacific gets cooler (warmer) and eastern Pacific becomes warmer (cooler) during the El Niño (La Niña) phase.

## 11.5 Exercises

**Exercise 11.1** Identify the following conic sections and eliminate the cross-product terms by an appropriate rotation.

- (a)  $x^2 + 5xy + 3y^2 = 1$ ,
- (b)  $x^2 - xy + 2y^2 = 4$ ,
- (c)  $x^2 + 2xy + y^2 = 1$ , what is the graph generated?

**Exercise 11.2** Find a new expression for the standard hyperbola  $y^2 - x^2 = 1$  if a anti-clockwise rotation of 60 degrees is done. What about a reflection along the  $x$ -axis?

**Exercise 11.3** Three-dimensional quadrics can also be treated in a similar fashion to two-dimensional conic sections. Find the length of three axes for an ellipsoid  $x^2 + y^2 + z^2 + 0.5xy - yz + 0.5xz = 1$  by doing an orthogonal coordinate transformation. What is the requirement for a three-dimensional quadratic form to represent an ellipsoid?

**Exercise 11.4** Use the Sylvester's Law of Inertia to argue that a symmetric matrix  $B$  is positive-definite if and only if  $B = P^T P$  for some invertible matrix  $P$ . Hint: Consider  $P^T I P$ .

**Exercise 11.5** Find the covariance matrix for three pressure time-series over 10 days at cities  $X, Y, Z$  (in hPa, relative to 1000 hPa), which take the values

$X$	$Y$	$Z$
17.1	19.2	22.0
18.5	16.9	25.4
14.8	15.3	17.3
19.7	21.6	23.5
24.1	22.3	26.8
21.6	20.9	23.2
28.0	26.7	29.5
24.3	25.0	22.5
20.3	21.5	27.2
23.4	22.4	24.6

Find the variance of  $W = X - 0.5Y - 0.5Z$ .

**Exercise 11.6** Carry out Principal Component Analysis on the data set above. Find the Principal Directions and the ratio of explained variance for each of

them. Reconstruct the data using the first principal component with largest variance only.

**Exercise 11.7** Download the ERA5 datasets for 200 hPa and 850 hPa zonal winds, as well as total column water over some 10 to 20 years from 15 °N to 15 °S and 60 °E to 160 °E. Standardize the three variables via dividing them by their standard deviations respectively, concatenate them and follow the procedure outlined in Section 11.4 to do the so-called *combined EOFs*. Recover the physical patterns through multiplying the entries of EOF modes back by the standard deviations of corresponding variables. You should be able to observe the appearance of *Madden-Julian Oscillation (MJO)* from the first two EOFs.



## Chapter 12

# Inner Product Spaces

---

In Chapter 6, we have discussed about what it means to be a vector space. Previously, it is accompanied by the dot product operation as defined in Section 4.2.1 that gives rise to the notion of *Euclidean* distance in  $\mathbb{R}^n$  (complex dot product for  $\mathbb{C}^n$ ). In this chapter, we will show that the usual dot product is not the only way to measure distance between vectors: we can equip any vector space with a so-called *inner product* that fulfills certain criteria and leads to an alternative expression for the length of vectors, in place of the dot product. This generalizes a vector space to an *inner product space*, and many concepts for a vector space, like orthogonality, can be adapted to be applied in inner product spaces. Particularly, we will have the *adjoint* as an inner product space equivalent to the usual transpose. Finally, we will talk about *special polynomials* which are generated by considering suitable inner products and are often used in Earth Science applications, such as solving *partial differential equations (PDEs)*.

## 12.1 Definition and Properties of Inner Product Spaces

### 12.1.1 Requirements of Inner Products

As introduced in the beginning, an *inner product space* is a vector space (either real or complex) that comes with an *inner product* operation that is akin to the usual dot product. To qualify as a valid inner product, it has to fulfill four requirements as suggested below.

**Definition 12.1.1** (Inner Product (Space)). An inner product on a vector space  $\mathcal{V}$  over  $\mathbb{R}$  ( $\mathbb{C}$ ) is a function that takes a pair of vectors  $\vec{u}, \vec{v} \in \mathcal{V}$  as the input and returns an  $\mathbb{R}$  ( $\mathbb{C}$ ) number, denoted by  $\langle \vec{u}, \vec{v} \rangle$ . Then, for any  $\vec{u}, \vec{v}, \vec{w} \in \mathcal{V}$  and scalar  $a \in \mathbb{R}$  ( $\mathbb{C}$ ), the following four axioms have to hold.

1.  $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$  ( $\langle \vec{u}, \vec{v} \rangle = \overline{\langle \vec{v}, \vec{u} \rangle}$ ) ((Conjugate) Symmetry);
2.  $\langle \vec{u} + \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle + \langle \vec{v}, \vec{w} \rangle$  (Additivity);
3.  $\langle a\vec{u}, \vec{v} \rangle = a\langle \vec{u}, \vec{v} \rangle$  (Homogeneity); and
4.  $\langle \vec{v}, \vec{v} \rangle \geq 0$ , and  $\langle \vec{v}, \vec{v} \rangle = 0$  if and only if  $\vec{v} = \mathbf{0}$  is the zero vector. (Positivity)

The second and third condition can be combined into *linearity*: given another scalar  $b \in \mathbb{R}$  ( $\mathbb{C}$ ), we have  $\langle a\vec{u} + b\vec{v}, \vec{w} \rangle = a\langle \vec{u}, \vec{w} \rangle + b\langle \vec{v}, \vec{w} \rangle$ . A real (complex) vector space with an inner product is then known as a real (complex) inner product space.

Short Exercise: Show that  $\langle a\vec{u}, b\vec{v} \rangle = ab\langle \vec{u}, \vec{v} \rangle$  for any complex inner product space.<sup>1</sup>

For now, we will limit ourselves to finite-dimensional vector/inner product spaces until Section 12.1.3. It is not hard to verify that the above axioms hold

---

<sup>1</sup>  $\langle a\vec{u}, b\vec{v} \rangle = a\langle \vec{u}, b\vec{v} \rangle = a\overline{\langle b\vec{v}, \vec{u} \rangle} = ab\overline{\langle \vec{v}, \vec{u} \rangle} = ab\overline{\langle \vec{v}, \vec{u} \rangle} = ab\langle \vec{u}, \vec{v} \rangle$  by the first and third axiom.

for the usual dot product, so  $\mathbb{R}^n$  (as well as  $\mathbb{C}^n$ ) is automatically an inner product space when the (complex) dot product is equipped. In this case they are known as the *standard inner product* and *Euclidean n-space*. However,  $\mathbb{R}^n$  will be a different inner product space when another inner product is used. It can be shown that all alternative inner products that can be applied to  $\mathbb{R}^n$  are precisely positive-definite symmetric bilinear forms as the extension of positive-definite quadratic forms introduced in the last chapter. Given such a positive-definite quadratic form  $B$ , then it can be verified that  $\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$  will satisfy the real inner product axioms.<sup>2</sup> Below are some other properties of inner product extended from the axioms that can be compared to Properties 4.2.3 and 8.2.3.

**Properties 12.1.2.** For vectors  $\vec{u}, \vec{v}, \vec{w} \in \mathcal{V}$  and scalar  $b \in \mathbb{R} (\mathbb{C})$  in a real (complex) inner product space, we have

1.  $\langle \vec{u} \pm \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle \pm \langle \vec{v}, \vec{w} \rangle;$
2.  $\langle \vec{u}, \vec{v} \pm \vec{w} \rangle = \langle \vec{u}, \vec{v} \rangle \pm \langle \vec{u}, \vec{w} \rangle;$
3.  $\langle \vec{v}, \mathbf{0} \rangle = \langle \mathbf{0}, \vec{v} \rangle = 0;$
4.  $\langle \vec{u}, b\vec{v} \rangle = b\langle \vec{u}, \vec{v} \rangle$  ( $\langle \vec{u}, b\vec{v} \rangle = \bar{b}\langle \vec{u}, \vec{v} \rangle$ );
5. if  $\langle \vec{u}, \vec{v} \rangle = \langle \vec{u}, \vec{w} \rangle$  for all  $\vec{u}$ , then  $\vec{v} = \vec{w}$ .

*Proof.* We will skip (1). (2):  $\langle \vec{u}, \vec{v} \pm \vec{w} \rangle = \overline{\langle \vec{v} \pm \vec{w}, \vec{u} \rangle} = \overline{\langle \vec{v}, \vec{u} \rangle} \pm \overline{\langle \vec{w}, \vec{u} \rangle} = \langle \vec{u}, \vec{v} \rangle \pm \langle \vec{u}, \vec{w} \rangle$  by the first and second axiom. (3):  $\langle \mathbf{0}, \vec{v} \rangle = \langle 0\vec{u}, \vec{v} \rangle = 0\langle \vec{u}, \vec{v} \rangle = 0$  using arbitrary  $\vec{u}$  and the third axiom. (4) simply follows from the last short exercise with  $a = 1$ . For (5):

$$\begin{aligned} \langle \vec{u}, \vec{v} \rangle &= \langle \vec{u}, \vec{w} \rangle \\ \langle \vec{u}, \vec{v} - \vec{w} \rangle &= 0 \end{aligned} \quad (\text{By (1)})$$

---

<sup>2</sup>For (1):  $\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v} = (\vec{u}^T B \vec{v})^T$  since it is only a scalar, and then  $(\vec{u}^T B \vec{v})^T = \vec{v}^T B^T \vec{u} = \vec{v}^T B \vec{u} = \langle \vec{v}, \vec{u} \rangle$  as  $B$  is symmetric. (2) and (3) are obvious. For (4), as  $B$  is required to be positive-definite,  $\langle \vec{v}, \vec{v} \rangle = \vec{v}^T B \vec{v} > 0$  by Definition 11.1.3 as long as  $\vec{v} \neq \mathbf{0}$  and it is apparent that  $\langle \vec{v}, \vec{v} \rangle = \mathbf{0} B \mathbf{0} = 0$  when  $\vec{v} = \mathbf{0}$ .

Then by letting  $\vec{u} = \vec{v} - \vec{w}$  and using the last axiom we get  $\vec{v} - \vec{w} = \mathbf{0}$  and thus  $\vec{v} = \vec{w}$ .  $\square$

### 12.1.2 Generalization of Length and Orthogonality via Inner Products

As noted earlier, the idea of inner products extends the usual dot product and we may ask how the notion of vector length, which can be expressed via the dot product of the vector with itself (Properties 4.2.2), is carried over to an inner product space. The most natural generalization is to simply replace the dot product by an inner product in Properties 4.2.2 when computing such a "length", which is now more properly known as a *norm*. This makes physical sense as the last axiom in Definition 12.1.1 forces the norm to always be positive (0 if it is the zero vector) just like the usual length.

**Properties 12.1.3.** The norm of a vector in an inner product space is induced by

$$\|\vec{v}\| = \sqrt{\langle \vec{v}, \vec{v} \rangle} \quad \text{or equivalently} \quad \|\vec{v}\|^2 = \langle \vec{v}, \vec{v} \rangle$$

Unit vectors are then conveniently created using this definition of norm.

**Definition 12.1.4.** The unit vector of a non-zero vector  $\vec{v}$  in an inner product space is denoted as  $\hat{v}$  and is given by

$$\hat{v} = \frac{1}{\|\vec{v}\|} \vec{v}$$

where the norm  $\|\vec{v}\|$  is now defined as in Properties 12.1.3.

The notion of orthogonality in Properties 4.2.5 is also transferred to an inner product space by the same essence of replacing the dot product by an inner product.

**Properties 12.1.5.** Two vectors  $\vec{u}, \vec{v} \in \mathcal{V}$  in an inner product space are said to be orthogonal with respect to the inner product when  $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle = 0$ .

Some other related results derived using dot product are also valid when inner products are used instead and we note them below.

**Theorem 12.1.6** (Cauchy–Schwarz Inequality). Given two vectors  $\vec{u}, \vec{v} \in \mathcal{V}$  in an inner product space, we have

$$|\langle \vec{u}, \vec{v} \rangle| \leq \|\vec{u}\| \|\vec{v}\| = \sqrt{\langle \vec{u}, \vec{u} \rangle} \sqrt{\langle \vec{v}, \vec{v} \rangle}$$

The proof is essentially the same as the one in Theorem 4.2.6 but with all the dot product expressions replaced by the inner product. Also

**Properties 12.1.7.** Non-zero orthogonal vectors with respect to any inner product are linearly independent.

Again, the proof follows the same line of arguments as in Properties 6.3.12 with the dot product changed to an inner product.

**Example 12.1.1.** Show that  $\vec{u} = (1, 2)^T$  and  $\vec{v} = (-3, 4)^T$  in  $\mathbb{R}^2$  are orthogonal to each other if the inner product used is given by

$$\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$$

where

$$B = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

*Solution.* First we have to make sure the inner product defined as a symmetric bilinear form above is indeed valid, particularly it has to be positive-definite. By Theorem 11.1.4, it simply amounts to check if the eigenvalues are all positive.

A simple calculation reveals that  $\lambda = \frac{3+\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2} > 0$  so we can proceed to calculate

$$\begin{aligned}\langle \vec{u}, \vec{v} \rangle &= \vec{u}^T B \vec{v} \\ &= [1 \ 2] \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 4 \end{bmatrix} \\ &= [1 \ 2] \begin{bmatrix} -2 \\ 1 \end{bmatrix} = 0\end{aligned}$$

Hence by Properties 12.1.5 the two vectors are orthogonal with respect to the said inner product. Obviously they will not be orthogonal if the usual dot product is used instead.  $\square$

### 12.1.3 Infinite-dimensional Inner Product Spaces

We have been staying in the realm of finite-dimensional vector/inner product spaces until now but the utility of inner product spaces only become the most significant when they are infinite-dimensional. As suggested by Properties 6.2.4, if there is a basis for an infinite-dimensional vector space<sup>3</sup>, it must be infinite. An example of infinite-dimensional inner product spaces will be  $C^0[a, b]$ , the vector space of all continuous functions in one variable over the interval  $a \leq x \leq b$ , equipped with the frequently used inner product of

$$\langle f, g \rangle = \int_a^b f(x) \overline{g(x)} dx \quad (12.1)$$

where  $a < b$  are some real constants. Checking the validity of this inner product formulation is not hard.<sup>4</sup> Unfortunately, the above example lacks the most

---

<sup>3</sup>Actually, there always exists some basis for any infinite-dimensional vector space due to the *Zorn's Lemma*, or equivalently the *Hausdorff's Maximal Principle*.

<sup>4</sup>We will only justify (4) in Definition 12.1.1 and leave the remaining axioms to the readers. It is obvious that if  $f(x) = 0$  is the zero function then  $\langle f, f \rangle = \int_a^b (0)^2 dx = 0$ . Assume that  $f(x)$  is not everywhere zero, then by continuity from elementary calculus, we know that  $f(c) \neq 0$  and hence  $|f(c)|^2 > 0$  at some point  $a \leq c \leq b$  and there exists  $\delta > 0$  such that  $|f(x)|^2 > \frac{|f(c)|^2}{2}$  where  $c - \delta \leq x \leq c + \delta$ , therefore  $\langle f, f \rangle \geq \frac{|f(c)|^2}{2}(2\delta) = |f(c)|^2\delta > 0$ .

desirable attribute of being a **Hilbert space**. The detailed explanation of how Hilbert spaces work belongs to the area of Functional Analysis and is very much out of the scope of this book<sup>5</sup>, and we will take the liberty to assume that an infinite-dimensional inner product space we work on is a *separable* Hilbert space from time to time. For instance, the  $L^2[a, b]$  space for all square-integrable functions<sup>6</sup> equipped with the same inner product in Equation (12.1) is such a Hilbert space. The reason why we care so much about Hilbert spaces, particularly those are separable, is that they always admit a countable orthonormal basis that is complete.

**Properties 12.1.8.** An infinite-dimensional separable Hilbert space  $\mathcal{H}$  always have a countably infinite orthonormal basis  $\{\varphi_j\}_{j=1}^{\infty}$  where  $\varphi_j$  denotes the  $j$ -th basis vector and  $j$  is an integer enumerated from 1 to infinity. It is *complete* in the sense that there exists no more other non-zero vector  $\tilde{\varphi}$  can be included in the basis such that  $\langle \tilde{\varphi}, \varphi_j \rangle = 0$  for all  $j$  without making the set linearly dependent.

Equivalently, it means that any vector/function in the Hilbert space can be expanded into an infinite sum of orthonormal vectors  $f = c_1\varphi_1 + c_2\varphi_2 + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n c_j \varphi_j$ .<sup>7</sup> For these infinite-dimensional vector spaces, we often loosen the restriction so that any infinite sum of their basis vectors also makes up the span. This is known as a *Schauder basis*. As noted before, the formal treatment of Hilbert spaces is out of our reach and we will invoke the relevant properties as we see fit.

Back to infinite-dimensional inner product spaces in general, the properties and theorems given earlier this section still hold for them as we have defined inner products in a way without regard to the (in)finiteness of dimensions.

<sup>5</sup>To put shortly, a Hilbert space is a *complete* inner product space where all *Cauchy sequences* are convergent. The meaning of complete here is different from that in Properties 12.1.8.

<sup>6</sup>A square-integrable function  $f$  means that  $\int |f|^2 < \infty$  is finite, and the  $L^2[a, b]$  space is actually the "completed" (in the sense of the footnote above) version of  $C^0[a, b]$ .

<sup>7</sup>(Note: again this only holds if the Hilbert space is *separable*, but we will not go into the details.) If there exists such a vector  $\tilde{\varphi}$  that satisfies  $\langle \tilde{\varphi}, \varphi_j \rangle = 0$  for all  $j$ , then consider  $f = \tilde{\varphi}$  and take the inner product with  $\tilde{\varphi}$  on both sides of  $\varphi = c_1\varphi_1 + c_2\varphi_2 + \dots$ . All the terms on the R.H.S. will become zero but the L.H.S. is  $\langle \tilde{\varphi}, \tilde{\varphi} \rangle > 0$  by the last axiom in Definition 12.1.1, a contradiction.

**Example 12.1.2.** Verify that the Cauchy-Schwarz Inequality (Theorem 12.1.6) holds for  $\varphi_1 = x$  and  $\varphi_2 = x^2$  in the  $C^0[0, 1]$  space where the inner product is defined by Equation (12.1).

*Solution.* This is to check

$$\begin{aligned} |\langle \varphi_1, \varphi_2 \rangle| &\leq \|\varphi_1\| \|\varphi_2\| \\ |\langle x, x^2 \rangle| &\leq \|x\| \|x^2\| \end{aligned}$$

by computing the three quantities in the above inequality:

$$\begin{aligned} |\langle x, x^2 \rangle| &= \left| \int_0^1 x \overline{x^2} dx \right| \\ &= \left| \int_0^1 x^3 dx \right| \\ &= \left| \left[ \frac{1}{4} x^4 \right]_0^1 \right| = \left| \frac{1}{4} \right| = \frac{1}{4} \\ \|x\| &= \sqrt{\int_0^1 x \overline{x} dx} \\ &= \sqrt{\int_0^1 x^2 dx} \\ &= \sqrt{\left[ \frac{1}{3} x^3 \right]_0^1} = \sqrt{\frac{1}{3}} = \frac{1}{\sqrt{3}} \end{aligned}$$

similarly  $\|x^2\| = \frac{1}{\sqrt{5}}$ . Hence  $|\langle x, x^2 \rangle| = \frac{1}{4} = \frac{1}{\sqrt{16}} < \frac{1}{\sqrt{15}} = (\frac{1}{\sqrt{3}})(\frac{1}{\sqrt{5}}) = \|x\| \|x^2\|$  and the inequality holds in this case.  $\square$

Short Exercise: Show that the same<sup>8</sup> two functions  $\varphi_1 = x$  and  $\varphi_2 = x^2$  become orthogonal to each other in  $C^0[-1, 1]$  where the inner product still takes the same form of Equation (12.1) but integrated from  $-1$  to  $1$  instead.<sup>9</sup>

<sup>8</sup>Actually they are not the same functions as before since their domain changes from  $[0, 1]$  to  $[-1, 1]$ , but we use the word for the sake of convenience.

<sup>9</sup> $\int_{-1}^1 x \overline{x^2} = \int_{-1}^1 x^3 = 0$  as  $x^3$  is an odd function and the integration limits are symmetric.

## 12.2 Adjoints and Hermitian/Unitary Operators

### 12.2.1 Definition of Adjoints

With the usual (complex) dot product as the inner product for  $\mathbb{R}^n$  ( $\mathbb{C}^n$ ), we have  $\langle A\vec{u}, \vec{v} \rangle = \langle \vec{u}, A^T \vec{v} \rangle$  ( $\langle \vec{u}, A^* \vec{v} \rangle$ ) by Properties 4.2.3 (8.2.7). Here the square matrix  $A$  in the first argument can be moved to the second argument by applying a (conjugate) transpose on it. In this context,  $A^T$  ( $A^*$ ) is known as the *adjoint* of  $A$  with respect to the dot product. Since any square matrix essentially represents a linear operator behind the scene (see Chapter 7), we can extend this idea for a vector space  $\mathcal{V}$  in general, where  $\vec{u}, \vec{v} \in \mathcal{V}$  and the linear operator is  $T : \mathcal{V} \rightarrow \mathcal{V}$  now. Subsequently, the adjoint of  $T$  will then be another linear operator  $T^*$  defined by the relationship  $\langle T(\vec{u}), \vec{v} \rangle = \langle \vec{u}, T^*(\vec{v}) \rangle$ . Again, if the standard inner product is employed, then we know that if the matrix representation of  $T$  in the basis  $\beta$  is  $[T]_\beta$ , then

$$\begin{aligned}\langle T(\vec{u}), \vec{v} \rangle &\equiv \langle [T]_\beta[\vec{u}]_\beta, [\vec{v}]_\beta \rangle \\ &= \langle [\vec{u}]_\beta, ([T]_\beta)^*[\vec{v}]_\beta \rangle \stackrel{\text{def}}{\equiv} \langle \vec{u}, T^*(\vec{v}) \rangle\end{aligned}$$

so that we identify the adjoint operator  $T^*$  with a conjugate transpose matrix representation of  $([T]_\beta)^*$ .

**Definition 12.2.1** (Adjoint). The adjoint of a linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  with respect to some inner product  $\langle \cdot, \cdot \rangle$  is another linear operator denoted by  $T^*$  such that

$$\langle T(\vec{u}), \vec{v} \rangle = \langle \vec{u}, T^*(\vec{v}) \rangle$$

holds for all  $\vec{u}, \vec{v} \in \mathcal{V}$ . Such an adjoint  $T^*$  is always unique for a given  $T$ .

It is easy to check the linearity and uniqueness of an adjoint.<sup>10</sup> Now let's look at how adjoint is found when inner products other than the standard one are used.

<sup>10</sup>Linearity (Definition 7.1.1): Consider  $\langle \vec{u}, T^*(a\vec{v} + b\vec{w}) \rangle = \langle T(\vec{u}), a\vec{v} + b\vec{w} \rangle = a\langle T(\vec{u}), \vec{v} \rangle + b\langle T(\vec{u}), \vec{w} \rangle$  by Definitions 12.2.1 and 12.1.1. Subsequently  $a\langle T(\vec{u}), \vec{v} \rangle + b\langle T(\vec{u}), \vec{w} \rangle = a\langle \vec{u}, T^*(\vec{v}) \rangle + b\langle \vec{u}, T^*(\vec{w}) \rangle = \langle \vec{u}, aT^*(\vec{v}) \rangle + \langle \vec{u}, bT^*(\vec{w}) \rangle = \langle \vec{u}, aT^*(\vec{v}) + bT^*(\vec{w}) \rangle$  by those

First, if  $\mathcal{V}$  is a finite-dimensional complex vector space and a new inner product is defined via a positive-definite Hermitian form  $\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v} = (B^T \vec{u}) \cdot \vec{v}$ , then for a linear operator  $T$  with a matrix representation of  $[T]_\beta$  in the  $\beta$  system, by Definition 8.2.1 and Properties 8.2.7:

$$\begin{aligned}\langle T(\vec{u}), \vec{v} \rangle &\equiv (B^T [T]_\beta [\vec{u}]_\beta) \cdot ([\vec{v}]_\beta) \\ &= (B^T [T]_\beta (B^T)^{-1} B^T [\vec{u}]_\beta) \cdot ([\vec{v}]_\beta) \\ &= (B^T [\vec{u}]_\beta) \cdot ((B^T [T]_\beta (B^T)^{-1})^* [\vec{v}]_\beta) \\ &\equiv \langle \vec{u}, (B^T [T]_\beta (B^T)^{-1})^* \vec{v} \rangle = \langle \vec{u}, (\bar{B})^{-1} [T]_\beta^* \bar{B} \vec{v} \rangle\end{aligned}$$

and thus we identify the matrix representation of the adjoint as  $T^* \equiv (\bar{B})^{-1} [T]_\beta^* \bar{B}$  (notice that  $[T]_\beta^*$  and  $[T^*]_\beta$  are not the same) with respect to such an inner product. For an infinite-dimensional inner product space, the inner product used is usually in the form of an integral like the one in Equation (12.1), additionally with a weighting function, and a linear operator  $T$  can be more general, including differentiation and multiplication by some function, or a mix of them. As a result, the adjoint is often found via the technique of *integration by parts*, producing boundary terms as a by-product. Conventionally, we will use  $\mathcal{L}$  in place of  $T$  to denote the operator when the vectors are functions.

**Properties 12.2.2.** For a general integral inner product:

$$\langle f, g \rangle = \int_a^b w(x) f(x) \overline{g(x)} dx \quad (12.2)$$

where  $w(x) > 0$  is a positive-definite *real* weighting function, the unique adjoint  $\mathcal{L}^*$  of a linear operator  $\mathcal{L}$  is another one that satisfies

$$\langle f, \mathcal{L}[g] \rangle = \int_a^b w(x) f(x) \overline{(\mathcal{L}[g(x)])} dx$$

---

definitions again. Since this has to hold for any  $\vec{u}$  (plus  $\vec{v}$  and  $\vec{w}$ ),  $\langle \vec{u}, T^*(a\vec{v} + b\vec{w}) \rangle = \langle \vec{u}, aT^*(\vec{v}) + bT^*(\vec{w}) \rangle$  means that  $T^*(a\vec{v} + b\vec{w}) = aT^*(\vec{v}) + bT^*(\vec{w})$  by Properties 12.1.2. Uniqueness: Assume that there is another adjoint  $S^*$  satisfies  $\langle T(\vec{u}), \vec{v} \rangle = \langle \vec{u}, S^*(\vec{v}) \rangle$  for all  $\vec{u}$  and  $\vec{v}$ , then by Properties 12.1.2  $T^*(\vec{v}) = S^*(\vec{v})$  for all  $\vec{v}$ . Hence  $T^*$  and  $S^*$  must be the same operator.

$$\begin{aligned}
 &= \int_a^b w(x) \mathcal{L}^*[f(x)] \overline{g(x)} dx + \text{boundary terms} \\
 &= \langle \mathcal{L}^*[f], g \rangle + \text{boundary terms}
 \end{aligned}$$

for all functions  $f, g \in \mathcal{V}$  and the boundary terms are evaluated at the end-points  $a$  and  $b$ .

Equation (12.1) is then simply a special case of (12.2) with a constant weight of  $w = 1$ .

**Example 12.2.1.** Find the adjoint of the linear operator  $\mathcal{L}[f] = x \frac{d}{dx}[f]$  with respect to the inner product in Equation (12.1).

*Solution.* We start with  $\langle f, \mathcal{L}[g] \rangle = \int_a^b f(x) \overline{\left( x \frac{d}{dx}(g(x)) \right)} dx$  and aim to rewrite it into the form of  $\int_a^b \mathcal{L}^*[f(x)] \overline{g(x)} dx = \langle \mathcal{L}^*[f], g \rangle$ , plus possibly some boundary term(s). As suggested by above we can try to apply integration by parts:

$$\begin{aligned}
 \langle f, \mathcal{L}[g] \rangle &= \int_a^b f(x) \overline{\left( x \frac{d}{dx}(g(x)) \right)} dx \\
 &= \int_a^b x f(x) \frac{d}{dx} \overline{(g(x))} dx \\
 &= [x f(x) \overline{(g(x))}]_a^b - \int_a^b \frac{d}{dx}(x f(x)) \overline{(g(x))} dx \\
 &= \int_a^b -\frac{d}{dx}(x f(x)) \overline{(g(x))} dx + [x f(x) \overline{(g(x))}]_a^b
 \end{aligned}$$

After rearrangement, the boundary term is put after the integral, from which we deduce that  $\mathcal{L}^*[f] = -\frac{d}{dx}(x[f])$  by comparing it with  $\int_a^b \mathcal{L}^*[f(x)] \overline{(g(x))} dx$ .  $\square$

Finally we note some properties of adjoints that can be compared to Properties 8.2.6.

**Properties 12.2.3.** For two linear operators  $T$  and  $U$  in the same inner product space with adjoints  $T^*$  and  $U^*$  respectively, we have

1.  $(cT)^* = \bar{c}T^*$ , where  $c$  is any complex scalar,
2.  $(T^*)^* = T$ ,
3.  $(T \pm U)^* = T^* \pm U^*$ ,
4.  $(T^*)^{-1} = (T^{-1})^*$  if  $T$  (hence  $T^*$ ) is invertible,
5.  $(TU)^* = U^*T^*$ .

We will briefly show the last item here. Using Definition 12.2.1 twice, we have

$$\begin{aligned}\langle TU(\vec{u}), \vec{v} \rangle &= \langle U(\vec{u}), T^*(\vec{v}) \rangle \\ &= \langle \vec{u}, U^*T^*(\vec{v}) \rangle\end{aligned}$$

so we identify that  $(TU)^* = U^*T^*$ .

## 12.2.2 Hermitian Operators

As suggested by the last subsection, adjoints are the inner product counterpart of (conjugate) transposes. Therefore, it is natural to ask if the concept of symmetric or Hermitian in the matrix world is also applicable to an adjoint. Correspondingly, when a linear operator  $T$  has an adjoint  $T^*$  which is equal to itself, i.e.  $T^* = T$ , it is called **self-adjoint**, as the inner product equivalent of Hermitian. If the inner product space is finite-dimensional with a positive-definite Hermitian form  $B$  as its inner product, then according to what we have just derived we simply need to check if  $T^* \equiv (\overline{B})^{-1}[T]_\beta^*\overline{B} = [T]_\beta \equiv T$ . The problem becomes a bit more complicated when we consider the integral inner product in Equation (12.2) because even when  $\mathcal{L}^* = \mathcal{L}$  is self-adjoint there can be boundary terms. We need an even stronger condition where there is no boundary term so that  $\mathcal{L}$  becomes "nicer" to work with and gives desirable

properties. In this situation, the self-adjoint  $\mathcal{L}$  is further known as a **Hermitian** operator.

**Definition 12.2.4.** A linear operator  $T$  is self-adjoint if its adjoint  $T^* = T$  equals to itself. A linear operator  $\mathcal{L}$  is Hermitian if it is self-adjoint and all boundary terms vanish.

The absence of any boundary term can be due to the structure of  $\mathcal{L}$  or  $\mathcal{L}^*$  itself, or the boundary condition(s) imposed on the input functions.

**Example 12.2.2.** Show that  $\mathcal{L} = -i\hbar \frac{d}{dx}$ , where  $\hbar$  is a real constant, is a self-adjoint operator with respect to the inner product in Equation (12.1) over the entire  $x$ -axis. Find the form of its eigenfunctions (treating functions as (eigen)vectors).

*Solution.*

$$\begin{aligned}
 \langle f, \mathcal{L}[g] \rangle &= \int_{-\infty}^{\infty} f(x) \overline{(\mathcal{L}[g(x)])} dx \\
 &= \int_{-\infty}^{\infty} f(x) \overline{\left( -i\hbar \frac{d}{dx}(g(x)) \right)} dx \\
 &= \int_{-\infty}^{\infty} i\hbar f(x) \frac{d}{dx}(g(x)) dx \\
 &= [i\hbar f(x)g(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} i\hbar \frac{d}{dx}(f(x))g(x) dx \quad (\text{Integration by parts}) \\
 &= \int_{-\infty}^{\infty} -i\hbar \frac{d}{dx}(f(x))g(x) dx + [i\hbar f(x)g(x)]_{-\infty}^{\infty} \\
 &= \int_{-\infty}^{\infty} \mathcal{L}[f(x)]g(x) dx + [i\hbar f(x)g(x)]_{-\infty}^{\infty}
 \end{aligned}$$

So we see that  $\mathcal{L}^* = \mathcal{L}$  is self-adjoint. The eigenfunctions  $\varphi$  of  $\mathcal{L}$  can be found from solving the ODE:

$$\mathcal{L}[\varphi] = -i\hbar \frac{d\varphi}{dx} = m\varphi$$

$$\begin{aligned} - \int i\hbar \frac{d\varphi}{\varphi} &= \int m dx \\ -i\hbar \ln \varphi &= mx + C \\ \therefore \varphi &= Ke^{\frac{i}{\hbar}mx} \end{aligned}$$

where  $K = e^{\frac{i}{\hbar}C}$  is any scaling constant. Notice that the eigenvalue  $m$  can take any value ranging from  $-\infty$  to  $\infty$  as there is no boundary condition being enforced. In this case, the eigenvalues form a *continuous spectrum* and we note that the vector space generated by those eigenfunctions, which are uncountable in this problem, are not separable. In fact, this operator  $\mathcal{L} = -i\hbar \frac{d}{dx}$ , is famously referred to as the *momentum operator* in Quantum Mechanics.  $\square$

The above example shows that eigenfunctions for a self-adjoint linear operator may be uncountable. In this case, a function cannot be represented by an infinite sum  $f = c_1\varphi_1 + c_2\varphi_2 + \dots = \sum_{j=1}^{\infty} c_j\varphi_j$  as suggested by Properties 12.1.8 but rather requires to be expressed by an integral  $f(x) = \int_0^{\infty} c_m\varphi_m(x)dm$  over a continuous index  $m$ , which is much more troublesome to deal with. This is the exact reason why we introduce Hermitian operators in this part, as they carry the desirable properties of possessing countably infinite eigenfunctions that form a complete orthonormal basis and a *discrete spectrum*, provided that the Hilbert space where the Hermitian operator acts on is already separable. The completeness of the eigenfunctions of a Hermitian operator, again should be delegated to the world of Functional Analysis. But first, we can show an essential property of Hermitian operators that their eigenvalues are always real, which can be compared to Properties 10.4.6.

**Properties 12.2.5.** The eigenvalues of any Hermitian operator  $T^* = T$  must be real.

*Proof.* (WIP)  $\square$

Moreover, we can briefly prove the orthogonality between its eigenfunctions.

**Properties 12.2.6.** Any two eigenfunctions of a Hermitian operator  $\mathcal{L}^* = \mathcal{L}$  corresponding to two distinct eigenvalues are always orthogonal to each other.

This closely parallels Properties 10.2.3 and the proof is also very similar.

*Proof.* Denote the two eigenfunctions as  $\varphi_1$  and  $\varphi_2$  where the respective eigenvalues are  $m_1$  and  $m_2$ . Subsequently we have

$$\begin{aligned}\langle \varphi_1, \mathcal{L}[\varphi_2] \rangle &= \langle \varphi_1, m_2 \varphi_2 \rangle \\ &= \bar{m}_2 \langle \varphi_1, \varphi_2 \rangle = m_2 \langle \varphi_1, \varphi_2 \rangle\end{aligned}\quad (\text{Properties 12.2.5})$$

but also

$$\begin{aligned}\langle \varphi_1, \mathcal{L}[\varphi_2] \rangle &= \langle \mathcal{L}^*[\varphi_1], \varphi_2 \rangle && (\text{Properties 12.2.2}) \\ &= \langle \mathcal{L}[\varphi_1], \varphi_2 \rangle && (\text{Definition 12.2.4}) \\ &= \langle m_1 \varphi_1, \varphi_2 \rangle \\ &= m_1 \langle \varphi_1, \varphi_2 \rangle\end{aligned}$$

So

$$\begin{aligned}m_1 \langle \varphi_1, \varphi_2 \rangle &= m_2 \langle \varphi_1, \varphi_2 \rangle \\ (m_1 - m_2) \langle \varphi_1, \varphi_2 \rangle &= 0\end{aligned}$$

but the two eigenvalues are taken to be distinct,  $m_1 \neq m_2$ , hence it must be that  $\langle \varphi_1, \varphi_2 \rangle = 0$  and  $\varphi_1, \varphi_2$  are orthogonal with respect to the inner product (Properties 12.1.5).  $\square$

We can then divide the eigenfunctions by their norm as in Definition 12.1.4 to make them have unit length and hence become orthonormal. Even when some of the  $k$  eigenfunctions have the same eigenvalue (in this context they are said to be *k-fold degenerate*), we can apply the Gram-Schmidt Orthogonalization process adapted for inner product space (to be discussed soon in Section 12.3.2), over those eigenfunctions, in a fashion very similar to the idea suggested in the discussion below Properties 10.2.3. Now let's see a very standard example of a Hermitian operator leading to a complete set of orthonormal eigenfunctions.

**Example 12.2.3.** For the separable Hilbert space  $L^2[-\pi, \pi]$  of square-integrable functions along  $-\pi \leq x \leq \pi$ , show that the linear operator  $\mathcal{L}[f] = \frac{d^2}{dx^2}[f]$  will be Hermitian with respect to the inner product in Equation (12.1) if the (eigen)functions are picked in a way so that the boundary terms vanish.

*Solution.* First, we have to check if  $\mathcal{L}$  is self-adjoint, from

$$\begin{aligned}
 \langle f, \mathcal{L}[g] \rangle &= \int_{-\pi}^{\pi} f(x) \overline{(\mathcal{L}[g(x)])} dx \\
 &= \int_{-\pi}^{\pi} f(x) \overline{\frac{d^2}{dx^2}([g(x)])} dx \\
 &= \int_{-\pi}^{\pi} f(x) \frac{d^2}{dx^2} \overline{([g(x)])} dx \\
 &= [f(x) \frac{d}{dx} \overline{([g(x)])}]_{-\pi}^{\pi} - \int_{-\pi}^{\pi} \frac{d}{dx}(f(x)) \frac{d}{dx} \overline{([g(x)])} dx \\
 &= [f(x) \frac{d}{dx} \overline{([g(x)])}]_{-\pi}^{\pi} - [\frac{d}{dx}(f(x)) \overline{g(x)}]_{-\pi}^{\pi} \\
 &\quad + \int_{-\pi}^{\pi} \frac{d^2}{dx^2}(f(x)) \overline{g(x)} dx \\
 &= [f(x) \frac{d}{dx} \overline{([g(x)])}]_{-\pi}^{\pi} - [\frac{d}{dx}(f(x)) \overline{g(x)}]_{-\pi}^{\pi} \\
 &\quad + \int_{-\pi}^{\pi} \mathcal{L}(f(x)) \overline{g(x)} dx
 \end{aligned}$$

and thus  $\mathcal{L}^* = \frac{d^2}{dx^2} = \mathcal{L}$ . Again from basic ODE we know that the eigenfunctions  $\varphi$  of  $\mathcal{L}[f] = \frac{d^2}{dx^2}[f]$  will be in the form of  $\sin(mx)$  and  $\cos(mx)$  as  $\frac{d^2}{dx^2}(\sin(mx)) = -m^2 \sin(mx)$  and  $\frac{d^2}{dx^2}(\cos(mx)) = -m^2 \cos(mx)$ . For the two boundary terms to vanish,  $m = 0, 1, 2, \dots$  must be a non-negative integer<sup>11</sup>, and

---

<sup>11</sup>Without loss of generality, let  $f = \sin(m_1 x)$  and  $g = \cos(m_2 x)$ , then

$$[f(x) \frac{d}{dx} \overline{([g(x)])}]_{-\pi}^{\pi} - [\frac{d}{dx}(f(x)) \overline{g(x)}]_{-\pi}^{\pi}$$

hence the countably infinite, orthogonal<sup>12</sup> (not yet normalized) basis of eigenfunctions are  $\{\sin(x), \sin(2x), \sin(3x), \dots, 1, \cos(x), \cos(2x), \cos(3x), \dots\}$ . This basis, consisting of sines and cosines with discrete, equally spaced frequencies, is famously known as the **Fourier basis**.  $\square$

### 12.2.3 Unitary Operators

Another class of matrices we want to generalize for inner product spaces is the orthogonal/unitary one introduced in Chapter 10. For those matrices, let's say  $A$ , two defining properties are that  $A^{-1} = A^*$  and the preservation of distance noted in Properties 10.1.5. Therefore, we want a unitary operator  $T$  to satisfy the same properties where  $T^*$  is now its adjoint and the notion of distance is with respect to the given inner product.

**Definition 12.2.7.** A unitary operator  $T$  is a linear operator such that its inverse operator equals to its adjoint  $T^{-1} = T^*$  with respect to the inner product used.

With this definition, the distance-preserving property can be readily derived.

$$\begin{aligned} &= [-m_2 \sin(m_1 x) \sin(m_2 x)]_{-\pi}^{\pi} - [m_1 \cos(m_1 x) \cos(m_2 x)]_{-\pi}^{\pi} \\ &= \left[ -\frac{m_2}{2} (\cos((m_1 - m_2)x) - \cos((m_1 + m_2)x)) \right]_{-\pi}^{\pi} \\ &\quad - \left[ \frac{m_1}{2} (\cos((m_1 - m_2)x) + \cos((m_1 + m_2)x)) \right]_{-\pi}^{\pi} \end{aligned}$$

using trigonometric identities. From this we see that if the two boundary terms have to be zero,  $m_1 - m_2$  and  $m_1 + m_2$  must both be odd or even at the same time, which implies that  $m_1$  and  $m_2$  have to be integers. And since  $\sin(-m_1 x) = -\sin(m_1 x)$  and  $\cos(-m_2 x) = \cos(m_2 x)$ , we only need to take the integers that are positive.

<sup>12</sup>Again, without the loss of generality, we let  $f = \sin(m_1 x)$  and  $g = \cos(m_2 x)$  where  $m_1$  and  $m_2$  are now positive integers. Then the orthogonality is verified by computing

$$\begin{aligned} \int_{-\pi}^{\pi} \sin(m_1 x) \cos(m_2 x) dx &= \int_{-\pi}^{\pi} \frac{1}{2} (\sin((m_1 - m_2)x) + \sin((m_1 + m_2)x)) dx \\ &= \left[ \frac{1}{2} \left( -\frac{\cos((m_1 - m_2)x)}{m_1 - m_2} - \frac{\cos((m_1 + m_2)x)}{m_1 + m_2} \right) \right]_{-\pi}^{\pi} \end{aligned}$$

which yields 0 when the end-points  $-\pi, \pi$  are substituted into the expression.

**Properties 12.2.8.** Transformation by a unitary operator  $T$  on a vector is length-preserving.

which should be compared to Properties 10.1.5.

*Proof.* Denote the original vector as  $\vec{v}$  and the newly transformed vector be  $T(\vec{v})$ , then its length, given as the norm in Properties 12.1.3,

$$\begin{aligned}\|T(\vec{v})\|^2 &= \langle T(\vec{v}), T(\vec{v}) \rangle = \langle \vec{v}, T^*(T(\vec{v})) \rangle && \text{(Definition 12.2.1)} \\ &= \langle \vec{v}, T^{-1}(T(\vec{v})) \rangle && \text{(Definition 12.2.7)} \\ &= \langle \vec{v}, \text{id}(\vec{v}) \rangle && \text{(Definition 7.1.10)} \\ &= \langle \vec{v}, \vec{v} \rangle = \|\vec{v}\|^2\end{aligned}$$

is shown to be equal throughout the unitary transformation.  $\square$

**Example 12.2.4.** Show that for the  $L^2[a, b]$  space with the inner product of Equation (12.1), the linear operator  $\mathcal{L}[f] = e^{ikx}[f]$  where  $k$  is any real number, is unitary.

*Solution.* It is apparent that the inverse of  $\mathcal{L}$  is  $\mathcal{L}^{-1}[f] = e^{-ikx}[f]$  so that  $(\mathcal{L}^{-1} \circ \mathcal{L})[f] = e^{-ikx}(e^{ikx}[f]) = (1)f = \text{id}[f]$ , and we have to show that it equals to the adjoint  $\mathcal{L}^*$ :

$$\begin{aligned}\langle f, \mathcal{L}[g] \rangle &= \int_a^b f(x) \overline{\mathcal{L}[g(x)]} dx \\ &= \int_a^b f(x) \overline{e^{ikx} g(x)} dx \\ &= \int_a^b e^{-ikx} f(x) \overline{g(x)} dx && (\overline{e^{ikx}} = e^{-ikx}) \\ &= \int_a^b \mathcal{L}^{-1}[f(x)] \overline{g(x)} dx\end{aligned}$$

From this we readily infer that  $\mathcal{L}^{-1} = \mathcal{L}^*$  is unitary.  $\square$

## 12.3 Revisiting Orthogonal Projections

### 12.3.1 Orthogonal Projections for an Inner Product Space

Since we have defined orthogonality and adjoints, we are now ready to derive the form of orthogonal projections for any inner product space, ultimately allowing us to establish the inner product space version of Spectral Theorem. Remember in Section 5.2.1 we derive the orthogonal projection of a vector onto another vector with the dot product, and in the same essence we can obtain the expression of any one-dimensional orthogonal projection with respect to an inner product by simply replacing the dot product with that inner product.

**Definition 12.3.1.** The orthogonal projection of a vector  $\vec{v}$  onto  $\vec{u}$  with respect to an inner product  $\langle \cdot, \cdot \rangle$  is

$$\overrightarrow{\text{proj}}_{\vec{u}} \vec{v} = \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{u}\|^2} \vec{u}$$

This definition is consistent in the sense that  $\vec{u}$  and the component of  $\vec{v}$  normal to  $\vec{u}$  are orthogonal:

$$\begin{aligned} \langle \vec{u}, \vec{v} - \overrightarrow{\text{proj}}_{\vec{u}} \vec{v} \rangle &= \langle \vec{u}, \vec{v} - \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{u}\|^2} \vec{u} \rangle \\ &= \langle \vec{u}, \vec{v} \rangle - \langle \vec{u}, \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{u}\|^2} \vec{u} \rangle \\ &= \langle \vec{u}, \vec{v} \rangle - \frac{\overline{\langle \vec{v}, \vec{u} \rangle}}{\|\vec{u}\|^2} \langle \vec{u}, \vec{u} \rangle \quad (\text{Properties 12.1.2}) \\ &= \langle \vec{u}, \vec{v} \rangle - \frac{\langle \vec{u}, \vec{v} \rangle}{\|\vec{u}\|^2} \|\vec{u}\|^2 \\ &= \langle \vec{u}, \vec{v} \rangle - \langle \vec{u}, \vec{v} \rangle = 0 \end{aligned}$$

Similar to the expression (10.1), the orthogonal projection operator  $T$  onto a subspace  $\mathcal{W} \subseteq \mathcal{V}$  of an inner product space with an orthonormal basis

$\{\vec{w}^{(1)}, \vec{w}^{(2)}, \dots\}$  that may be finite or countably infinite, is the sum of the one-dimensional projectors onto each of the basis vectors according to the definition above:

$$T(\vec{v}) = \frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \dots \quad (12.3)$$

Similar to Properties 10.3.1 and 10.3.2, a linear operator  $T$  represents an orthogonal projection if and only if it has an adjoint  $T^*$  so that  $T^2 = T = T^*$ .

**Properties 12.3.2.** A linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  over an inner product space, is an orthogonal projection with respect to the inner product if and only if  $T^2 = T = T^*$  where  $T^*$  is its adjoint.

The proof for the first equality  $T^2 = T$  is the same one in Properties 10.3.1 except that  $\mathcal{V}$ , as well as  $\mathcal{W}_1$  and  $\mathcal{W}_2$  may be infinite-dimensional now. Meanwhile for the second equality  $T = T^*$ , we need to replace the dot product in the original proof for Properties 10.3.2 by the inner product, and additionally show that  $\mathcal{N}(T)^\perp = \mathcal{R}(T)$  for the "if" part when  $\mathcal{V}$  is infinite-dimensional, which is not too far from our reach.<sup>13</sup> We now show that the expression of the orthogonal projector  $T$  given in Equation (12.3) satisfies the above requirement  $T^2 = T = T^*$ :

$$\begin{aligned} T^2(\vec{v}) &= T(T(\vec{v})) = \frac{\langle T(\vec{v}), \vec{w}^{(1)} \rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\langle T(\vec{v}), \vec{w}^{(2)} \rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \dots \\ &= \frac{\left\langle \frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \dots, \vec{w}^{(1)} \right\rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} \end{aligned}$$

---

<sup>13</sup>We take it for granted that  $\mathcal{W} \subseteq \mathcal{W}^{\perp\perp}$  for any subspace  $\mathcal{W}$ , so that  $\mathcal{N}(T)^\perp = \mathcal{R}(T)^{\perp\perp} \supseteq \mathcal{R}(T)$ , and the remaining task is to show  $\mathcal{N}(T)^\perp \subseteq \mathcal{R}(T)$  so that  $\mathcal{N}(T)^\perp = \mathcal{R}(T)$ , i.e. for any  $\vec{w} \in \mathcal{N}(T)^\perp$ ,  $\vec{w} \in \mathcal{R}(T)$ . Consider  $\|\vec{w} - T(\vec{w})\|^2 = \langle \vec{w} - T(\vec{w}), \vec{w} - T(\vec{w}) \rangle = \langle \vec{w}, \vec{w} - T(\vec{w}) \rangle - \langle T(\vec{w}), \vec{w} - T(\vec{w}) \rangle$ . The first term is zero since  $\vec{w} - T(\vec{w}) \in \mathcal{N}(T)$  is orthogonal to  $\vec{w} \in \mathcal{N}(T)^\perp$ , and the second term is also zero as  $\langle T(\vec{w}), \vec{w} - T(\vec{w}) \rangle = \langle \vec{w}, T^*(\vec{w} - T(\vec{w})) \rangle = \langle \vec{w}, T(\vec{w} - T(\vec{w})) \rangle = \langle \vec{w}, T(\vec{w}) - T^2(\vec{w}) \rangle = \langle \vec{w}, \mathbf{0} \rangle = 0$  since  $T^* = T$  by the "if" condition and  $T^2 = T$  from the first part. Thus  $\|\vec{w} - T(\vec{w})\| = 0$  and this implies that  $T(\vec{w}) = \vec{w}$  so  $\vec{w} \in \mathcal{R}(T)$ .

$$\begin{aligned}
 & + \frac{\left\langle \vec{v}, \vec{w}^{(1)} \right\rangle \vec{w}^{(1)} + \left\langle \vec{v}, \vec{w}^{(2)} \right\rangle \vec{w}^{(2)} + \cdots, \vec{w}^{(2)} \right\rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \cdots \\
 & = \frac{\frac{\left\langle \vec{v}, \vec{w}^{(1)} \right\rangle}{\|\vec{w}^{(1)}\|^2} \|\vec{w}^{(1)}\|^2 + (0)}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} \\
 & \quad + \frac{(0) + \frac{\left\langle \vec{v}, \vec{w}^{(2)} \right\rangle}{\|\vec{w}^{(2)}\|^2} \|\vec{w}^{(2)}\|^2 + (0)}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \cdots \\
 & \quad (\vec{w}^{(j)} \text{ are orthogonal}) \\
 & = \frac{\left\langle \vec{v}, \vec{w}^{(1)} \right\rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\left\langle \vec{v}, \vec{w}^{(2)} \right\rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \cdots = T(\vec{v})
 \end{aligned}$$

and

$$\begin{aligned}
 \langle \vec{u}, T(\vec{v}) \rangle &= \langle \vec{u}, \frac{\left\langle \vec{v}, \vec{w}^{(1)} \right\rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\left\langle \vec{v}, \vec{w}^{(2)} \right\rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \cdots \rangle \\
 &= \frac{\left\langle \vec{v}, \vec{w}^{(1)} \right\rangle}{\|\vec{w}^{(1)}\|^2} \langle \vec{u}, \vec{w}^{(1)} \rangle + \frac{\left\langle \vec{v}, \vec{w}^{(2)} \right\rangle}{\|\vec{w}^{(2)}\|^2} \langle \vec{u}, \vec{w}^{(2)} \rangle + \cdots \\
 &= \frac{\langle \vec{w}^{(1)}, \vec{v} \rangle}{\|\vec{w}^{(1)}\|^2} \langle \vec{u}, \vec{w}^{(1)} \rangle + \frac{\langle \vec{w}^{(2)}, \vec{v} \rangle}{\|\vec{w}^{(2)}\|^2} \langle \vec{u}, \vec{w}^{(2)} \rangle + \cdots
 \end{aligned}$$

but

$$\begin{aligned}
 \langle T(\vec{u}), \vec{v} \rangle &= \left\langle \frac{\langle \vec{u}, \vec{w}^{(1)} \rangle}{\|\vec{w}^{(1)}\|^2} \vec{w}^{(1)} + \frac{\langle \vec{u}, \vec{w}^{(2)} \rangle}{\|\vec{w}^{(2)}\|^2} \vec{w}^{(2)} + \cdots, \vec{v} \right\rangle \\
 &= \frac{\langle \vec{u}, \vec{w}^{(1)} \rangle}{\|\vec{w}^{(1)}\|^2} \langle \vec{w}^{(1)}, \vec{v} \rangle + \frac{\langle \vec{u}, \vec{w}^{(2)} \rangle}{\|\vec{w}^{(2)}\|^2} \langle \vec{w}^{(2)}, \vec{v} \rangle + \cdots \\
 &= \langle \vec{u}, T(\vec{v}) \rangle
 \end{aligned}$$

Hence we identify  $T^*$  with  $T$ .

### 12.3.2 Revisiting Gram-Schmidt Orthogonalization

After obtaining the formula for orthogonal projections with respect to an inner product space, the next step is to extend the method of Gram-Schmidt Orthogonalization so that a countably infinite orthonormal basis can be produced for an infinite-dimensional separable Hilbert space as predicted by Properties 12.1.8, given that the procedure keeps going on indefinitely. Attentive readers should already be able to conceive that we simply have to replace all the dot products in Definition 7.2.3 by an appropriate inner product.

**Definition 12.3.3.** Given a countably infinite basis  $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \dots\}$ , where  $\vec{u}^{(j)} \in \mathcal{V}$  belongs to an inner product space, it is transformed into an orthogonal basis  $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \dots\}$ ,  $\vec{v}^{(j)} \in \mathcal{V}$ , by Gram-Schmidt Orthogonalization according to the following formulae:

$$\vec{v}^{(1)} = \vec{u}^{(1)}$$

$$\vec{v}^{(2)} = \vec{u}^{(2)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(2)} = \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)}$$

$$\begin{aligned}\vec{v}^{(3)} &= \vec{u}^{(3)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(3)} - \text{proj}_{\vec{v}^{(2)}} \vec{u}^{(3)} \\ &= \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)}\end{aligned}$$

 $\vdots$ 

$$\begin{aligned}\vec{v}^{(n)} &= \vec{u}^{(n)} - \text{proj}_{\vec{v}^{(1)}} \vec{u}^{(n)} - \text{proj}_{\vec{v}^{(2)}} \vec{u}^{(n)} - \dots - \text{proj}_{\vec{v}^{(n-1)}} \vec{u}^{(n)} \\ &= \vec{u}^{(n)} - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(2)} \rangle}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)} - \dots - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(n-1)} \rangle}{\|\vec{v}^{(n-1)}\|^2} \vec{v}^{(n-1)}\end{aligned}$$

For  $j \geq 2$ , the  $j$ -th new orthogonal basis vector is computed by

$$\vec{v}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \text{proj}_{\vec{v}^{(k)}} \vec{u}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \frac{\langle \vec{u}^{(j)}, \vec{v}^{(k)} \rangle}{\|\vec{v}^{(k)}\|^2} \vec{v}^{(k)}$$

where the expression of a vector projection in an inner product space now follows

Definition 12.3.1 and  $j$  can be arbitrary large. To make it an orthonormal basis we simply normalize each  $\vec{v}^{(j)}$  by its norm as suggested by Definition 12.1.4.

However, for now we will go through the example of a general, finite-dimensional inner product space first, in which the procedure truncates at the  $n$ -th step where  $n$  is the dimension, and reserve the infinite-dimensional case until the next section.

**Example 12.3.1.** For the  $\mathbb{R}^3$  space with an inner product defined according to the symmetric bilinear form as

$$\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$$

where

$$B = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

apply the Gram-Schmidt Orthogonalization over the standard basis for  $\mathbb{R}^3$ ,  $\hat{e}^{(1)} = (1, 0, 0)^T$ ,  $\hat{e}^{(2)} = (0, 1, 0)^T$ ,  $\hat{e}^{(3)} = (0, 0, 1)^T$ , to transform it into an orthonormal basis.

*Solution.* We leave to the readers to check that the eigenvalues of  $B$  are  $\lambda = 2 - \sqrt{2}, 2, 2 + \sqrt{2}$  all positive so that  $B$  is positive-definite by Theorem 11.1.4 and the inner product makes sense. We calculate each of the expressions appearing in Definition 12.3.3:

$$\begin{aligned} \|\vec{v}^{(1)}\|^2 &= [1 \ 0 \ 0] \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 2 \\ \vec{v}^{(2)} &= \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} \end{aligned}$$

$$\begin{aligned}
 &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \frac{\left( [0 \ 1 \ 0] \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
 &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} \\
 \left\| \vec{v}^{(2)} \right\|^2 &= \begin{bmatrix} -\frac{1}{2} & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} = \frac{3}{2} \\
 \vec{v}^{(3)} &= \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\| \vec{v}^{(1)} \|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\| \vec{v}^{(2)} \|^2} \vec{v}^{(2)} \\
 &= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \frac{\left( [0 \ 0 \ 1] \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
 &\quad - \frac{\left( [0 \ 0 \ 1] \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} \right)}{\frac{3}{2}} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} \\
 &= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - (0) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix} \\
 \left\| \vec{v}^{(3)} \right\|^2 &= \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix} = \frac{4}{3}
 \end{aligned}$$

Therefore, the required orthonormal basis is  $\{(\frac{1}{\sqrt{2}}, 0, 0)^T, (-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, 0)^T, (\frac{1}{2\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{3}{2\sqrt{3}})^T\}$  by dividing each of the  $\vec{v}^{(j)}$  by  $\| \vec{v}^{(j)} \|$ . This example shows that the standard unit vectors are usually not orthogonal to each other when an

inner product other than the standard one is used.  $\square$

### 12.3.3 Spectral Theorem for Hermitian Operators

As in the symmetric matrix case, the Spectral Theorem is the most principal result that can be derived for Hermitian operators, which are the inner product counterpart of symmetric matrices. Its statements are therefore analogous to Theorem 10.3.4 and are listed below.

**Theorem 12.3.4** (Spectral Theorem for Hermitian Operators). For a Hermitian linear operator  $T : \mathcal{V} \rightarrow \mathcal{V}$  where  $T^* = T$  and the inner product space  $\mathcal{V}$  is a separable Hilbert space, denote its eigenvalues by  $\lambda_j$ ,  $j = 1, 2, \dots$ , and  $\varphi^{(j)}$  be the corresponding eigenvectors that are orthonormal (by Properties 12.2.6, or made from the Gram-Schmidt process if necessary). The dimension of this Hilbert space,  $n$ , and hence  $j$ , may be finite (if so, remove the limit  $\lim_{n \rightarrow \infty}$ ) or countably infinite. Refer the one-dimensional eigenspaces generated by each of the  $\varphi^{(j)}$  to as  $\mathcal{E}_j$  and denote the orthogonal projection onto  $\mathcal{E}_j$  by  $T_j$ , then we have:

- (a)  $\mathcal{V} = \mathcal{E}_1 \oplus \mathcal{E}_2 \oplus \dots = \lim_{n \rightarrow \infty} \bigoplus_{j=1}^n \mathcal{E}_j$ ;
- (b)  $(\bigoplus_{j \in J} \mathcal{E}_j)^\perp = \bigoplus_{j \notin J} \mathcal{E}_j$  where  $J$  is a countable index set;
- (c)  $T_j T_{j'} = \begin{cases} T_j & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$ ;
- (d)  $I = T_1 + T_2 + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n T_j$ ; and
- (e)  $T = \lambda_1 T_1 + \lambda_2 T_2 + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n \lambda_j T_j$ , where  $\lambda_j$  are all real numbers.

Note that (a) and subsequently (d) is simply a restatement of the fact that a Hermitian operator has a complete orthonormal basis formed by countably infinite eigenvectors/eigenfunctions which we have taken for granted to avoid getting involved with Functional Analysis. (a) essentially means that any vector  $\vec{v} = c_1 \varphi^{(1)} + c_2 \varphi^{(2)} + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n c_j \varphi^{(j)} \in \mathcal{V}$  in the inner product

space can be written as an unique (infinite) sum of the eigenvectors which is exactly what a (Schauder) basis indicates. We put the justification of (b) in the footnote below<sup>14</sup> and (c) follows its counterpart in Theorem 10.3.4 but the group indices there are now replaced by individual indices. The resolution of the identity (d) is then derived similarly as in Theorem 10.3.4: express any  $\vec{v}$  as  $\vec{v} = c_1\varphi^{(1)} + c_2\varphi^{(2)} + \dots = \lim_{n \rightarrow \infty} \bigoplus_{j=1}^n c_j\varphi^{(j)}$  and by Definition 12.3.1, we have

$$\begin{aligned} T_j(\vec{v}) &= \langle \vec{v}, \varphi^{(j)} \rangle \varphi^{(j)} \\ &= \langle \dots + c_{j-1}\varphi^{(j-1)} + c_j\varphi^{(j)} + c_{j+1}\varphi^{(j+1)} + \dots, \varphi^{(j)} \rangle \varphi^{(j)} \\ &= (\dots(0) + c_j(1) + \dots(0))\varphi^{(j)} = c_j\varphi^{(j)} \end{aligned}$$

and thus

$$\begin{aligned} I(\vec{v}) &= \vec{v} = c_1\varphi^{(1)} + c_2\varphi^{(2)} + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n c_j\varphi^{(j)} \\ &= T_1(\vec{v}) + T_2(\vec{v}) + \dots = (T_1 + T_2 + \dots)(\vec{v}) = \left( \lim_{n \rightarrow \infty} \sum_{j=1}^n T_j \right)(\vec{v}) \end{aligned}$$

So  $I = \lim_{n \rightarrow \infty} \sum_{j=1}^n T_j$ . (e) is also shown in a similar fashion:

$$\begin{aligned} T(\vec{v}) &= T(c_1\varphi^{(1)} + c_2\varphi^{(2)} + \dots) = T\left(\lim_{n \rightarrow \infty} \sum_{j=1}^n c_j\varphi^{(j)}\right) \\ &= T(c_1\varphi^{(1)}) + T(c_2\varphi^{(2)}) + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n c_j T(\varphi^{(j)}) \quad (T \text{ is linear}) \\ &= \lambda_1 c_1 \varphi^{(1)} + \lambda_2 c_2 \varphi^{(2)} + \dots = \lim_{n \rightarrow \infty} \sum_{j=1}^n \lambda_j c_j \varphi^{(j)} \\ &= \lambda_1 T_1(\vec{v}) + \lambda_2 T_2(\vec{v}) + \dots \quad (\text{as in deriving (d)}) \end{aligned}$$

---

<sup>14</sup>For  $\vec{v} \in \bigoplus_{j \in \{J\}} \mathcal{E}_j$ , it will be in the form of  $\vec{v} = \sum_{j \in \{J\}} c_j \varphi^{(j)}$  which is clearly orthogonal to  $\vec{w} = \sum_{j' \notin \{J\}} c_{j'} \varphi^{(j')} \in \bigoplus_{j \notin \{J\}} \mathcal{E}_j$ , seen by expanding and computing  $\langle \vec{v}, \vec{w} \rangle = 0$ , where each  $\langle \varphi^{(j)}, \varphi^{(j')} \rangle$  term is zero when  $j \neq j'$ , as  $\varphi^{(j)}$  and  $\varphi^{(j')}$  are orthogonal to each other.

$$= (\lambda_1 T_1 + \lambda_2 T_2 + \cdots) \vec{v} = \left( \lim_{n \rightarrow \infty} \sum_{j=1}^n \lambda_j T_j \right) (\vec{v})$$

Be wary that in (e), the linear operator  $T$  applied on  $\vec{v}$  is required to be Hermitian, not only self-adjoint. This means that there will be problem if boundary terms appear when an integral-like inner product is used. This issue is going to be raised in the next example and short exercise.

Recall that in Example 12.2.3 we have derived the Fourier basis  $\{\sin(x), \sin(2x), \sin(3x), \dots, 1, \cos(x), \cos(2x), \cos(3x), \dots\}$  for the  $L^2[-\pi, \pi]$  Hilbert space with the inner product of Equation (12.1). To normalize it, note that each eigenfunction has a norm of  $\frac{1}{\sqrt{\pi}}$ <sup>15</sup> ( $\frac{1}{\sqrt{2\pi}}$  for 1) and hence the orthonormal Fourier basis is  $\{\frac{1}{\sqrt{\pi}} \sin(x), \frac{1}{\sqrt{\pi}} \sin(2x), \frac{1}{\sqrt{\pi}} \sin(3x), \dots, \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos(x), \frac{1}{\sqrt{\pi}} \cos(2x), \frac{1}{\sqrt{\pi}} \cos(3x), \dots\}$ . According to (a) of the Spectral Theorem 12.3.4, any function  $f \in L^2[-\pi, \pi]$  can thus be expanded in a so-called **Fourier series**, which is often expressed in the form of

$$\begin{aligned} f &= \frac{a_0}{2} + a_1 \cos(x) + a_2 \cos(2x) + a_3 \cos(3x) + \cdots \\ &\quad + b_1 \sin(x) + b_2 \sin(2x) + b_3 \sin(3x) + \cdots \\ &= \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx) \end{aligned} \tag{12.4}$$

---

<sup>15</sup>We only show this for the cosines but the calculation is the same for the sines. For integer  $m$ ,

$$\begin{aligned} \|\cos(mx)\|^2 &= \int_{-\pi}^{\pi} \cos(mx) \overline{\cos(mx)} dx \\ &= \int_{-\pi}^{\pi} \cos^2(mx) dx \\ &= \int_{-\pi}^{\pi} \frac{1}{2}(1 + \cos(2mx)) dx \\ &= \left[ \frac{1}{2}x + \frac{1}{4m} \sin(2mx) \right]_{-\pi}^{\pi} \\ &= \frac{1}{2}(2\pi) + (0) = \frac{1}{\pi} \end{aligned}$$

an (infinite) sum of these sinusoidal eigenfunctions (without the  $\frac{1}{\sqrt{\pi}}$  normalization factor). To compute the **Fourier coefficients**  $a_m$  and  $b_n$ , we use (d) of the Spectral Theorem where the  $T_j$  are in the form of Equation (12.3), leading to

$$\begin{aligned} f &= \sum_j \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \\ &= \left\langle f, \frac{1}{\sqrt{2\pi}} \right\rangle \left( \frac{1}{\sqrt{2\pi}} \right) \\ &\quad + \left\langle f, \frac{1}{\sqrt{\pi}} \cos(x) \right\rangle \left( \frac{1}{\sqrt{\pi}} \cos(x) \right) + \left\langle f, \frac{1}{\sqrt{\pi}} \cos(2x) \right\rangle \left( \frac{1}{\sqrt{\pi}} \cos(2x) \right) + \cdots + \\ &\quad + \left\langle f, \frac{1}{\sqrt{\pi}} \sin(x) \right\rangle \left( \frac{1}{\sqrt{\pi}} \sin(x) \right) + \left\langle f, \frac{1}{\sqrt{\pi}} \sin(2x) \right\rangle \left( \frac{1}{\sqrt{\pi}} \sin(2x) \right) + \cdots \\ &= \frac{1}{2\pi} \langle f, 1 \rangle + \frac{1}{\pi} \langle f, \cos(x) \rangle \cos(x) + \frac{1}{\pi} \langle f, \cos(2x) \rangle \cos(2x) + \cdots \\ &\quad + \frac{1}{\pi} \langle f, \sin(x) \rangle \sin(x) + \frac{1}{\pi} \langle f, \sin(2x) \rangle \sin(2x) + \cdots \end{aligned}$$

Comparing this expression with the form of Fourier series in Equation (12.4), we yield the following formulae for the coefficients.

**Properties 12.3.5** (Fourier Series). For a function  $f(x)$  in the  $L^2[-\pi, \pi]$  space, it can be written as a Fourier series of

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx)$$

where the Fourier coefficients are given by

$$a_m = \frac{1}{\pi} \langle f, \cos(mx) \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(mx) dx \quad (12.5)$$

$$b_n = \frac{1}{\pi} \langle f, \sin(nx) \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx \quad (12.6)$$

**Example 12.3.2.** Expand the function  $x$  as a Fourier series in the interval  $[-\pi, \pi]$ .

*Solution.* It is simply to compute  $a_m$  and  $b_n$  mechanically according to Equations (12.5) and (12.6). But note that since  $x$  is an odd function and cosines are even, the integrals for  $a_m$  where the interval is symmetric about the origin and the resulting integrands are odd, are thus all zero. Now we just have to calculate the general form of  $b_n$ , as below.

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} x \sin(nx) dx \\ &= \frac{1}{\pi} \left[ -\frac{1}{n} x \cos(nx) \right]_{-\pi}^{\pi} - \frac{1}{\pi} \int_{-\pi}^{\pi} \left( -\frac{1}{n} \cos(nx) \right) dx \\ &= \frac{1}{\pi} \left[ -\frac{1}{n} \pi \cos(n\pi) + \frac{1}{n} (-\pi) \cos(n(-\pi)) \right] + \frac{1}{n\pi} \int_{-\pi}^{\pi} \cos(nx) dx \\ &= -\frac{1}{n} \cos(n\pi) - \frac{1}{n} \cos(-n\pi) + \frac{1}{n^2\pi} [\sin(nx)]_{-\pi}^{\pi} \\ &= -\frac{2(-1)^n}{n} + (0) = -\frac{2(-1)^n}{n} \end{aligned}$$

as  $\cos(n\pi) = \cos(-n\pi) = (-1)^n$ . Therefore the Fourier series of  $x$  is

$$x = -2 \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin(nx)$$

Short Exercise: It seems that if we apply part (e) of the Spectral Theorem 12.3.4 where  $T = \frac{d^2}{dx^2}$  as in Example 12.2.3 on the Fourier series of  $x$  above, the L.H.S. will readily become zero but the R.H.S. will still contain non-trivial sine terms under the twice differentiation so they are obviously unequal. Why does this paradox occur?<sup>16</sup> □

---

<sup>16</sup>It is because the Hermicity of  $T$ , required by (e) of the Spectral Theorem, is not satisfied for the function  $x$ . The boundary terms derived in Example 12.2.3

$$[f(x) \frac{d}{dx} \overline{([g(x)])}]_{-\pi}^{\pi} - [\frac{d}{dx} (f(x)) \overline{g(x)}]_{-\pi}^{\pi}$$

## 12.4 Special Polynomials

### 12.4.1 Sturm-Liouville Equations

The theory of Hermitian operators is most widely applied in the context of ***Sturm-Liouville Equations*** that are frequently encountered when solving ***Partial Differential Equations (PDEs)*** with the technique of *separation of variables*. A Sturm-Liouville equation takes the general form of

$$\frac{d}{dx} \left( p(x) \frac{dy}{dx} \right) + q(x)y + \lambda w(x)y = 0 \quad (12.7)$$

where  $p(x)$ ,  $q(x)$  are real functions of  $x$  and  $w(x)$  is a real positive-definite weighting function by  $x$ . It can be written in terms of the ***Sturm-Liouville operator***  $\mathcal{L}[f] = -(\frac{d}{dx}(p(x)\frac{d}{dx}) + q(x))[f]$  as

$$\mathcal{L}[y] = \lambda w(x)y$$

which apparently poses an eigenvalue-eigenfunction problem with a weighting  $w(x)$  where  $\lambda$  is the eigenvalue. The Sturm-Liouville operator is self-adjoint with respect to the inner product (12.1) and can become Hermitian under suitable boundary conditions as shown below:

$$\begin{aligned} \langle f, g \rangle &= \int_a^b f(x) \overline{\mathcal{L}[g(x)]} dx \\ &= \int_a^b f(x) \overline{\left( -\left( \frac{d}{dx}(p(x)\frac{dg}{dx}) + q(x)g(x) \right) \right)} dx \end{aligned}$$

---

will vanish only if  $f(x)$  and  $g(x)$  and their derivatives  $f'(x)$  and  $g'(x)$  are periodic, i.e. equal at the two end-points  $-\pi, \pi$ . Clearly  $x$  is not a periodic function so it does not satisfy the boundary conditions and this part of the Spectral Theorem is not applicable. For reference, in general, a Fourier series can be differentiated term by term if the function  $f(x)$  is continuous, takes the same value at the two end points  $f(-\pi) = f(\pi)$  and its derivative  $f'(x)$  is piecewise continuous. A Fourier series essentially extends the given function by repeating it with a period of  $2\pi$ , so actually such a periodic extension of the function  $x$  will be discontinuous at the boundaries between the repeated graphs of  $x$ .

$$\begin{aligned}
 &= - \int_a^b f(x) \frac{d}{dx} \left( p(x) \frac{d\overline{g(x)}}{dx} \right) + \int_a^b f(x) q(x) \overline{g(x)} dx \\
 &= - [f(x) p(x) \frac{d\overline{g(x)}}{dx}]_a^b + \int_a^b \frac{df(x)}{dx} \left( p(x) \frac{d\overline{g(x)}}{dx} \right) dx \\
 &\quad + \int_a^b f(x) q(x) \overline{g(x)} dx \\
 &= - [f(x) p(x) \frac{d\overline{g(x)}}{dx}]_a^b + [\frac{df(x)}{dx} p(x) \overline{g(x)}]_a^b \\
 &\quad - \int \frac{d}{dx} \left( p(x) \frac{df(x)}{dx} \right) g(x) + \frac{d\overline{g(x)}}{dx} dx + \int_a^b f(x) q(x) \overline{g(x)} dx \\
 &= - [f(x) p(x) \frac{d\overline{g(x)}}{dx}]_a^b + [\frac{df(x)}{dx} p(x) \overline{g(x)}]_a^b \\
 &\quad + \int - \left( \frac{d}{dx} \left( p(x) \frac{df(x)}{dx} \right) + q(x) f(x) \right) \overline{g(x)} dx \\
 &= - [f(x) p(x) \frac{d\overline{g(x)}}{dx}]_a^b + [\frac{df(x)}{dx} p(x) \overline{g(x)}]_a^b + \int_a^b \mathcal{L}[f(x)] \overline{g(x)} dx
 \end{aligned}$$

So  $\mathcal{L}^* = \mathcal{L}$  is self-adjoint, and the boundary condition for the boundary terms to vanish is  $[f(x) p(x) \frac{d\overline{g(x)}}{dx}]_a^b = 0$  for any pair of  $f(x)$  and  $g(x)$ <sup>17</sup> so that  $\mathcal{L}$  can be Hermitian. This boundary condition will also have to be obeyed by the eigenfunctions  $\varphi_j$  where  $\mathcal{L}[\varphi_j] = \lambda w(x) \varphi_j$ . However, sometimes  $p(x)$  may take a form such that at some end-points its value may be zero so that  $\mathcal{L}$  is automatically Hermitian over this *natural interval*. The reality of eigenvalues (again, left as a short exercise) and the orthogonality of eigenfunctions are enabled by the Hermicity of the Sturm-Liouville operator as in Properties 12.2.6 but with an extra factor from the weighting  $w(x)$ :

$$\int_a^b \varphi_1 \overline{\mathcal{L}[\varphi_2]} = \int_a^b \mathcal{L}[\varphi_1] \overline{\varphi_2}$$

---

<sup>17</sup>This implies that  $[g(x) p(x) \frac{d\overline{f(x)}}{dx}]_a^b = 0$  as well when the roles of  $f(x)$  and  $g(x)$  are interchanged. Conjugating this relation gives  $[g(x) p(x) \frac{d\overline{f(x)}}{dx}]_a^b = [\frac{df(x)}{dx} p(x) \overline{g(x)}]_a^b = 0$  so the second boundary term also vanishes.

$$\begin{aligned}\int_a^b \varphi_1 \overline{\lambda_2 w(x) \varphi_2} dx &= \int_a^b \lambda_1 w(x) \varphi_1 \overline{\varphi_2} dx \\ \int_a^b \lambda_2 w(x) \varphi_1 \overline{\varphi_2} dx &= \int_a^b \lambda_1 w(x) \varphi_1 \overline{\varphi_2} dx \\ (\lambda_1 - \lambda_2) \int_a^b w(x) \varphi_1 \overline{\varphi_2} dx &= 0\end{aligned}$$

so that  $\int_a^b w(x) \varphi_1 \overline{\varphi_2} dx = 0$  and the two different eigenfunctions are orthogonal with respect to the inner product (12.2) assumed that  $\lambda_1 \neq \lambda_2$ . To transform a general second-order ODE  $P(x) \frac{d^2y}{dx^2} + R(x) \frac{dy}{dx} + Q(x)y + \lambda w(x)y = 0$  into the Sturm-Liouville form can be done by multiplying the integrating factor  $F(x) = \exp\left(\int \frac{R(x)-P'(x)}{P(x)} dx\right)$ , so that it becomes

$$\begin{aligned}[F(x)P(x)y']' + F(x)Q(x)y + \lambda F(x)w(x)y \\ = [p(x)y']' + q(x)y + \lambda F(x)w(x)y = 0\end{aligned}$$

<sup>18</sup> where  $p(x) = F(x)P(x)$ ,  $q(x) = F(x)Q(x)$ , and the new weighting is  $F(x)w(x)$  which is still positive-definite as  $F(x)$  is an exponential function.

**Example 12.4.1.** Convert the Hermite's Equation

$$y'' - 2xy' + 2\gamma y = 0$$

<sup>18</sup>  $F(x)$  is derived such that  $F(x)P(x)y'' + F(x)R(x)y' = [F(x)P(x)y']'$ :

$$\begin{aligned}[F(x)P(x)y']' &= F(x)P(x)y'' + F(x)R(x)y' = F(x)P(x)y'' + F(x)P'(x)y' + F'(x)P(x)y' \\ F(x)R(x)y' &= F(x)P'(x)y' + F'(x)P(x)y' \\ F(x)(R(x) - P'(x)) &= F'(x)P(x) \\ \frac{dF(x)}{F(x)} &= \frac{R(x) - P'(x)}{P(x)} \\ \ln F(x) &= \int \frac{R(x) - P'(x)}{P(x)} dx \\ \Rightarrow F(x) &= \exp\left(\int \frac{R(x) - P'(x)}{P(x)} dx\right)\end{aligned}$$

to the Sturm-Liouville form and find its eigenfunctions.

*Solution.* The integrating factor is

$$\begin{aligned} F(x) &= \exp\left(\int \frac{(-2x) - (0)}{(1)} dx\right) \\ &= \exp\left(\int -2x dx\right) = e^{-x^2} \end{aligned}$$

and hence by multiplying it to the Hermite's Equation

$$\begin{aligned} e^{-x^2} y'' - 2xe^{-x^2} y' + 2\lambda e^{-x^2} y &= 0 \\ (e^{-x^2} y')' + 2\lambda e^{-x^2} y &= 0 \end{aligned}$$

yields the Sturm-Liouville form with  $p(x) = e^{-x^2}$ ,  $q(x) = 0$ ,  $\lambda = 2\nu$ ,  $w(x) = e^{-x^2}$ . We can find the eigenfunctions by the method of *series solution*. Assume a series solution of  $y = a_0 + a_1x + a_2x^2 + \dots = \sum_{n=0}^{\infty} a_n x^n$ , then

$$\begin{aligned} y' &= \sum_{n=1}^{\infty} n a_n x^{n-1} \\ y'' &= \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} \end{aligned}$$

Substituting them into the original form of Hermite's equation, we have

$$\begin{aligned} \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} - 2x \sum_{n=1}^{\infty} n a_n x^{n-1} + 2\nu \sum_{n=0}^{\infty} a_n x^n &= 0 \\ \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n - 2 \sum_{n=1}^{\infty} n a_n x^n + 2\nu \sum_{n=0}^{\infty} a_n x^n &= 0 \end{aligned}$$

which gives a recurrence relation of

$$(n+2)(n+1)a_{n+2} - 2na_n + 2\nu a_n = 0$$

for  $n \geq 1$ . For  $n = 0$  it is simply  $2a_2 + 2\nu a_0 = 0$ . Rearranging we have

$$a_{n+2} = \frac{2(n-\nu)}{(n+2)(n+1)} a_n$$

the form of which indicates that there will be two series, one for the odd indices and another for the even. Note that for the corresponding Sturm-Liouville operator to become Hermitian, the boundary term

$$[\varphi_1 p(x) \frac{d\varphi_2}{dx}]_a^b = [e^{-x^2} \varphi_1 \frac{d\varphi_2}{dx}]_a^b = 0$$

has to vanish, and the  $e^{-x^2}$  factor means that the natural interval would be  $(-\infty, \infty)$  along the entire real axis. The eigenfunctions must therefore grow slower than  $e^{x^2}$ , which happens when one of the series solution is truncated to a polynomial when  $\nu = n$ . These polynomials are subsequently known as the *Hermite's polynomial*. Here we compute the first four of them ( $n = 0, 1, 2, 3$ ). For  $n = 0$ , we simply have  $y = a_0$  as  $2a_2 + 2(0)a_0 = 0$  means  $a_2 = 0$  and the series is immediately terminated. For  $n = 1$ , we similarly have  $y = a_1 x$  as  $a_3 = \frac{2(1-1)}{(3)(2)} a_1 = (0)a_1 = 0$ . Going up to  $n = 2$ , we have

$$\begin{aligned} a_2 &= \frac{2(0-2)}{(2)(1)} a_0 = -2a_0 \\ a_4 &= \frac{2(2-2)}{(4)(3)} a_2 = 0 \end{aligned}$$

So  $y = a_0 - 2a_0 x^2 = a_0(1 - 2x^2)$ . In the same fashion, when  $n = 3$ , we have

$$\begin{aligned} a_3 &= \frac{2(1-3)}{(3)(2)} a_1 = -\frac{2}{3} a_1 \\ a_5 &= \frac{2(3-3)}{(5)(4)} a_3 = 0 \end{aligned}$$

hence  $y = a_1 x - \frac{2}{3} a_1 x^3 = a_1(x - \frac{2}{3}x^3)$ . By convention, the Hermite's polynomial is scaled in the way such that the leading highest degree term has a coefficient of  $2^n$ . Hence the first four of Hermite's polynomial are

$$H_0(x) = 1$$

$$\begin{aligned}H_1(x) &= 2x \\H_2(x) &= 4x^2 - 2 \\H_3(x) &= 8x^3 - 12x\end{aligned}$$

Short Exercise: Find the Hermite's polynomial of degree 4,  $H_4(x)$ , corresponding to  $\nu = n = 4$ .<sup>19</sup>  $\square$

### 12.4.2 Generating Special Polynomials by Gram-Schmidt Orthogonalization

Since a Sturm-Liouville operator is Hermitian given appropriate boundary conditions and Hermicity leads to a complete orthonormal basis that is countable if the underlying Hilbert space is separable, which we assume to be that so, we can derive the corresponding orthogonal **special polynomials** to the Sturm-Liouville equation, e.g. Hermite's polynomials in the previous part, by applying the Gram-Schmidt process on the standard polynomial basis  $\{1, x, x^2, x^3, \dots\}$  to any degree with respect to the inner product in Equation (12.2). The details of calculation is shown below.

**Example 12.4.2.** Compute the first four Hermite's polynomials as introduced in Example 12.4.1 by Gram-Schmidt Orthogonalization.

*Solution.* The Hermite's polynomials has a weighting of  $w(x) = e^{-x^2}$  and are integrated over the entire real axis  $(-\infty, \infty)$ . Hence it is instructive to first note

---

<sup>19</sup>Using the recurrence relation, we have

$$\begin{aligned}a_2 &= \frac{2(0-4)}{(2)(1)}a_0 = -4a_0 \\a_4 &= \frac{2(2-4)}{(4)(3)}a_2 = -\frac{1}{3}a_2 = \frac{4}{3}a_0 \\a_6 &= \frac{2(4-4)}{(6)(5)}a_4 = 0\end{aligned}$$

hence  $y = a_0 - 4a_0x^2 + \frac{4}{3}a_0x^4$  and after scaling it becomes  $H_4(x) = 16x^4 - 48x^2 + 12$ .

that for any non-negative integer  $m$

$$\int_{-\infty}^{\infty} x^{2m+1} e^{-x^2} dx = 0$$

since  $e^{-x^2}$  is even and  $x^{2m+1}$  is odd, and

$$\int_{-\infty}^{\infty} x^{2m} e^{-x^2} dx = \frac{(1)(3)(5) \cdots (2m-1)\sqrt{\pi}}{2^m}$$

(see footnote below)<sup>20</sup> Now we apply Gram-Schmidt Orthogonalization on the standard polynomials  $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \vec{u}^{(4)}, \dots\} = \{1, x, x^2, x^3, \dots\}$  according to Definition 12.3.3. The zeroth degree Hermite's polynomial is trivially  $\vec{v}^{(1)} = H_0(x) = 1$  and  $\|\vec{v}^{(1)}\|^2 = \sqrt{\pi}$ . Next, for  $n = 1$ , we have

$$\vec{v}^{(2)} = \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)}$$

---

<sup>20</sup>It is a well-known result from multivariable calculus that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

for the Gaussian integral. By mathematical induction and integration by parts, we have

$$\begin{aligned} \int_{-\infty}^{\infty} x^{2(m+1)} e^{-x^2} dx &= \int_{-\infty}^{\infty} x^{2m+2} e^{-x^2} dx \\ &= \int_{-\infty}^{\infty} -\frac{1}{2} x^{2m+1} (-2x e^{-x^2}) dx \\ &= \int_{-\infty}^{\infty} -\frac{1}{2} x^{2m+1} d(e^{-x^2}) \\ &= [-\frac{1}{2} x^{2m+1} e^{-x^2}]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-\frac{2m+1}{2} x^{2m}) e^{-x^2} dx \\ &= (0) + \frac{2m+1}{2} \int_{-\infty}^{\infty} x^{2m} e^{-x^2} dx \\ &= \frac{(1)(3)(5) \cdots (2m-1)(2m+1)\sqrt{\pi}}{2^{m+1}} \quad (\text{via the induction hypothesis}) \end{aligned}$$

so the formula is established.

$$\begin{aligned}
 &= x - \frac{\int_{-\infty}^{\infty} e^{-x^2}(x)(1)dx}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} \\
 &= x - \frac{\int_{-\infty}^{\infty} xe^{-x^2} dx}{\sqrt{\pi}} (1) = x - (0)(1) = x
 \end{aligned}$$

So the Hermite's polynomial of degree 1 is just  $\vec{v}^{(2)} = H_1(x) = 2x$  (scaled by convention, similar for the followings). Now

$$\begin{aligned}
 \|\vec{v}^{(2)}\|^2 &= \int_{-\infty}^{\infty} (2x)^2 e^{-x^2} dx \\
 &= 4 \int_{-\infty}^{\infty} x^2 e^{-x^2} dx \\
 &= 4\left(\frac{1}{2}\sqrt{\pi}\right) = 2\sqrt{\pi}
 \end{aligned}$$

For  $n = 2$ , we have

$$\begin{aligned}
 \vec{v}^{(3)} &= \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)} \\
 &= x^2 - \frac{\int_{-\infty}^{\infty} (x^2)(1)e^{-x^2} dx}{\sqrt{\pi}} (1) - \frac{\int_{-\infty}^{\infty} (x^2)(2x)e^{-x^2} dx}{2\sqrt{\pi}} (2x) \\
 &= x^2 - \frac{\int_{-\infty}^{\infty} x^2 e^{-x^2} dx}{\sqrt{\pi}} (1) - \frac{2 \int_{-\infty}^{\infty} x^3 e^{-x^2} dx}{2\sqrt{\pi}} (2x) \\
 &= x^2 - \frac{\left(\frac{1}{2}\sqrt{\pi}\right)}{\sqrt{\pi}} (1) - (0)(2x) \\
 &= x^2 - \frac{1}{2}
 \end{aligned}$$

Hence  $\vec{v}^{(3)} = H_2(x) = 4x^2 - 2$ , and  $\|\vec{v}^{(3)}\|^2 = \int_{-\infty}^{\infty} (4x^2 - 2)^2 e^{-x^2} dx = \int_{-\infty}^{\infty} (16x^4 - 16x^2 + 4)e^{-x^2} dx = 16\left(\frac{3}{4}\sqrt{\pi}\right) - 16\left(\frac{1}{2}\sqrt{\pi}\right) + 4(\sqrt{\pi}) = 8\sqrt{\pi}$ . Finally, for  $n = 3$

$$\vec{v}^{(4)} = \vec{u}^{(4)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(1)} \rangle}{\|\vec{v}^{(1)}\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(2)} \rangle}{\|\vec{v}^{(2)}\|^2} \vec{v}^{(2)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(3)} \rangle}{\|\vec{v}^{(3)}\|^2} \vec{v}^{(3)}$$

$$\begin{aligned}
 &= x^3 - \frac{\int_{-\infty}^{\infty} (x^3)(1)e^{-x^2} dx}{\sqrt{\pi}}(1) - \frac{\int_{-\infty}^{\infty} (x^3)(2x)e^{-x^2} dx}{2\sqrt{\pi}}(2x) \\
 &\quad - \frac{\int_{-\infty}^{\infty} (x^3)(4x^2 - 2)e^{-x^2} dx}{8\sqrt{\pi}}(4x^2 - 2) \\
 &= x^3 - (0)(1) - \frac{2 \int_{-\infty}^{\infty} x^4 e^{-x^2} dx}{2\sqrt{\pi}}(2x) - (0)(4x^2 - 2) \\
 &= x^3 - \left(\frac{3\sqrt{\pi}}{4\sqrt{\pi}}\right)(2x) = x^3 - \frac{3}{2}x
 \end{aligned}$$

thus  $\vec{v}^{(4)} = H_3(x) = 8x^3 - 12x$ . □

## 12.5 Earth Science Applications

**Example 12.5.1.** Consider the *linearized shallow water system* along the Equator:

$$\begin{aligned}
 \frac{\partial u}{\partial t} - \beta y v &= -\frac{\partial \Phi}{\partial x} \\
 \frac{\partial v}{\partial t} + \beta y u &= -\frac{\partial \Phi}{\partial y} \\
 \frac{\partial \Phi}{\partial t} &= -gH\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)
 \end{aligned}$$

The first two equations are the momentum equations and the last one is the continuity equation.  $u, v, \Phi$  are the zonal, meridional winds and the geopotential. The constants include  $\beta$  which denotes the beta effect (the change of  $f$  over latitudes), and  $g, H$  as the gravitational acceleration and equivalent depth. By assuming a travelling wave solution (written in the complex form,  $i$  is the imaginary number)

$$\begin{aligned}
 u &= \hat{u}(y) \exp(i(kx - \omega t)) \\
 v &= \hat{v}(y) \exp(i(kx - \omega t))
 \end{aligned}$$

$$\Phi = \hat{\Phi}(y) \exp(i(kx - \omega t))$$

so that  $\hat{u}(y)$ ,  $\hat{v}(y)$ ,  $\hat{\Phi}(y)$  are now functions of  $y$  only. Substitute them into the three shallow water equations above, simplify them and apply some suitable changes of variable to arrive at the Hermite's equation and hence derive the modes of the linearized shallow water system.

*Solution.* Plugging the ansatz in, the system of equations become

$$\begin{aligned} -i\omega\hat{u}(y)e^{i(kx-\omega t)} - \beta y\hat{v}(y)e^{i(kx-\omega t)} &= -ik\hat{\Phi}(y)e^{i(kx-\omega t)} \\ -i\omega\hat{u}(y) - \beta y\hat{v}(y) &= -ik\hat{\Phi}(y) \end{aligned} \quad (12.8)$$

$$\begin{aligned} -i\omega\hat{v}(y)e^{i(kx-\omega t)} + \beta y\hat{u}(y)e^{i(kx-\omega t)} &= -\frac{d\hat{\Phi}(y)}{dy}e^{i(kx-\omega t)} \\ -i\omega\hat{v}(y) + \beta y\hat{u}(y) &= -\frac{d\hat{\Phi}(y)}{dy} \end{aligned} \quad (12.9)$$

and

$$\begin{aligned} -i\omega\hat{\Phi}(y)e^{i(kx-\omega t)} &= -gH \left( ik\hat{u}(y)e^{i(kx-\omega t)} + \frac{d\hat{v}(y)}{dy}e^{i(kx-\omega t)} \right) \\ -i\omega\hat{\Phi}(y) &= -gH \left( ik\hat{u}(y) + \frac{d\hat{v}(y)}{dy} \right) \end{aligned} \quad (12.10)$$

Differentiating (12.10) gives

$$-i\omega \frac{d\hat{\Phi}(y)}{dy} = -gH \left( ik \frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2} \right) \quad (12.11)$$

Substituting (12.9) into (12.11) leads to

$$\begin{aligned} i\omega(-i\omega\hat{v}(y) + \beta y\hat{u}(y)) &= -gH \left( ik \frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2} \right) \\ \omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y) &= -gH \left( ik \frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2} \right) \end{aligned} \quad (12.12)$$

Using (12.8) in (12.12) gives

$$\begin{aligned}\omega^2 \hat{v}(y) + i\omega\beta y \hat{u}(y) &= -gH\left(-\frac{k}{\omega} \frac{d}{dy}(-ik\hat{\Phi}(y) + \beta y \hat{v}(y)) + \frac{d^2 \hat{v}(y)}{dy^2}\right) \\ \omega^2 \hat{v}(y) + i\omega\beta y \hat{u}(y) &= \frac{-igHk^2}{\omega} \frac{d\hat{\Phi}(y)}{dy} + \frac{gHk}{\omega} \left(\beta \hat{v}(y) + \beta y \frac{d\hat{v}}{dy}\right) \\ &\quad - gH \frac{d^2 \hat{v}(y)}{dy^2}\end{aligned}\tag{12.13}$$

Applying both (12.9) and (12.10) in (12.13) yields

$$\begin{aligned}\omega^2 \hat{v}(y) + i\omega\beta y \hat{u}(y) &= \frac{igHk^2}{\omega} (-i\omega \hat{v}(y) + \beta y \hat{u}(y)) + \frac{gHk}{\omega} \left(\beta \hat{v}(y) + \beta y \left(\frac{i\omega}{gH} \hat{\Phi}(y) - ik \hat{u}(y)\right)\right) \\ &\quad - gH \frac{d^2 \hat{v}(y)}{dy^2}\end{aligned}$$

Rearranging then gives

$$\begin{aligned}\omega^2 \hat{v}(y) + i\omega\beta y \hat{u}(y) &= gHk^2 \hat{v}(y) + \frac{igHk^2}{\omega} \beta y \hat{u}(y) + \frac{gHk}{\omega} \beta \hat{v}(y) + ik\beta y \hat{\Phi}(y) \\ &\quad - \frac{igHk^2}{\omega} \beta y \hat{u}(y) - gH \frac{d^2 \hat{v}(y)}{dy^2} \\ &= gHk^2 \hat{v}(y) + \frac{gHk}{\omega} \beta \hat{v}(y) + ik\beta y \hat{\Phi}(y) - gH \frac{d^2 \hat{v}(y)}{dy^2}\end{aligned}$$

Finally, using (12.8) again, we have

$$\omega^2 \hat{v}(y) - \beta^2 y^2 \hat{v}(y) = gHk^2 \hat{v}(y) + \frac{gHk}{\omega} \beta \hat{v}(y) - gH \frac{d^2 \hat{v}(y)}{dy^2}$$

Cleaning this up, we obtain

$$\frac{d^2 \hat{v}(y)}{dy^2} + \left[\left(\frac{\omega^2}{gH} - k^2 - \beta \frac{k}{\omega}\right) - \frac{\beta^2}{gH} y^2\right] \hat{v}(y) = 0$$

Finally, with a change of variable  $\tilde{y} = y/y_0$ ,  $y_0 = (\sqrt{gH}/\beta)^{1/2}$ , it becomes

$$\frac{d^2\hat{v}(\tilde{y})}{d\tilde{y}^2} + \left[ \frac{\sqrt{gH}}{\beta} \left( \frac{\omega^2}{gH} - k^2 - \beta \frac{k}{\omega} \right) - \tilde{y}^2 \right] \hat{v}(\tilde{y}) = 0$$

Letting

$$\mu = \frac{\sqrt{gH}}{\beta} \left( \frac{\omega^2}{gH} - k^2 - \beta \frac{k}{\omega} \right)$$

simplifies the equation to  $\hat{v}'' + (\mu - \tilde{y}^2)\hat{v} = 0$ . Further, if

$$\hat{v} = \hat{w} \exp\left(-\frac{\tilde{y}^2}{2}\right)$$

then

$$\hat{v}' = \hat{w}' \exp\left(-\frac{\tilde{y}^2}{2}\right) - \tilde{y}\hat{w} \exp\left(-\frac{\tilde{y}^2}{2}\right)$$

and

$$\begin{aligned} \hat{v}'' &= \hat{w}'' \exp\left(-\frac{\tilde{y}^2}{2}\right) - \tilde{y}\hat{w}' \exp\left(-\frac{\tilde{y}^2}{2}\right) - \hat{w} \exp\left(-\frac{\tilde{y}^2}{2}\right) \\ &\quad - \tilde{y}\hat{w}' \exp\left(-\frac{\tilde{y}^2}{2}\right) + \tilde{y}^2\hat{w} \exp\left(-\frac{\tilde{y}^2}{2}\right) \\ &= (\hat{w}'' - 2\tilde{y}\hat{w}' + (\tilde{y}^2 - 1)\hat{w}) \exp\left(-\frac{\tilde{y}^2}{2}\right) \end{aligned}$$

So

$$\hat{v}'' + (\mu - \tilde{y}^2)\hat{v} = 0$$

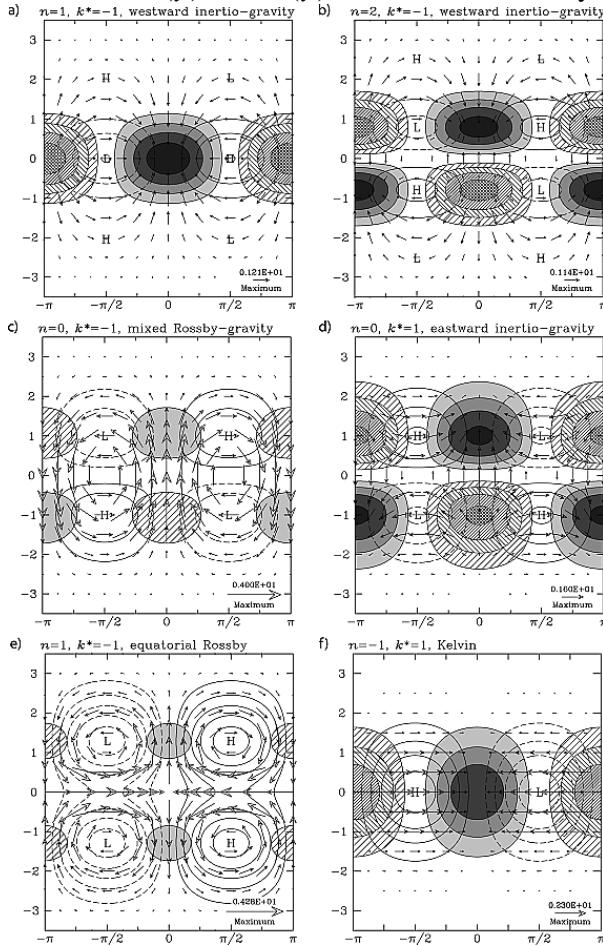
becomes

$$\begin{aligned} (\hat{w}'' - 2\tilde{y}\hat{w}' + (\tilde{y}^2 - 1)\hat{w}) \exp\left(-\frac{\tilde{y}^2}{2}\right) + (\mu - \tilde{y}^2)\hat{w} \exp\left(-\frac{\tilde{y}^2}{2}\right) &= 0 \\ \hat{w}'' - 2\tilde{y}\hat{w}' + (\mu - 1)\hat{w} &= 0 \end{aligned}$$

This is the Hermite's equation in disguise where we have  $\mu = 2\nu + 1$ ,  $\nu = 0, 1, 2, \dots$  if the solutions have to be bounded. Hence the equatorial wave modes are given in terms of

$$\hat{v}(\tilde{y}) = v_0 H_n(\tilde{y}) \exp\left(-\frac{\tilde{y}^2}{2}\right)$$

From this,  $\hat{u}(\tilde{y})$  and  $\hat{\Phi}(\tilde{y})$  can be found via any two of (12.8), (12.9) and (12.10).



Horizontal structures of selected zonally propagating wave solutions to the shallow water equations on an equatorial  $\beta$  plane. (Kiladis et al., 2009)  $\square$

## 12.6 Python Programming

Here we demonstrate how to compute a real inner product supplied by a symmetric bilinear form  $B$  and carry out Gram-Schmidt Orthogonalization with respect to it. We use Example 12.3.1 as the test case. We first define a function to calculate the inner product given two input vectors and the matrix  $B$ :

```
import numpy as np

def real_inner_prod(u, v, B):
    if np.all(B == B.T): # Check if symmetric
        return(u @ B @ v)
    else:
        print("Not symmetric bilinear form!")
        return(None)
```

Let's check it with that  $B$  in Example 12.3.1 and  $\vec{u} = (1, 0, 1)^T$ ,  $\vec{v} = (0, 2, -1)^T$ . Then

```
u = np.array([1., 0., 1.])
v = np.array([0., 2., -1.])
B = np.array([[2., 1., 0.],
              [1., 2., 1.],
              [0., 1., 2.]])  
  
print(real_inner_prod(u,v,B))
```

gives 2.0 which turns out to be correct. For convenience we also define a function to compute norm, which is simply a wrapped version of `real_inner_prod`:

```
def norm(v, B):
    return(np.sqrt(real_inner_prod(v, v, B)))
```

Now come the main part of executing the Gram-Schmidt procedure. The inner loop subtract the parallel components of each previous vector from the current vector and the outer loop iterates the calculation for every vector.

```
def GS_inner_prod(vecs, B):
    """
    Gram-Schmidt Orthogonalization with respect to an inner
    product (finite-dimensional)
    vecs: A list containing the vectors
```

```
B: The symmetric matrix for the inner product
"""
n_vecs = len(vecs)
for jj in np.arange(n_vecs):
    for ii in np.arange(jj):
        vecs[jj] -= real_inner_prod(vecs[jj], vecs[ii], B)
            / norm(vecs[ii], B)**2 * vecs[ii]
return(vecs)
```

Trying this with Example 12.3.1

```
vecs = [np.array([1., 0., 0.]),
        np.array([0., 1., 0.]),
        np.array([0., 0., 1.])]
print(GS_inner_prod(vecs, B))
```

produces

```
[array([1., 0., 0.]), array([-0.5, 1., 0.]), array([
0.33333333, -0.66666667, 1.])]
```

which matches our answer in the example.

## 12.7 Exercises

**Exercise 12.1** Show that the set of all  $n \times n$  (complex) matrices  $\mathcal{V} = \mathcal{M}_{n \times n}(\mathbb{C})$  is a vector space and the definition  $\langle A, B \rangle = \text{tr}(AB^*)$  satisfies the requirements of an inner product for this vector space. This form of inner product is better known as the *Frobenius inner product*.

**Exercise 12.2** Show that  $\vec{u} = (-1, 0, 2)^T$  and  $\vec{v} = (1, -1, 1)^T$  in  $\mathbb{R}^3$  are orthogonal to each other if an inner product of

$$\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$$

where

$$B = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

is used. Also, find the norm  $\|\vec{u}\|$  and  $\|\vec{v}\|$  of both  $\vec{u}$  and  $\vec{v}$  with respect to this inner product.

**Exercise 12.3** Let  $\mathcal{V} = \mathbb{R}^3$ . Show that

$$\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$$

where

$$B = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

is a valid inner product for all  $\vec{u}, \vec{v} \in \mathcal{V}$  and turns  $\mathcal{V}$  into an inner product space. Hence derive an orthonormal basis for  $\mathbb{R}^3$  with respect to this inner product using Gram-Schmidt Orthogonalization.

**Exercise 12.4** Prove the Triangular Inequality

$$\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$$

now for any inner product.

**Exercise 12.5** Show that  $x^3$  and  $\cos(2x)$  is orthogonal to each other in the  $L^2[-\pi, \pi]$  space with respect to the inner product of Equation (12.1).

**Exercise 12.6** Find the adjoint of  $\mathcal{L}[f] = \frac{d}{dx}(x \frac{d}{dx}[f])$  with respect to the inner product of Equation (12.2)

$$\langle f, g \rangle = \int_a^b w(x) f(x) \overline{g(x)} dx$$

where  $w(x) =$  (a) 1, and (b)  $x$ , with  $a > 0$ .

**Exercise 12.7** Show that given a real inner product induced by  $\langle \vec{u}, \vec{v} \rangle = \vec{u}^T B \vec{v}$  where

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

The linear operator

$$T = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

is self-adjoint.

**Exercise 12.8** Find the Fourier series of (a)  $x^2$ , (b)  $x^3$  and (c)  $4 + \sin(3x) + 2\cos(8x)$  over  $[-\pi, \pi]$ .

**Exercise 12.9** Convert the *Legendre's Equation*

$$(1 - x^2)y'' - 2xy' + \lambda y = 0$$

into the Sturm-Liouville form, what should be the natural interval such that the corresponding Sturm-Liouville operator is Hermitian? Then, show that when  $\lambda = n(n + 1)$ ,  $n$  is a non-negative integer, the series solution truncates and produces the *Legendre polynomials* as its eigenfunctions. Find the first five of them. Alternatively, apply the Gram-Schmidt procedure over the standard polynomial basis over the natural interval to come up with the same set of Legendre polynomials.

## Chapter 13

# Least-Square Approximation

---

As discussed in Chapter 3, given a linear system, if the number of equations (rows) is greater than the number of unknowns (columns), then it is overdetermined. Generally, it would be inconsistent and there will be no solution. However, we can salvage this by finding an approximated solution such that the so-called squared error is minimized. This is known as the *least-square approximation*. Its most common application is *linear regression* in which we predict a dependent variable using an optimal linear equation of some independent variable(s).

### 13.1 Mathematical Ideas of Least-Square Approximation

For a linear system  $A\vec{x} = \vec{h}$  where  $\vec{h}$  does not lie in  $C(A)$  the column space of  $A$ , by Properties 6.3.7 it is inconsistent and there will be no exact solution. Nevertheless, a best-fit vector  $\vec{x}_f$  can be found, where  $A\vec{x}_f = \vec{h}_f$ , in the sense that  $\vec{h}_f$  will be the closest vector in the column space of  $A$  to  $\vec{h}$  in distance, i.e. the squared error

$$\left\| \vec{h} - \vec{h}_f \right\|^2 = \left\| \vec{h} - A\vec{x}_f \right\|^2$$

is minimized and  $\vec{h}_f = A\vec{x}_f$  (or  $\vec{x}_f$ ) is referred to as the *least-square approximation* to the system. From Properties 10.3.3, we know that such a shortest

distance will be achieved by the orthogonal projection of  $\vec{x}$  onto the column space of  $A$ . Notice that the distance and orthogonality can now be defined with respect to a general inner product other than the usual dot product. Therefore,  $\vec{h} - A\vec{x}_f$  will be in the orthogonal complement  $C(A)^\perp$ <sup>1</sup> of the column space  $C(A)$  of  $A$ , and we have

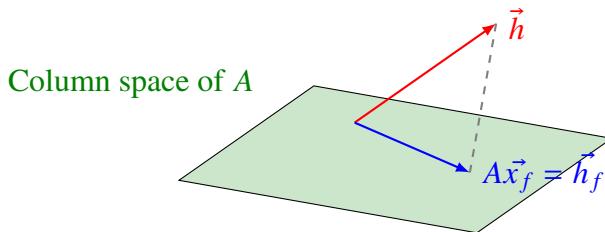
$$\langle A\vec{x}, \vec{h} - A\vec{x}_f \rangle = \langle \vec{x}, A^*(\vec{h} - A\vec{x}_f) \rangle = 0$$

Since this holds for any  $\vec{x}$ , by the last item of Properties 12.1.2 we have

$$A^*(\vec{h} - A\vec{x}_f) = \mathbf{0}$$

$$A^*A\vec{x}_f = A^*\vec{h}$$

This is called the **normal equation** due to the appearance of  $A^*A$ . So any  $\vec{x}_f$ <sup>2</sup> satisfying this equation will produce the least-square error. We may be tempted to multiply both sides of the equation by the inverse  $(A^*A)^{-1}$  to arrive at a formula for the least-square solution  $\vec{x}_f = (A^*A)^{-1}A^*\vec{h}$ . However, this is only allowed when this inverse indeed exists, and we now discuss under what condition it will happen.



Geometric view for the least-square approximation problem, where  $\vec{h}$  and the orthogonal projection of  $\vec{h}$  onto the two-dimensional column space of the  $3 \times 2$  matrix  $A$ , that is,  $\vec{h}_f = A\vec{x}_f$ , are  $\mathbb{R}^3$ .

**Properties 13.1.1.** For an  $m \times n$  matrix  $A$ ,  $A^*A$  and  $A$  has the same null space.

<sup>1</sup>which may not be equal to  $N(A^T)$  but rather  $N(A^*)$  if an inner product other than the standard one is used.

<sup>2</sup>The existence of at least one of such a vector is guaranteed by the uniqueness of orthogonal projection of  $\vec{h}$  onto  $C(A)$ .

*Proof.* It is obvious that if  $\vec{x} \in \mathcal{N}(A)$ ,  $A\vec{x} = \mathbf{0}$  then  $A^*A\vec{x} = \mathbf{0}$  so  $\vec{x} \in \mathcal{N}(A^*A)$  and  $\mathcal{N}(A) \subseteq \mathcal{N}(A^*A)$ . Now we only need to show that  $\mathcal{N}(A^*A) \subseteq \mathcal{N}(A)$ . For  $\vec{x} \in \mathcal{N}(A^*A)$ , we have  $A^*A\vec{x} = \mathbf{0}$  and hence  $\langle A^*A\vec{x}, \vec{x} \rangle = 0$ . Subsequently by the definition of an adjoint (Definition 12.2.1), we have  $\langle A\vec{x}, A\vec{x} \rangle = \|A\vec{x}\|^2 = 0$ . By Definition 12.1.1, we conclude that it must be  $A\vec{x} = \mathbf{0}$  so  $\vec{x} \in \mathcal{N}(A)$ , and thus  $\mathcal{N}(A^*A) \subseteq \mathcal{N}(A)$ . Hence  $\mathcal{N}(A^*A) = \mathcal{N}(A)$ , the null space of  $A^*A$  and  $A$  coincides.  $\square$

**Properties 13.1.2.** For an  $m \times n$  matrix  $A$ ,  $A^*A$  has the same rank as  $A$ . As a corollary, if  $A$  has linearly independent columns such that  $\text{rank}(A) = n$ , then  $A^*A$  is invertible.

*Proof.* By the Rank-nullity Theorem 6.3.9,  $\text{rank}(A^*A) + \text{nullity}(A^*A) = n = \text{rank}(A) + \text{nullity}(A)$  but by the previous properties  $\text{nullity}(A^*A) = \text{nullity}(A)$  so  $\text{rank}(A^*A) = \text{rank}(A)$  where  $A^*A$  is an  $n \times n$  matrix. Therefore if  $\text{rank}(A^*A) = \text{rank}(A) = n$ , then  $A^*A$  is full-rank and by Properties 6.3.10 it is invertible.  $\square$

Therefore, we have the following result.

**Theorem 13.1.3.** If  $A$  is a  $m \times n$  matrix with  $m \geq n$  and all its  $n$  column vectors are linearly independent, then for the system  $A\vec{x} = \vec{h}$ , there exists a unique best-fit solution

$$\vec{x}_f = (A^*A)^{-1}A^*\vec{h}$$

such that the square error  $\|\vec{h} - \vec{h}_f\|^2 = \|\vec{h} - A\vec{x}_f\|^2$  is minimized.

Notice that if  $\vec{h}$  already lies in the column space of  $A$ , then  $A\vec{x} = \vec{h}$  will have an exact solution and the best-fit solution will be identical to this exact solution. However, on the other extreme, if the column vectors in  $A$  are not linearly independent, then the best-fit solution will not be unique. Rather, the normal equation will still be consistent, but there are infinitely many possible solutions, each having the same least-square error. Also, if the standard real inner product

is used so that  $A^* = A^T$  and we by chance have the QR decomposition of  $A$ , then

$$\begin{aligned}\vec{x}_f &= (A^T A)^{-1} A^T \vec{h} \\ &= ((QR)^T (QR))^{-1} (QR)^T \vec{h} \\ &= (R^T Q^T QR)^{-1} (QR)^T \vec{h} \\ &= (R^T R)^{-1} R^T Q^T \vec{h} \\ &= R^{-1} (R^T)^{-1} R^T Q^T \vec{h} \\ &= R^{-1} Q^T \vec{h}\end{aligned}$$

where  $Q^T Q = I$  since  $Q$  is orthogonal matrix in which the column vectors form an orthonormal basis as indicated by Properties 7.2.5.

**Example 13.1.1.** Find the least-square solution to the overdetermined linear system

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix}$$

where the error is calculated with respect to the usual Euclidean distance, and also relative to the inner product as defined in Example 12.3.1.

*Solution.* If the standard inner product is used for defining lengths, then the least-square solution in Theorem 13.1.3 is reduced to

$$\begin{aligned}\vec{x}_f &= (A^T A)^{-1} A^T \vec{h} \\ &= \left( \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix}^T \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix}^T \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\ &= \begin{bmatrix} 11 & 18 \\ 18 & 36 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{11}{72} \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{19}{12} \end{bmatrix}$$

and the least-square error is

$$\begin{aligned} \|\vec{h} - A\vec{x}_f\|^2 &= \left\| \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{19}{12} \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} \frac{11}{3} \\ \frac{47}{6} \\ \frac{41}{6} \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} -\frac{2}{3} \\ \frac{1}{6} \\ \frac{1}{6} \end{bmatrix} \right\|^2 = \left( -\frac{2}{3}, \frac{1}{6}, \frac{1}{6} \right)^T \cdot \left( -\frac{2}{3}, \frac{1}{6}, \frac{1}{6} \right)^T = \frac{1}{2} \end{aligned}$$

For the inner product in Example 12.3.1, an appropriate adjoint to be used in this situation is

$$\begin{aligned} A^* &= C^{-1} A^T B \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \end{aligned}$$

where  $C$  can be picked to be any positive-definite matrix and we choose  $C = I$  for simplicity. Subsequently the least-square solution is

$$\begin{aligned} \vec{x}_f &= (A^* A)^{-1} A^* \vec{h} \\ &= \left( \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\ &= \begin{bmatrix} 34 & 62 \\ 62 & 120 \end{bmatrix}^{-1} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} \frac{30}{59} & -\frac{31}{118} \\ -\frac{31}{118} & \frac{17}{118} \end{bmatrix} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} \frac{10}{59} \\ \frac{103}{59} \end{bmatrix}$$

with the least-square error being

$$\begin{aligned} \|\vec{h} - A\vec{x}_f\|^2 &= (\vec{h} - A\vec{x}_f)^T B (\vec{h} - A\vec{x}_f) \\ &= \left( \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{10}{59} \\ \frac{103}{59} \end{bmatrix} \right)^T \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \left( \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{10}{59} \\ \frac{103}{59} \end{bmatrix} \right) \\ &= \begin{bmatrix} -\frac{39}{59} \\ \frac{30}{59} \\ -\frac{9}{59} \end{bmatrix}^T \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{39}{59} \\ \frac{30}{59} \\ -\frac{9}{59} \end{bmatrix} = \frac{36}{59} \end{aligned}$$

□

We close this section by confirming that the matrix  $T = A(A^*A)^{-1}A^*$  as in  $\vec{h}_f = A\vec{x}_f = A(A^*A)^{-1}A^*\vec{h}$  indeed represents an orthogonal projection (of  $\vec{h}$  onto the column space of  $A$ ). By Properties 12.3.2 we just need to check if  $T^2 = T = T^*$ . For the first equality, we have

$$\begin{aligned} T^2 &= (A(A^*A)^{-1}A^*)(A(A^*A)^{-1}A^*) \\ &= A(A^*A)^{-1}(A^*A)(A^*A)^{-1}A^* \\ &= A(A^*A)^{-1}(I)A^* = A(A^*A)^{-1}A^* = T \end{aligned}$$

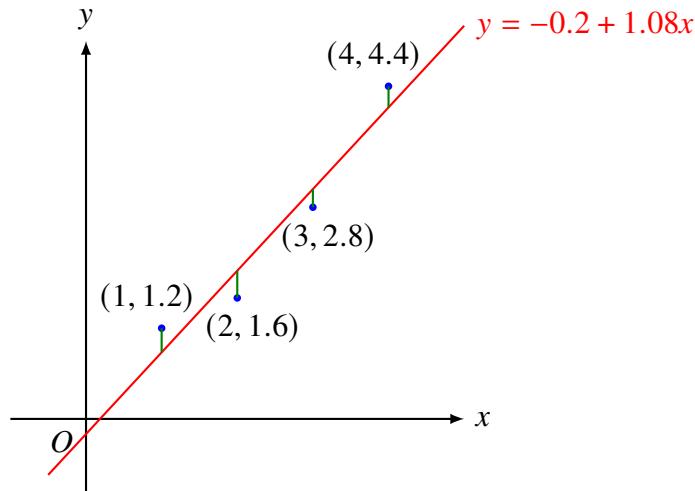
and for the second equality, we can use Properties 12.2.3 to get

$$\begin{aligned} T^* &= (A(A^*A)^{-1}A^*)^* \\ &= (A^*)^*((A^*A)^{-1})^*A^* \\ &= A((A^*A)^*)^{-1}A^* \\ &= A(A^*A)^{-1}A^* = T \end{aligned}$$

## 13.2 Linear Regression

### 13.2.1 Linear Regression for One Predictor Variable

**Linear regression** is a very important tool in Statistics that helps identify any linear trend in data and is based from least-square approximation. The simplest type of linear regression is fitting a straight line  $y = \alpha + \beta x$ , where  $\alpha$  and  $\beta$  are its intercept and slope, to  $n$  pairs of observation,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  such that the sum of squared errors  $\sum_{k=1}^n (y_k - (\alpha + \beta x_k))^2$  is minimized. In this context  $x$  and  $y$  are known as the *explanatory* and *response variable* respectively.



A toy linear regression model for 4 data points. The red straight line represents the best linear fit, and the green lines are the distances, or errors between the actual data and the regression line, whose sum of square is minimized.

To see how we can apply the results in the last section, we first rewrite the system into matrix form. The actual sampled values are given by  $\vec{y} = (y_1, y_2, \dots, y_n)^T$ , while the values predicted by the best-fit straight line will be in the form of  $\vec{y}^{\text{pred}} = (\alpha \mathbf{1} + \beta \vec{x})^T = (\alpha + \beta x_1, \alpha + \beta x_2, \dots, \alpha + \beta x_n)^T$ , where  $\mathbf{1}$  is a column

vector filled with ones,  $\alpha$  and  $\beta$  are the intercept and slope of the best-fit line to be determined. In other words, we are trying to find an optimal solution for the unknown parameters  $\alpha$  and  $\beta$  in the matrix system

$$\alpha \mathbf{1} + \beta \vec{x} = \vec{y}$$

or alternatively

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

so that the sum of squared errors  $\sum_{k=1}^n (y_k - y_k^{\text{pred}})^2 = \sum_{k=1}^n (y_k - (\alpha + \beta x_k))^2$  is minimized. Such a system is conveniently denoted as  $[X]\vec{\beta} = \vec{y}$ , where the first and second column of  $[X] = [\mathbf{1}|\vec{x}]$  represent the two predictors, the constant term which is a "hidden" predictor that gives rise to the  $y$ -intercept, and the linear term of  $x$ . Now, the sum of squared errors

$$\sum_{k=1}^n (y_k - (\alpha + \beta x_k))^2 = \left\| \vec{y} - [X]\vec{\beta} \right\|^2$$

will be minimized by the best-fit parameters which are computed via

$$\vec{\beta}_f = ([X]^T [X])^{-1} [X]^T \vec{y}$$

according to Theorem 13.1.3 where  $A = [X]$ , and simply  $A^* = [X]^T$  as the error is computed based on the usual dot product/Euclidean distance. Writing out the matrices in the formula, the parameters for single variable linear regression are

$$\begin{bmatrix} \alpha_f \\ \beta_f \end{bmatrix} = \left( \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\begin{aligned}
 &= \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} \\
 &= \frac{1}{n(\sum X^2) - (\sum X)^2} \begin{bmatrix} \sum X^2 & -\sum X \\ -\sum X & n \end{bmatrix} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} \\
 &= \frac{1}{n(\sum X^2) - (\sum X)^2} \begin{bmatrix} (\sum Y)(\sum X^2) - (\sum X)(\sum XY) \\ n(\sum XY) - (\sum X)(\sum Y) \end{bmatrix}
 \end{aligned}$$

where

$$\begin{aligned}
 \sum X &= x_1 + x_2 + \cdots + x_n \\
 \sum X^2 &= x_1^2 + x_2^2 + \cdots + x_n^2 \\
 \sum Y &= y_1 + y_2 + \cdots + y_n \\
 \sum XY &= x_1y_1 + x_2y_2 + \cdots + x_ny_n
 \end{aligned}$$

and we have used the results in Example 2.3.5 to calculate the inverse from the second to third line.

**Properties 13.2.1.** The best-fit parameters for single predictor variable linear regression are

$$\begin{aligned}
 \alpha_f &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2} \\
 \beta_f &= \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}
 \end{aligned}$$

such that the regression line  $y = \alpha_f + \beta_fx$  achieves the least-square error, i.e. the value of  $\sum_{k=1}^n (y_k - (\alpha_f + \beta_fx_k))^2$  is the smallest.

**Example 13.2.1.** Find the best linear fit for five  $(x, y)$  data points, which are  $(2, 4), (3, 6), (4, 7), (5, 9), (7, 11)$ .

*Solution.* We first compute the following quantities.

$$\sum X = 2 + 3 + 4 + 5 + 7 = 21$$

$$\sum(X^2) = 2^2 + 3^2 + 4^2 + 5^2 + 7^2 = 103$$

$$\sum Y = 4 + 6 + 7 + 9 + 11 = 37$$

$$\sum(XY) = (2)(4) + (3)(6) + (4)(7) + (5)(9) + (7)(11) = 176$$

Using Properties 13.2.1, with  $n = 5$ , the required parameters are

$$\alpha_f = \frac{\sum Y \sum(X^2) - \sum X \sum(XY)}{n \sum(X^2) - (\sum X)^2} = \frac{(37)(103) - (21)(176)}{(5)(103) - (21)^2} \approx 1.554$$

$$\beta_f = \frac{n \sum(XY) - \sum X \sum Y}{n \sum(X^2) - (\sum X)^2} = \frac{(5)(176) - (21)(37)}{(5)(103) - (21)^2} \approx 1.392$$

So the best linear fit is around  $y = 1.554 + 1.392x$ . □

### 13.2.2 Linear Regression for Multiple Predictor Variables

Sometimes we may need to predict a variable  $y$  with multiple predictor variables  $x^{(1)}, x^{(2)}, \dots$  as there can be different factors contributing to a phenomenon in Earth Science scenario. Linear regression for multiple predictor variables will then produce a best fit equation in the form of  $y = \alpha + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \dots$ . Extending the earlier results from Theorem 13.1.3, the computation of the parameters utilizes the same formula

$$\vec{\beta}_f = ([X]^T [X])^{-1} [X]^T \vec{y}$$

where in this case the quantities now involves more predictor variables,  $\vec{\beta}_f^T = (\alpha_f, \beta_{1f}, \beta_{2f}, \dots)^T$ , and the matrix  $[X] = [\mathbf{1} | \vec{x}^{(1)} | \vec{x}^{(2)} | \dots]$  holds the observed values of those multiple predictor variables column by column.

**Example 13.2.2.** The university carries out an experiment and try to quantify the effects of IQ and studying time on the GPA of students. 5 students participate and below are the data.

Students	IQ	Studying Time per Week (Hours)	GPA
Lily	104	5.2	3.56
Christy	111	6.1	3.71
Sabrina	107	8.3	3.73
Julia	106	3.4	3.34
Emily	109	9.6	3.88

Carry out a linear regression on the students' GPA against their IQ and studying time.

*Solution.* Denote the IQ and studying hours of the students by the variables  $x^{(1)}$  and  $x^{(2)}$ , and  $y$  for their GPA, we want to fit a linear relationship  $y = \alpha + \beta_1 x^{(1)} + \beta_2 x^{(2)}$ . The matrix  $[X] = [\mathbf{1} | \vec{x}^{(1)} | \vec{x}^{(2)}]$ , consisting of the observed predictors, is then

$$[X] = \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix}$$

The first column represents the intercept term. Subsequently, the fit is derived by

$$\begin{aligned} \vec{\beta}_f &= ([X]^T [X])^{-1} [X]^T \vec{y} \\ &= \left( \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix}^T \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix}^T \begin{bmatrix} 3.56 \\ 3.71 \\ 3.73 \\ 3.34 \\ 3.88 \end{bmatrix} \\ &\approx \begin{bmatrix} 1.4463 \\ 0.0162 \\ 0.0710 \end{bmatrix} \end{aligned}$$

So the regression model is  $\text{GPA} = 1.4463 + 0.0162(\text{IQ}) + 0.0710(\text{Studying Hrs})$ .

□

**Example 13.2.3.** Find a quadratic fit for Example 13.2.1.

*Solution.* The predictor variables are the constant term, the linear term  $x$ , as well as the newly added quadratic term  $x^2$ . The desired parameters are

$$\begin{aligned}\vec{\beta}_f &= ([X]^T [X])^{-1} [X]^T \vec{y} \\ &= \left( \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 7 \\ 2^2 & 3^2 & 4^2 & 5^2 & 7^2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2^2 \\ 1 & 3 & 3^2 \\ 1 & 4 & 4^2 \\ 1 & 5 & 5^2 \\ 1 & 7 & 7^2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 7 \\ 2^2 & 3^2 & 4^2 & 5^2 & 7^2 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \\ 7 \\ 9 \\ 11 \end{bmatrix} \\ &\approx \begin{bmatrix} -0.009 \\ 2.201 \\ -0.089 \end{bmatrix}\end{aligned}$$

The quadratic fit is thus  $y = -0.009 + 2.201x - 0.089x^2$ . □

Notice that in the above example despite we use the quadratic term  $x^2$  as a predictor, the regression is still *linear* in the sense that the regression model is a *linear* combination of these predictors. Usually, fitting a degree  $p$  polynomials in a single variable  $x$  to  $n$  points,  $p < n$ , will involve a matrix in the form like the one above:

$$[X] = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

This class of matrices is called **Vandermonde matrices**.<sup>3</sup>

---

<sup>3</sup>As long as  $p < n$  and the sampled points  $x_i$  are distinct, the columns of a Vandermonde matrix will be linearly independent.

### 13.2.3 Properties of Linear Regression

Before going to the next topic, we derive some features of linear regression. The errors, or called the *residuals*  $e_i = h_i - (h_f)_i$ , have a mean of zero. Using matrix notation, it means that the sum of components in the vector  $\vec{e} = \vec{h} - \vec{h}_f$  is zero. This can be seen from the very beginning of our derivation for the best-fit problem  $A\vec{x} = \vec{h}$ , where the condition has been  $A^*(\vec{h} - A\vec{x}_f) = A^*(\vec{h} - \vec{h}_f) = \mathbf{0}$ . For linear regression,  $A^* = [X]^T$  and it becomes  $[X]^T \vec{e} = \mathbf{0}$ . However, since the first column in  $[X]$  is  $\mathbf{1}$ , the first entry in  $[X]^T \vec{e}$  is just  $\mathbf{1} \cdot \vec{e}$ , which is just the sum of errors. As  $[X]^T \vec{e} = \mathbf{0}$ , the sum and hence the mean of errors must be zero.

Another property is that, if we denote the mean of  $y$  as  $\bar{y}$ , each actual data and predicted values as  $y_i$  and  $\hat{y}_i = y_i^{\text{pred}}$ , then

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

The term at the left hand side is called the *SST/Sum of Squares Total*, while the two terms at the right hand side are known as the *SSE/Sum of Squares Error* and *SSR/Sum of Squares Regression*. To prove this, we expand the SST, which gives

$$\begin{aligned} \text{SST} &= \sum_i (y_i - \bar{y})^2 \\ &= \sum_i ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) \\ &= \text{SSE} + \text{SSR} + 2 \sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) \end{aligned}$$

The remaining step is to show that the last term equals to zero. For simplicity, we work with the case of single predictor variable so there are two parameters  $\alpha$  and  $\beta$  only, but it can be easily generalized to multiple predictors and  $\beta_j$ . Expanding the product gives

$$\sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) = \sum_i (\hat{y}_i(y_i - \hat{y}_i)) - \bar{y} \sum_i (y_i - \hat{y}_i)$$

$$\begin{aligned}
 &= \sum_i ((\alpha + \beta x_i)(y_i - \hat{y}_i)) - \bar{y} \sum_i (y_i - \hat{y}_i) \\
 &= \beta \sum_i (x_i(y_i - \hat{y}_i)) - (\bar{y} - \alpha) \sum_i (y_i - \hat{y}_i) \\
 &= \beta \sum_i (x_i e_i) - (\bar{y} - \alpha) \sum_i e_i
 \end{aligned}$$

Using the same logic when we are investigating  $[X]^T \vec{e} = \mathbf{0}$  before, we know that

$$\begin{aligned}
 \mathbf{1} \cdot \vec{e} &= \sum_i e_i = 0 \\
 \vec{x} \cdot \vec{e} &= \sum_i (x_i e_i) = 0
 \end{aligned}$$

These two relations can also be derived by setting  $\partial(\sum_i e_i^2)/\partial\alpha = 0$  and  $\partial(\sum_i e_i^2)/\partial\beta = 0$  as the sum of squared errors reaches minimum at the point of best-fit. Substituting this two equations readily implies that the concerned quantity  $\sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) = 0$  is zero, and thus  $SST = SSE + SSR$ . We repeat these two results as follows.

**Properties 13.2.2.** The mean error of any linear regression is zero. Moreover, its sum of squares total (SST) is equal to sum of squares error (SSE) plus sum of squares regression (SSR), i.e.

$$\begin{aligned}
 \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\
 SST &= SSE + SSR
 \end{aligned}$$

The ratio of SSR to SST is known as the *coefficient of determination* and is denoted by  $R^2$  (hence sometimes we also simply call it *R-squared*). It is the proportion of variance in the response variable that can be explained by the predictors and thus indicates how good the linear fit is. It ranges from 0 to 1. The larger the value of  $R^2$ , the better the linear regression at predicting the target variable. If  $R^2$  is close to one, then the fit is almost perfect. However, it

is cautioned that there may be the problem of *overfitting*<sup>4</sup>, which means that the regression does well on the dataset it is trained on but fails horribly when extrapolated to data points outside this training set. On the other hand, if  $R^2$  is low, then the linear regression is ineffective, but it simply means that there is no apparent linear trend and the possibility of other forms of relationship, like a logarithmic one, existing between the explanatory variables and response variable, cannot be ruled out.

**Example 13.2.4.** Find the R-squared for the linear regression in Example 13.2.2.

*Solution.* We compute the SST and SSR as follows.

$$\begin{aligned}\bar{y} &= \frac{1}{5}(3.56 + 3.71 + 3.73 + 3.34 + 3.88) \\ &= 3.644\end{aligned}$$

$$\begin{aligned}\text{SST} &= \sum_i (y_i - \bar{y})^2 \\ &= (3.56 - 3.644)^2 + (3.71 - 3.644)^2 + (3.73 - 3.644)^2 \\ &\quad + (3.34 - 3.644)^2 + (3.88 - 3.644)^2 \\ &\approx 0.167\end{aligned}$$

$$\begin{aligned}\text{SSR} &= \sum_i (\hat{y}_i - \bar{y})^2 \\ &= ((1.4463 + 0.0162(104) + 0.0710(5.2)) - 3.644)^2 \\ &\quad + ((1.4463 + 0.0162(111) + 0.0710(6.1)) - 3.644)^2 \\ &\quad + ((1.4463 + 0.0162(107) + 0.0710(8.3)) - 3.644)^2 \\ &\quad + ((1.4463 + 0.0162(106) + 0.0710(3.4)) - 3.644)^2 \\ &\quad + ((1.4463 + 0.0162(109) + 0.0710(9.6)) - 3.644)^2 \\ &= (3.500 - 3.644)^2 + (3.678 - 3.644)^2 + (3.769 - 3.644)^2\end{aligned}$$

---

<sup>4</sup>For  $n$  distinct sets of data, it is always possible to achieve a perfect fit ( $R^2 = 1$ ) using  $n - 1$  predictors (plus the intercept term) in the linear regression even when they may not be really related to the predictand and this is an extreme case of overfitting.

$$\begin{aligned} & + (3.405 - 3.644)^2 + (3.894 - 3.644)^2 \\ & \approx 0.157 \end{aligned}$$

Hence  $R^2 \approx 0.157/0.167 = 0.94$  which indicates a good fit.  $\square$

### 13.3 Earth Science Applications

**Example 13.3.1.** Find the least-square approximation to the temperature measurement problem in Example 1.4.4.

*Solution.* In Example 3.3.3 we have found that the overdetermined system has no exact solution. To find the least-square solution we invoke the formula in Theorem 13.1.3 where

$$\vec{x}_f = (A^T A)^{-1} A^T \vec{h}$$

with

$$A = \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \quad \text{and} \quad \vec{h} = \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix}$$

So

$$\begin{aligned} \vec{x}_f &= \left( \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix}^T \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \right)^{-1} \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix}^T \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix} \\ &\approx \begin{bmatrix} 0.01357 \\ 0.00058 \end{bmatrix} \end{aligned}$$

So the approximated solution to the temperature gradients that minimizes the squared errors is  $\partial T / \partial x = 1.36 \times 10^{-2} \text{ } ^\circ\text{C/km}$  and  $\partial T / \partial y = 0.58 \times 10^{-3} \text{ } ^\circ\text{C/km}$ .  $\square$

## 13.4 Python Programming

We will use Example 13.2.2 to demonstrate how to do a linear regression in Python. The `statsmodels` module is recommended for this. First, let's provide the data.

```
import numpy as np
import statsmodels.api as sm

X = np.array([[104, 5.2],
              [111, 6.1],
              [107, 8.3],
              [106, 3.4],
              [109, 9.6]])
Y = np.array([3.56, 3.71, 3.73, 3.34, 3.88])
```

We need to append the constant term by using the `sm.add_constant` function.

```
Xc = sm.add_constant(X, prepend=False)
```

Subsequently, create an `sm.OLS` (stands for "Ordinary Least Squares") object and call the `fit` method that takes the predictand/predictors in the first/second argument.

```
mod = sm.OLS(Y, Xc)
res = mod.fit()
```

The `summary` method will output a table that provides relevant parameters about the linear regression like the R-squared and p-values.

```
print(res.summary())
```

and the `predict` method will return the predicted values from an input dataset according to the linear regression.

```
print(res.predict(Xc))
```

This gives `[3.495 3.672 3.764 3.400 3.888]` which is slightly different from what we have obtained in Example 13.2.4 as our manual calculation of the regression coefficients inevitably contains round-off errors.

OLS Regression Results									
Dep. Variable:	y	R-squared:	0.938						
Model:	OLS	Adj. R-squared:	0.875						
Method:	Least Squares	F-statistic:	15.04						
Date:	Mon, 21 Oct 2024	Prob (F-statistic):	0.0623						
Time:	11:12:06	Log-Likelihood:	8.3427						
No. Observations:	5	AIC:	-10.69						
Df Residuals:	2	BIC:	-11.86						
Df Model:	2								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
-----									
x1	0.0162	0.015	1.090	0.389	-0.048	0.080			
x2	0.0710	0.016	4.370	0.049	0.001	0.141			
const	1.4463	1.548	0.934	0.449	-5.216	8.109			
-----									
Omnibus:		nan	Durbin-Watson:		0.883				
Prob(Omnibus):		nan	Jarque-Bera (JB):		0.451				
Skew:		0.142	Prob(JB):		0.798				
Kurtosis:		1.556	Cond. No.			5.17e+03			
-----									

## 13.5 Exercise

**Exercise 13.1** Show that if  $A$  is an invertible  $n \times n$  square matrix then the least-square solution of  $A\vec{x} = \vec{h}$  given in Theorem 13.1.3

$$\vec{x}_f = (A^* A)^{-1} A^* \vec{h}$$

will be reduced to simply

$$\vec{x}_f = A^{-1} \vec{h}$$

the exact solution as derived in Section 3.2.2.

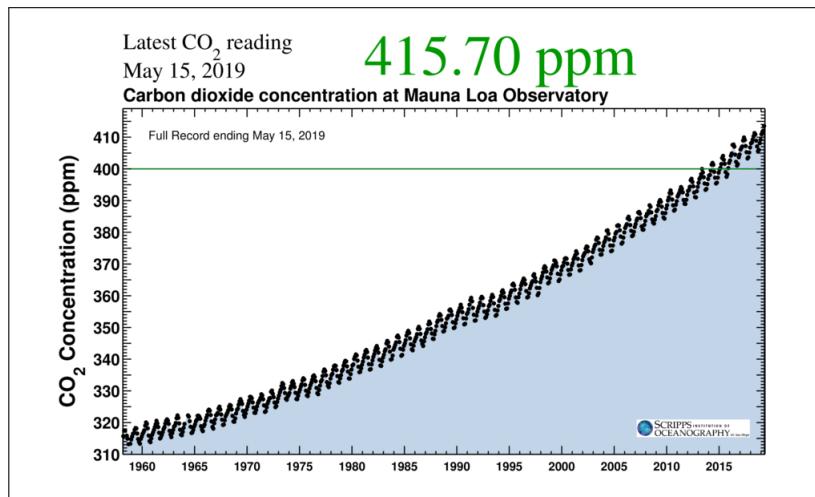
**Exercise 13.2** Find the least-square solution to the overdetermined linear system below.

$$\begin{cases} x + 2y - 3z = 4 \\ x - y + 3z = 5 \\ 2x + y - z = 1 \\ x + y + z = 2 \end{cases}$$

**Exercise 13.3** Find a linear fit for the following data about sea level pressure and temperature measured at a weather station.

Temperature (° C)	10	12	12	13	16	17
Pressure (hPa)	1022.1	1019.5	1018.9	1017.6	1014.3	1013.5

**Exercise 13.4** Find a linear fit and a quadratic fit for the following atmospheric data regarding global carbon dioxide level. Also, calculate the root mean square error for each fit.



Years passed since 1960	0	5	10	15	20	25
CO <sub>2</sub> level (ppm)	316.9	320.0	325.7	331.1	338.8	346.4
Years passed since 1960	30	35	40	45	50	55
CO <sub>2</sub> level (ppm)	354.4	361.0	369.7	380.0	390.1	401.0

(Data from: NOAA ([https://gml.noaa.gov/webdata/ccgg/trends/co2/co2\\_annmean\\_mlo.csv](https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_annmean_mlo.csv)))

**Exercise 13.5** Radioactive decay is modelled by  $N = N_0 e^{-kt}$ , where  $N_0$  and  $k$  are the initial concentration and the decay constant respectively. While the formula is exponential, not linear, the technique of linear regression can still be applied if the data undergoes linearization. Show that by the substitution  $n = \ln N$  the equation can be transformed into a linear equation  $n = \ln N = \ln N_0 - kt = n_0 - kt$ . Hence find the best linear fit on  $(t, n)$  by finding the parameters  $(n_0, k)$  from the experimental data on the radioactive isotope Sodium-24 below and recover the decay constant and initial mass.

Time passed (hr)	6	8	12	24	36	48
Mass (g)	75.8	69.1	57.3	33.0	18.8	10.8

**Exercise 13.6** A commercial study investigates eight companies that sell the same type of products and are also similar in size. The following table summarizes their revenues, R&D expense, employee wages, and amounts of advertisement (the last three items are normalized scores).

	Revenues	R&D	Wages	Advertisement
Company 1	135%	2.5	1.7	0.8
Company 2	128%	1.6	1.8	1.5
Company 3	119%	1.8	0.5	2.4
Company 4	121%	0.3	1.5	1.2
Company 5	126%	1.9	1.6	1.3
Company 6	112%	0.8	1.1	0.2
Company 7	143%	2.2	2.4	1.1
Company 8	135%	1.5	2.2	2.3

Construct a linear regression model for the revenue against the three factors that follow. What is the  $R^2$  of this regression?



## Chapter 14

# Discrete Fourier Transform (DFT)

---

We now discuss a powerful mathematical tool that can be approached from a least-square approximation view point. In the last chapter, we have seen how to fit a polynomial curve to some data. It is then natural to ask further, whether there are other suitable types of curves that can be used for data fitting as well. Remember in Chapter 12 we have derived the Fourier series which can expand any reasonable function in sines and cosines. Coincidentally, in the area of Earth Science, many phenomena can be described by the notion of *waves*, which are often taken to be *sinusoidal* during their derivation, e.g. atmospheric gravity wave, seismic waves, electromagnetic wave. However, in real-life our sampling of data will be discrete. Therefore, we may want to know if we can do something similar and interpolate a discrete time-series by fitting sines and cosines to it, and this is the central idea of what is known as the *Discrete Fourier Transform*.

## 14.1 Mathematical Ideas of DFT

### 14.1.1 From Fourier Series to a Prototype of DFT

By Properties 12.3.5, we know that any function  $f(x)$  in the interval  $[-\pi, \pi]$  can be written as

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx)$$

where the Fourier coefficients  $a_m$  and  $b_n$  are given by Formulae (12.5) and (12.6). By a change of variable  $x = \frac{2\pi}{N}t - \pi$ , the interval is scaled to  $t \in [0, N]$  where the cosines and sines are now in the form of

$$\begin{aligned}\cos\left(m\left(\frac{2\pi}{N}t - \pi\right)\right) &= \cos\left(\frac{2m\pi}{N}t\right) \\ \sin\left(n\left(\frac{2\pi}{N}t - \pi\right)\right) &= \sin\left(\frac{2n\pi}{N}t\right)\end{aligned}$$

The negative sign due to the  $m\pi$  or  $n\pi$  term inside is absorbed whenever needed. Since it involves a linear variable transformation only, these cosines and sines

$$\left\{\sqrt{\frac{2}{N}}\sin\left(\frac{2\pi}{N}t\right), \sqrt{\frac{2}{N}}\sin\left(\frac{2\pi(2)}{N}t\right), \sqrt{\frac{2}{N}}\sin\left(\frac{2\pi(3)}{N}t\right), \dots, \frac{1}{\sqrt{N}}, \right. \\ \left. \sqrt{\frac{2}{N}}\cos\left(\frac{2\pi}{N}t\right), \sqrt{\frac{2}{N}}\cos\left(\frac{2\pi(2)}{N}t\right), \sqrt{\frac{2}{N}}\cos\left(\frac{2\pi(3)}{N}t\right), \dots\right\}$$

are still an orthonormal basis to the new  $L^2[0, N]$  space mapped from the initial  $L^2[-\pi, \pi]$ <sup>1</sup>. The Fourier coefficients are now computed by

$$a_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos\left(\frac{2m\pi}{N}t\right) dt = \frac{2}{N} \int_0^N f(t) \cos\left(\frac{2m\pi}{N}t\right) dt \quad (14.1)$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin\left(\frac{2n\pi}{N}t\right) dt = \frac{2}{N} \int_0^N f(t) \sin\left(\frac{2n\pi}{N}t\right) dt \quad (14.2)$$

where we have written  $f(t)$  in place of  $f(x) = f(\frac{2\pi}{N}t - \pi)$ . The partial sum of the Fourier expansion of  $f(t)$  up to degree  $p$

$$S_p(t) = \frac{a_0}{2} + \sum_{m=1}^p a_m \cos\left(\frac{2m\pi}{N}t\right) + \sum_{n=1}^p b_n \sin\left(\frac{2n\pi}{N}t\right) \quad (14.3)$$

---

<sup>1</sup>Denote the original Fourier basis functions by  $\varphi_j(x) \in L^2[-\pi, \pi]$ . We will only show the part of linear independence and omit the justification for span and orthogonality. By Theorem 6.1.9, since they form a basis as given,  $c_1\varphi_1(x) + c_2\varphi_2(x) + \dots = 0(x) \equiv 0$  where  $0(x)$  is the zero function, has the trivial solution  $c_j = 0$  as its only solution. Assume the form of the linear change in variable is  $x = \alpha t + \beta = X(t)$ . Replacing  $x$  by  $X(t)$  (possible since  $\alpha \neq 0$ ) then immediately produces the equality  $c_1\varphi_1(X(t)) + c_2\varphi_2(X(t)) + \dots = 0(X(t)) \equiv 0$  which automatically inherits the desired property of possessing only the trivial solution as well and shows that the new basis functions are now  $\hat{\varphi}_j(t) = \varphi_j(X(t))$ .

$$\begin{aligned}
 &= \frac{a_0}{2} + a_1 \cos\left(\frac{2\pi}{N}t\right) + a_2 \cos\left(\frac{2\pi(2)}{N}t\right) + \cdots + a_p \cos\left(\frac{2\pi p}{N}t\right) \\
 &\quad + b_1 \sin\left(\frac{2\pi}{N}t\right) + b_2 \sin\left(\frac{2\pi(2)}{N}t\right) + \cdots + b_p \sin\left(\frac{2\pi p}{N}t\right)
 \end{aligned}$$

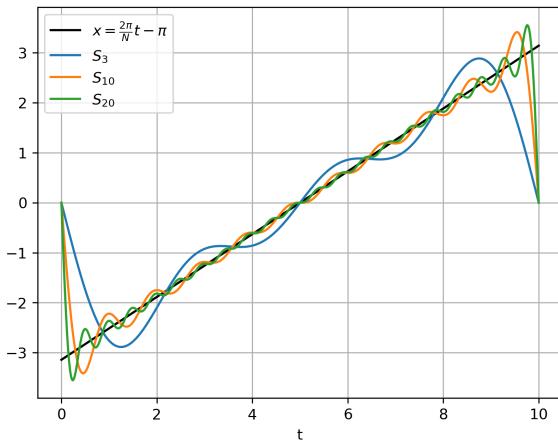
will then be the best approximation of the function  $f(t)$  using sines and cosines up to order  $p$  with distance defined with respect to the inner product of Equation (12.1), a.k.a. the best-fit trigonometric polynomial of degree  $p$ . This is known as the *Best Approximation Theorem* which directly follows from Properties 10.3.3 and related discussion about Fourier basis in Section 12.2 because such a Fourier expansion  $S_p(t)$  is essentially an orthogonal projection of  $f(t)$  onto the subspace spanned by the corresponding orthonormal trigonometric basis up to order  $p$ . Hence, higher the degree of the partial sum of the Fourier expansion, the more closer the approximation to the given function. We use Example 12.3.2 as an illustration. The appropriate Fourier series of  $x = f(t) = \frac{2\pi}{N}t - \pi$  is now

$$\begin{aligned}
 f(t) &= \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi}{N}t\right) \\
 &= -2 \sum_{n=1}^{\infty} \frac{1}{n} \sin\left(\frac{2n\pi}{N}t\right)
 \end{aligned}$$

where  $b_n = -\frac{2}{n}$ . The following plot shows the original function as well as the partial sum of its Fourier expansion up to degree 3, 10, 20. From this, we can see that as the degree  $p$  goes up,  $S_p(t)$  becomes a better approximation to the straight line  $y = \frac{2\pi}{N}t - \pi$ <sup>2</sup>.

---

<sup>2</sup>Notice that at the end points, the values of Fourier series of  $f(t) = \frac{2\pi}{N}t - \pi$  stay at 0. It is because the Fourier series works as the periodic extension of the original function (see Footnote 16 in Chapter 12) and will converge to the average of the left and right limits  $(f_-(t) + f_+(t))/2$  at point of discontinuity. Therefore, the end point is essentially a discontinuity where one of the side limits is  $\pi$  while the another is  $-\pi$ , so it converges to  $((-\pi) + \pi)/2 = 0$ .



Eventually, as  $p \rightarrow \infty$  and the expression tends to the full Fourier series,  $S_p(t)$  will converge to  $f(t)$  in the  $L^2$  sense<sup>3</sup>, with respect to the inner product of Equation (12.1). This is derived heuristically below.

$$\begin{aligned}
 & \lim_{p \rightarrow \infty} \|f(t) - S_p(t)\|^2 = \lim_{p \rightarrow \infty} \left( \int_0^N |f(t) - S_p(t)|^2 dt \right) \\
 &= \lim_{p \rightarrow \infty} \langle f(t) - S_p(t), f(t) - S_p(t) \rangle \\
 &= \lim_{p \rightarrow \infty} \langle (\lim_{p \rightarrow \infty} S_p(t)) - S_p(t), (\lim_{p \rightarrow \infty} S_p(t)) - S_p(t) \rangle \\
 &\quad ((d) \text{ of Theorem 12.3.4}) \\
 &= \lim_{p \rightarrow \infty} \left\langle \sum_{k=p+1}^{\infty} [a_k \cos\left(\frac{2k\pi}{N}t\right) + b_k \sin\left(\frac{2k\pi}{N}t\right)], \right. \\
 &\quad \left. \sum_{k=p+1}^{\infty} [a_k \cos\left(\frac{2k\pi}{N}t\right) + b_k \sin\left(\frac{2k\pi}{N}t\right)] \right\rangle \\
 &= \lim_{p \rightarrow \infty} \sum_{k=p+1}^{\infty} \frac{N}{2} (a_k^2 + b_k^2) \quad (\text{Orthogonality within the Fourier basis})
 \end{aligned}$$

<sup>3</sup>This is also called *convergence in mean* and is different from *pointwise convergence* that is more intuitive for most people. The rigorous treatment to this requires Measure Theory, and again, Functional Analysis and will not be pursued here.

$$= 0$$

Notice that the last step requires crucially that the series  $\{a_k^2 + b_k^2\}$ ,  $k = p+1, \dots, \infty$  will converge to zero eventually, i.e.  $a_k, b_k \rightarrow 0$  as  $k \rightarrow \infty$ . This will be justified when we introduce the Parseval's Theorem later on.

Returning back to practices in Earth Science, and other fields like Engineering, often we are not given a function that has a closed form to work with. Instead, we collect data from measurements at a fixed sampling rate, and what we obtain is a *discrete time series*. However, we can still try to apply the idea of Fourier, and try to approximate and interpolate the finite time series with sinusoidal functions. Assume that, we have  $N$  data points collected for the time series  $f(t)$ , evenly spaced at time  $t = 0, 1, 2, \dots, N-1$ . Further assume that we only use a few of sines and cosines for the approximation, such that the degree  $p$  much is less than the number of data points  $N$ , specifically,  $2p+1 \leq N$ . The suitable Fourier approximation then will be in the form of Equation (14.3) where  $N$  is the period of the time series. Since we only have finite data points, we cannot carry out the needed integration to compute (14.1) and (14.2). Nevertheless, we can borrow the idea in the last chapter and do the approximation in the way such that the Fourier partial sum will achieve the least-square error, summed over the sampling points. Subsequently, the system to be optimized will be

$$\left\{ \begin{array}{l} C_0 + A_1 \cos\left(\frac{2\pi}{N}(0)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(0)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(0)\right) \\ + B_1 \sin\left(\frac{2\pi}{N}(0)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(0)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(0)\right) \end{array} \right. = f(0)$$

$$\left\{ \begin{array}{l} C_0 + A_1 \cos\left(\frac{2\pi}{N}(1)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(1)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(1)\right) \\ + B_1 \sin\left(\frac{2\pi}{N}(1)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(1)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(1)\right) \end{array} \right. = f(1)$$

$$\left\{ \begin{array}{l} C_0 + A_1 \cos\left(\frac{2\pi}{N}(2)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(2)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(2)\right) \\ + B_1 \sin\left(\frac{2\pi}{N}(2)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(2)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(2)\right) \end{array} \right. = f(2)$$

$$\vdots$$

$$\left\{ \begin{array}{l} C_0 + A_1 \cos\left(\frac{2\pi}{N}(N-1)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(N-1)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(N-1)\right) \\ + B_1 \sin\left(\frac{2\pi}{N}(N-1)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(N-1)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(N-1)\right) \end{array} \right. = f(N-1)$$

where we have replaced the small letters  $a_k, b_k$  by capital letters  $A_k, B_k$  and  $a_0$  by  $C_0$ . Or in the form of a matrix system,  $G\vec{\beta} = \vec{d}$ , where

$$G = \begin{bmatrix} 1 & 1 & 0 & \cdots & 1 & 0 \\ 1 & \cos\left(\frac{2\pi}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \cdots & \cos\left(\frac{2\pi p}{N}\right) & \sin\left(\frac{2\pi p}{N}\right) \\ 1 & \cos\left(\frac{2\pi(2)}{N}\right) & \sin\left(\frac{2\pi(2)}{N}\right) & \cdots & \cos\left(\frac{2\pi p(2)}{N}\right) & \sin\left(\frac{2\pi p(2)}{N}\right) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \cos\left(\frac{2\pi(n-1)}{N}\right) & \sin\left(\frac{2\pi(n-1)}{N}\right) & \cdots & \cos\left(\frac{2\pi p(N-1)}{N}\right) & \sin\left(\frac{2\pi p(N-1)}{N}\right) \end{bmatrix}$$

is a  $N \times (2p + 1)$  matrix and

$$\vec{\beta} = \begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \\ A_p \\ B_p \end{bmatrix} \quad \vec{d} = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \\ f(N - 1) \end{bmatrix}$$

are vectors with  $2p + 1$  and  $N$  entries respectively. The least-square method then sets out to find the best-fit parameters  $\vec{\beta} = (C_0, A_1, B_1, \dots, A_p, B_p)^T$  for this system of equations. From Theorem 13.1.3, we know that the best parameters are found by

$$\vec{\beta}_f = (G^T G)^{-1} G^T \vec{d}$$

However, we can greatly simplify the expression, by noticing that every column vector of a sine/cosine series is orthogonal to each other. If we write each sine/cosine term as a complex exponential using Euler's formula in Definition 8.1.6, then the column vectors of a sine-cosine pair in  $G$  with a frequency of  $2\pi k/N$ , where both  $k$  and  $N$  are integers, can be expressed by the real and imaginary parts of

$$\left\{ \exp\left(i \frac{2\pi k}{N} t\right) \right\}_{t=0,1,2,\dots,N-1} = \left\{ \cos\left(\frac{2\pi k}{N} t\right) + i \sin\left(\frac{2\pi k}{N} t\right) \right\}_{t=0,1,2,\dots,N-1}$$

or in the other way around,

$$\left\{ \exp\left(-i\frac{2\pi k}{N}t\right) \right\}_{t=0,1,2,\dots,N-1} = \left\{ \cos\left(\frac{2\pi k}{N}t\right) - i \sin\left(\frac{2\pi k}{N}t\right) \right\}_{t=0,1,2,\dots,N-1}$$

We now first prove that the two column vectors of a sine-cosine pair that has the same frequency are orthogonal. We take the sum of squares of the first expression, which gives

$$\begin{aligned} \sum_{t=0}^{N-1} (\exp\left(i\frac{2\pi k}{N}t\right))^2 &= \sum_{t=0}^{N-1} (\cos\left(\frac{2\pi k}{N}t\right) + i \sin\left(\frac{2\pi k}{N}t\right))^2 \\ \sum_{t=0}^{N-1} \exp\left(i\frac{4\pi k}{N}t\right) &= \sum_{t=0}^{N-1} (\cos^2(\frac{2\pi k}{N}t) + 2i \sin(\frac{2\pi k}{N}t) \cos(\frac{2\pi k}{N}t) - \sin^2(\frac{2\pi k}{N}t)) \end{aligned}$$

Notice that the left hand side is a geometric sequence with a common ratio  $r = \exp(i4\pi k/N)$ , whose sum is seen to be

$$\frac{1 - r^N}{1 - r} = \frac{1 - \exp(i4\pi k)}{1 - \exp(i4\pi k/N)} = 0$$

as  $\exp(i4\pi k)$  is just 1. By comparing the real and imaginary parts, we know that

$$\sum_{t=0}^{N-1} \cos^2\left(\frac{2\pi k}{N}t\right) = \sum_{t=0}^{N-1} \sin^2\left(\frac{2\pi k}{N}t\right) \quad (14.4)$$

$$\sum_{t=0}^{N-1} \sin\left(\frac{2\pi k}{N}t\right) \cos\left(\frac{2\pi k}{N}t\right) = 0 \quad (14.5)$$

The second equation shows that the two column vectors representing the sine and cosine waves of the same frequency have a dot product of zero and hence are orthogonal.

Utilizing the complex formulations, we can also prove that column vectors of sine (or cosine) functions with different frequencies are orthogonal as well. Here we prove one of the cases, where the first series is a sine at a frequency of  $2\pi k/N$ ,

and the second series is also a sine, of a frequency of  $2\pi l/N$ , where  $k \neq l$  are both integers. We start by considering the sum of products between

$$\left\{ \exp\left(\iota \frac{2\pi k}{N} t\right) \right\}_{t=0,1,2,\dots,N-1} = \left\{ \cos\left(\frac{2\pi k}{N} t\right) + \iota \sin\left(\frac{2\pi k}{N} t\right) \right\}_{t=0,1,2,\dots,N-1}$$

and

$$\left\{ \exp\left(\iota \frac{2\pi l}{N} t\right) \right\}_{t=0,1,2,\dots,N-1} = \left\{ \cos\left(\frac{2\pi l}{N} t\right) + \iota \sin\left(\frac{2\pi l}{N} t\right) \right\}_{t=0,1,2,\dots,N-1}$$

The analysis is similar to the one above. Particularly, the L.H.S. is zero and by considering the real part of the expression on R.H.S., we have

$$\sum_{t=0}^{N-1} \cos\left(\frac{2\pi k}{N} t\right) \cos\left(\frac{2\pi l}{N} t\right) - \sum_{t=0}^{N-1} \sin\left(\frac{2\pi k}{N} t\right) \sin\left(\frac{2\pi l}{N} t\right) = 0 \quad (14.6)$$

as long as  $k + l$  is not the integer multiples of  $N$ . (why?)<sup>4</sup> We can also consider another sum of products between  $\left\{ \exp\left(\iota \frac{2\pi k}{N} t\right) \right\}_{t=0,1,2,\dots,N-1}$  and

$$\left[ \exp\left(-\iota \frac{2\pi l}{N} t\right) \right]_{t=0,1,2,\dots,N-1} = \left[ \cos\left(\frac{2\pi l}{N} t\right) - \iota \sin\left(\frac{2\pi l}{N} t\right) \right]_{t=0,1,2,\dots,N-1}$$

Again by looking at the real part, this yields another relation as<sup>5</sup>

$$\sum_{t=0}^{N-1} \cos\left(\frac{2\pi k}{N} t\right) \cos\left(\frac{2\pi l}{N} t\right) + \sum_{t=0}^{N-1} \sin\left(\frac{2\pi k}{N} t\right) \sin\left(\frac{2\pi l}{N} t\right) = 0 \quad (14.7)$$

---

<sup>4</sup>The sum of complex exponentials on L.H.S. will then become

$$\begin{aligned} \sum_{t=0}^{N-1} \exp\left(\iota \frac{2\pi k}{N} t\right) \exp\left(\iota \frac{2\pi l}{N} t\right) &= \sum_{t=0}^{N-1} \exp\left(\iota \frac{2\pi k}{N} t\right) \exp\left(\iota \frac{2\pi(qN - k)}{N} t\right) \\ &= \sum_{t=0}^{N-1} \exp(\iota 2\pi q t) = \sum_{t=0}^{N-1} 1 = N \neq 0 \end{aligned}$$

and the argument fails.

<sup>5</sup>Similarly we have the constraint that  $k$  and  $l$  are not differed by an integer multiple of  $N$ .

From the two derived equations (14.6) and (14.7), we can hence conclude that

$$\sum_{t=0}^{N-1} \cos\left(\frac{2\pi k}{N}t\right) \cos\left(\frac{2\pi l}{N}t\right) = 0$$

$$\sum_{t=0}^{N-1} \sin\left(\frac{2\pi k}{N}t\right) \sin\left(\frac{2\pi l}{N}t\right) = 0$$

The orthogonality relations can be proven between a sine and cosine of different frequencies as well in a very similar essence. We will now establish the last result, the dot product of any cosine (or sine) column vector with a specific frequency  $2\pi k/N$  with itself. We can consider the sum of products between

$$\left\{ \exp\left(i\frac{2\pi k}{N}t\right) \right\}_{t=0,1,2,\dots,N-1} \quad \text{and} \quad \left\{ \exp\left(-i\frac{2\pi k}{N}t\right) \right\}_{t=0,1,2,\dots,N-1}$$

This time, the L.H.S. is not a geometric series, but rather  $N$  terms of 1. The relation is then

$$\begin{aligned} \sum_{t=0}^{N-1} \cos^2\left(\frac{2\pi k}{N}t\right) + \sum_{t=0}^{N-1} \sin^2\left(\frac{2\pi k}{N}t\right) &= \sum_{t=0}^{N-1} \exp\left(i\frac{2\pi k}{N}t\right) \exp\left(-i\frac{2\pi k}{N}t\right) \\ &= \sum_{t=0}^{N-1} (1) \\ &= N \end{aligned} \tag{14.8}$$

Actually it can also be observed from the fact that the sum of a sine-cosine square pair is 1. Not long before, we have arrived at Equation (14.4)

$$\sum_{t=0}^{N-1} \cos^2\left(\frac{2\pi k}{N}t\right) = \sum_{t=0}^{N-1} \sin^2\left(\frac{2\pi k}{N}t\right)$$

Solving these two equations (14.4) and (14.8) yields

$$\sum_{t=0}^{N-1} \cos^2\left(\frac{2\pi k}{N}t\right) = \sum_{t=0}^{N-1} \sin^2\left(\frac{2\pi k}{N}t\right) = \frac{N}{2} \tag{14.9}$$

Hence the product  $G^T G$ , where each entry will be the dot product between the series of sines and cosines, will be

$$G^T G = \begin{bmatrix} N & 0 & 0 & \dots \\ 0 & \frac{N}{2} & 0 & \\ 0 & 0 & \frac{N}{2} & \\ \vdots & & & \ddots \end{bmatrix}$$

and

$$(G^T G)^{-1} = \frac{1}{N} \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & 2 & 0 & \\ 0 & 0 & 2 & \\ \vdots & & & \ddots \end{bmatrix}$$

So the best-fit parameters are

$$\begin{aligned} \vec{\beta}_f &= (G^T G)^{-1} G^T \vec{d} \\ \begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \end{bmatrix} &= \frac{1}{N} \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & 2 & 0 & \\ 0 & 0 & 2 & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \dots \\ 1 & \cos\left(\frac{2\pi}{N}\right) & \cos\left(\frac{2\pi(2)}{N}\right) & \\ 0 & \sin\left(\frac{2\pi}{N}\right) & \sin\left(\frac{2\pi(2)}{N}\right) & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} \\ \begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \end{bmatrix} &= \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \dots \\ 2 & 2 \cos\left(\frac{2\pi}{N}\right) & 2 \cos\left(\frac{2\pi(2)}{N}\right) & \\ 0 & 2 \sin\left(\frac{2\pi}{N}\right) & 2 \sin\left(\frac{2\pi(2)}{N}\right) & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} \end{aligned}$$

Detailed expressions are

$$\begin{aligned} C_0 &= \frac{1}{N}(f(0) + f(1) + f(2) + \dots + f(N-1)) \\ &= \frac{1}{N} \sum_{t=0}^{n-1} f(t) \end{aligned} \tag{14.10}$$

$$\begin{aligned} A_k &= \frac{2}{N}(f(0) + f(1) \cos\left(\frac{2\pi k}{N}\right) + \cdots + f(N-1) \cos\left(\frac{2\pi(N-1)k}{N}\right)) \\ &= \frac{2}{N} \sum_{t=0}^{N-1} f(t) \cos\left(\frac{2\pi kt}{N}\right) \end{aligned} \quad (14.11)$$

$$\begin{aligned} B_k &= \frac{2}{N}(f(1) \sin\left(\frac{2\pi k}{N}\right) + \cdots + f(N-1) \sin\left(\frac{2\pi(N-1)k}{N}\right)) \\ &= \frac{2}{N} \sum_{t=0}^{N-1} f(t) \sin\left(\frac{2\pi kt}{N}\right) \end{aligned} \quad (14.12)$$

That makes sense, at least they seem to be. But in the short exercise that follows the example below you will immediately notice there is a big caveat to this prototype.

**Example 14.1.1.** Fit the following time-series with the Fourier basis up to order  $p = 1$ , where  $f(0) = 4, f(1) = 1, f(2) = 2, f(3) = 3, f(4) = 1$ .

*Solution.* The degree is  $p = 1$  and means that there are only three components, which are the constant term, and a sine-cosine pair with a frequency of  $\frac{2\pi k}{N}$  where  $k = 1, N = 5$ . The best-fit parameters will be

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 \cos\left(\frac{2\pi}{5}\right) & 2 \cos\left(\frac{2\pi(2)}{5}\right) & 2 \cos\left(\frac{2\pi(3)}{5}\right) & 2 \cos\left(\frac{2\pi(4)}{5}\right) \\ 0 & 2 \sin\left(\frac{2\pi}{5}\right) & 2 \sin\left(\frac{2\pi(2)}{5}\right) & 2 \sin\left(\frac{2\pi(3)}{5}\right) & 2 \sin\left(\frac{2\pi(4)}{5}\right) \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \end{bmatrix}$$

$$C_0 = \frac{1}{5}(4 + 1 + 2 + 3 + 1) = \frac{11}{5}$$

$$\begin{aligned} A_1 &= \frac{2}{5}(4 + 1 \cos\left(\frac{2\pi}{5}\right) + 2 \cos\left(\frac{2\pi(2)}{5}\right) + 3 \cos\left(\frac{2\pi(3)}{5}\right) + 1 \cos\left(\frac{2\pi(4)}{5}\right)) \\ &\approx 0.229 \end{aligned}$$

$$B_1 = \frac{2}{5}(0 + 1 \sin\left(\frac{2\pi}{5}\right) + 2 \sin\left(\frac{2\pi(2)}{5}\right) + 3 \sin\left(\frac{2\pi(3)}{5}\right) + 1 \sin\left(\frac{2\pi(4)}{5}\right))$$

$$\approx -0.235$$

So the best trigonometric fit of order  $p = 1$  for the time-series concerned is

$$f(t) = \frac{11}{5} + 0.229 \cos\left(\frac{2\pi}{5}t\right) - 0.235 \sin\left(\frac{2\pi}{5}t\right)$$

□

Short Exercise: Find an improved approximation with order  $p = 2$ . What happens if  $p = 3$ ? <sup>6</sup>

---

<sup>6</sup>For  $p = 2$ , the best-fit coefficients are

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ A_2 \\ B_2 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 \cos\left(\frac{2\pi}{5}\right) & 2 \cos\left(\frac{2\pi(2)}{5}\right) & 2 \cos\left(\frac{2\pi(3)}{5}\right) & 2 \cos\left(\frac{2\pi(4)}{5}\right) \\ 0 & 2 \sin\left(\frac{2\pi}{5}\right) & 2 \sin\left(\frac{2\pi(2)}{5}\right) & 2 \sin\left(\frac{2\pi(3)}{5}\right) & 2 \sin\left(\frac{2\pi(4)}{5}\right) \\ 2 & 2 \cos\left(\frac{2\pi(2)}{5}\right) & 2 \cos\left(\frac{2\pi(2)(2)}{5}\right) & 2 \cos\left(\frac{2\pi(2)(3)}{5}\right) & 2 \cos\left(\frac{2\pi(2)(4)}{5}\right) \\ 0 & 2 \sin\left(\frac{2\pi(2)}{5}\right) & 2 \sin\left(\frac{2\pi(2)(2)}{5}\right) & 2 \sin\left(\frac{2\pi(2)(3)}{5}\right) & 2 \sin\left(\frac{2\pi(2)(4)}{5}\right) \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \end{bmatrix}$$

It is not hard to see that  $C_0, A_1, B_1$  will be the same and

$$\begin{aligned} A_2 &= \frac{2}{5}(4 + 1 \cos\left(\frac{2\pi(2)}{5}\right) + 2 \cos\left(\frac{2\pi(2)(2)}{5}\right) + 3 \cos\left(\frac{2\pi(2)(3)}{5}\right) + 1 \cos\left(\frac{2\pi(2)(4)}{5}\right)) \\ &\approx 1.571 \\ B_2 &= \frac{2}{5}(0 + 1 \sin\left(\frac{2\pi(2)}{5}\right) + 2 \sin\left(\frac{2\pi(2)(2)}{5}\right) + 3 \sin\left(\frac{2\pi(2)(3)}{5}\right) + 1 \sin\left(\frac{2\pi(2)(4)}{5}\right)) \\ &\approx 0.380 \end{aligned}$$

Hence the new approximation will be  $f(t) = \frac{11}{5} + 0.229 \cos\left(\frac{2\pi}{5}t\right) - 0.235 \sin\left(\frac{2\pi}{5}t\right) + 1.571 \cos\left(\frac{2\pi(2)}{5}t\right) + 0.380 \sin\left(\frac{2\pi(2)}{5}t\right)$ . When  $p$  is increased to 3, the matrix  $G$  involved in the best-fit formula becomes

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & \cos\left(\frac{2\pi}{5}\right) & \sin\left(\frac{2\pi}{5}\right) & \cos\left(\frac{2\pi(2)}{5}\right) & \sin\left(\frac{2\pi(2)}{5}\right) & \cos\left(\frac{2\pi(3)}{5}\right) & \sin\left(\frac{2\pi(3)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(2)}{5}\right) & \sin\left(\frac{2\pi(2)}{5}\right) & \cos\left(\frac{2\pi(2)(2)}{5}\right) & \sin\left(\frac{2\pi(2)(2)}{5}\right) & \cos\left(\frac{2\pi(3)(2)}{5}\right) & \sin\left(\frac{2\pi(3)(2)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(3)}{5}\right) & \sin\left(\frac{2\pi(3)}{5}\right) & \cos\left(\frac{2\pi(3)(2)}{5}\right) & \sin\left(\frac{2\pi(3)(2)}{5}\right) & \cos\left(\frac{2\pi(3)(3)}{5}\right) & \sin\left(\frac{2\pi(3)(3)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(4)}{5}\right) & \sin\left(\frac{2\pi(4)}{5}\right) & \cos\left(\frac{2\pi(2)(4)}{5}\right) & \sin\left(\frac{2\pi(2)(4)}{5}\right) & \cos\left(\frac{2\pi(3)(4)}{5}\right) & \sin\left(\frac{2\pi(3)(4)}{5}\right) \end{bmatrix}$$

### 14.1.2 Nyquist Frequency and Real DFT

While we may want to make the approximation by the Fourier basis as good as possible, we have to know how high the order  $p$  needs to be. On the other hand, as we can see in the last example, if  $p$  is set too large, the best-fit formula will become problematic and the approximation will contain duplicated terms, in the sense that they take equal values (or with an opposite sign) at all sampling points. For example, if  $k + l = N$ , then for the integer time steps  $t = 0, 1, \dots, N - 1$

$$\begin{aligned}\cos\left(\frac{2\pi l}{N}t\right) &= \cos\left(\frac{2\pi(N-k)}{N}t\right) \\ &= \cos\left(2\pi t - \frac{2\pi k}{N}t\right) \\ &= \cos\left(\frac{2\pi k}{N}t\right)\end{aligned}$$

and a routine computation will show that

$$G^T G = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{5}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{5}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{2} & 0 & \frac{5}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{2} & 0 & -\frac{5}{2} & 0 \\ 0 & 0 & 0 & \frac{5}{2} & 0 & \frac{5}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{5}{2} & 0 & \frac{5}{2} & 0 \end{bmatrix}$$

which is not invertible (the fourth/sixth rows are equal and the fifth/last rows are the negative of each other), so that the formula  $\vec{\beta}_f = (G^T G)^{-1} G^T \vec{d}$  will fail. If we forcefully use the expressions in (14.11) and (14.12) to compute  $A_3$  and  $B_3$  we will get

$$\begin{aligned}A_3 &= \frac{2}{5}(4 + 1 \cos\left(\frac{2\pi(3)}{5}\right) + 2 \cos\left(\frac{2\pi(3)(2)}{5}\right) + 3 \cos\left(\frac{2\pi(3)(3)}{5}\right) + 1 \cos\left(\frac{2\pi(3)(4)}{5}\right)) \\ &= A_2 \\ B_3 &= \frac{2}{5}(0 + 1 \sin\left(\frac{2\pi(3)}{5}\right) + 2 \sin\left(\frac{2\pi(3)(2)}{5}\right) + 3 \sin\left(\frac{2\pi(3)(3)}{5}\right) + 1 \sin\left(\frac{2\pi(3)(4)}{5}\right)) \\ &= -B_2\end{aligned}$$

So  $A_2$  ( $B_2$ ) and  $A_3$  ( $B_3$ ) carry the same information and one of them will be redundant.

so the two cosine waves, despite having different frequencies, coincide at every data point and will be indistinguishable from the time-series. The same problem arises similarly for the sine terms. The condition  $k + l = N$  above hints that the maximum value of  $p$  should be  $N/2$ . This corresponds to an angular frequency of  $\omega_{\text{Ny}} = \frac{2\pi}{N}(\frac{N}{2}) = \pi$ , which is known as the **Nyquist frequency**. Consequentially, we have the **Nyquist Sampling Theorem** as follows.

**Theorem 14.1.1** (Nyquist Sampling Theorem). For an evenly spaced time-series that has a time step of  $\Delta t = 1$ , any sinusoidal wave with an angular frequency exceeding the Nyquist frequency  $\omega > \omega_{\text{Ny}} = \pi$  cannot be properly detected.

So it means that the highest resolvable frequency in the time-series is the Nyquist frequency  $\omega_{\text{Ny}} = \pi$ , or in other words, the minimum period length has to cover at least two time steps. As a result, we only need the sine/cosine terms up to the order  $\lfloor N/2 \rfloor$ . We have just shown a part of the theorem that if  $k + l = N$ , then the sinusoidal series of a higher order  $l = \lfloor N/2 \rfloor + 1, \dots, N$  will be redundant in the derivation above. We will complete the theorem by verifying that the sine/cosine terms of an order  $l = N + 1, N + 2, \dots$  beyond will also lead to duplicated modes: let  $l = k + qN$  where  $q$  is a positive integer, then

$$\begin{aligned}\cos\left(\frac{2\pi l}{N}t\right) &= \cos\left(\frac{2\pi(k + qN)}{N}t\right) \\ &= \cos\left(\frac{2\pi k}{N}t + 2\pi qt\right) \\ &= \cos\left(\frac{2\pi k}{N}t\right)\end{aligned}$$

for all time steps  $t = 0, 1, \dots, N - 1$ . Again, it is similar for the sines. Another perspective to see the problem is that, if the "ground truth" function to be approximated by DFT indeed contains some sinusoidal component with a frequency higher than the Nyquist frequency, then its signal will "spill" into a corresponding lower frequency. Again, using cosine and the case  $k + l = N$  as an illustration, if  $k$  represents the lower frequency mode and  $l$  represents the

higher one (assumed to have a coefficient of  $a_l$ ), then (14.11) will yield

$$\begin{aligned}
 A_k &= \frac{2}{N} \sum_{t=0}^{N-1} (a_l \cos\left(\frac{2\pi l t}{N}\right)) \cos\left(\frac{2\pi k t}{N}\right) \\
 &= \frac{2}{N} \sum_{t=0}^{N-1} a_l \cos\left(\frac{2\pi(N-k)t}{N}\right) \cos\left(\frac{2\pi k t}{N}\right) \\
 &= \frac{2}{N} \sum_{t=0}^{N-1} a_l \cos\left(2\pi t - \frac{2\pi k}{N} t\right) \cos\left(\frac{2\pi k t}{N}\right) \\
 &= \frac{2}{N} \sum_{t=0}^{N-1} a_l \cos\left(\frac{2\pi k}{N} t\right) \cos\left(\frac{2\pi k t}{N}\right) \\
 &= \frac{2}{N} \sum_{t=0}^{N-1} a_l \cos^2\left(\frac{2\pi k t}{N}\right) \\
 &= \frac{2}{N} (a_l \left(\frac{N}{2}\right)) = a_l
 \end{aligned} \tag{Equation (14.9)}$$

Hence any signal with a frequency higher than the Nyquist frequency will contribute to the respective DFT frequency and contaminate it.

Now we can formally derive the (*real*) **Discrete Fourier Transform** for a time-series which is given by (14.10), (14.11), and (14.12). In the last part, we have already worked with an odd  $N$  (see Example 14.1.1) where the maximum resolvable degree will be  $p = \frac{N}{2} - 1$ . When  $N$  is even, then we have frequencies from zero to  $k = N/2$ . Notice that the constant term of zero frequency contributes a single parameter, every other frequency contribute two coefficients via a pair of sine and cosine, and the maximum frequency  $k_{Ny} = N/2$  only give rises to one coefficient from the cosine series which takes the form of alternating  $(1, -1, 1, -1, \dots, 1, -1)$  (the corresponding diagonal entry in the  $(G^T G)^{-1}$  term of the least-square formula will be  $1/N$  instead of  $2/N$ ), as the sine function will be always zero at the Nyquist frequency. In both cases, there can be at most  $N$  sinusoidal curves for fitting  $N$  data points and  $G$  will be square. Since the amount of parameters is the same as the number of data, it becomes an interpolation that passes through all the given data points. By convention, we will omit the factor of  $1/N$  in the computation of DFT.

**Example 14.1.2.** Find the Discrete Fourier Transform of the time series  $(1, 2, 1, -1, 3, 0, 2, -2)$ .

*Solution.* According to Theorem 14.1.1, the highest resolvable degree will be  $N/2 = 4$ . Its DFT is then given by

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ A_2 \\ B_2 \\ A_3 \\ B_3 \\ A_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 2 & 2\cos\left(\frac{2\pi}{8}\right) & 2\cos\left(\frac{2\pi(2)}{8}\right) & 2\cos\left(\frac{2\pi(3)}{8}\right) & \cdots & 2\cos\left(\frac{2\pi(6)}{8}\right) & 2\cos\left(\frac{2\pi(7)}{8}\right) \\ 0 & 2\sin\left(\frac{2\pi}{8}\right) & 2\sin\left(\frac{2\pi(2)}{8}\right) & 2\sin\left(\frac{2\pi(3)}{8}\right) & \cdots & 2\sin\left(\frac{2\pi(6)}{8}\right) & 2\sin\left(\frac{2\pi(7)}{8}\right) \\ 2 & 2\cos\left(\frac{2\pi(2)}{8}\right) & 2\cos\left(\frac{2\pi(2)(2)}{8}\right) & 2\cos\left(\frac{2\pi(2)(3)}{8}\right) & \cdots & 2\cos\left(\frac{2\pi(2)(6)}{8}\right) & 2\cos\left(\frac{2\pi(2)(7)}{8}\right) \\ 0 & 2\sin\left(\frac{2\pi(2)}{8}\right) & 2\sin\left(\frac{2\pi(2)(2)}{8}\right) & 2\sin\left(\frac{2\pi(2)(3)}{8}\right) & \cdots & 2\sin\left(\frac{2\pi(2)(6)}{8}\right) & 2\sin\left(\frac{2\pi(2)(7)}{8}\right) \\ 2 & 2\cos\left(\frac{2\pi(3)}{8}\right) & 2\cos\left(\frac{2\pi(3)(2)}{8}\right) & 2\cos\left(\frac{2\pi(3)(3)}{8}\right) & \cdots & 2\cos\left(\frac{2\pi(3)(6)}{8}\right) & 2\cos\left(\frac{2\pi(3)(7)}{8}\right) \\ 0 & 2\sin\left(\frac{2\pi(3)}{8}\right) & 2\sin\left(\frac{2\pi(3)(2)}{8}\right) & 2\sin\left(\frac{2\pi(3)(3)}{8}\right) & \cdots & 2\sin\left(\frac{2\pi(3)(6)}{8}\right) & 2\sin\left(\frac{2\pi(3)(7)}{8}\right) \\ 1 & -1 & 1 & -1 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \\ f(5) \\ f(6) \\ f(7) \end{bmatrix}$$

A direct computation then gives

$$(C_0, A_1, B_1, A_2, B_2, A_3, B_3, A_4)^T = (6, -2.586, 2.243, 2, 10, -5.414, 6.243, 8)^T$$

□

### 14.1.3 Complex DFT

The real DFT approach above has a drawback of using sine-cosine pairs for computation. In Section 8.1, we have learnt the power of complex exponentials to simultaneously represent sines and cosines through Euler's formula, which has also been exploited to derive the relationships between the sinusoidal functions when we are developing the DFT prototype. Therefore, it is incentive to explore the possibility of using complex exponentials to define Discrete Fourier Transform. This has the benefits of simplicity and also convenience when programming.

Continuing from the ideas built in the last section, we propose an interpolation scheme which uses  $\exp(i(2\pi k/N)t)$ , for a time-series with  $N$  data, evenly spaced by a time step of  $\Delta t = 1$ . The range of  $k$  will be from  $-\frac{N}{2}, -(\frac{N}{2}-1), \dots, 0, \frac{N}{2}-1$  for even  $N$  and  $-\frac{N-1}{2}, -(\frac{N-1}{2}-1), \dots, 0, \frac{N-1}{2}$  for odd  $N$ . Both ranges ensure

that the total number of complex exponentials used in the fitting process is exactly  $N$ . Each pair of  $k$  and  $-k$ , i.e.  $\exp(i(2\pi k/N)t)$  and  $\exp(-i(2\pi k/N)t)$  in combination, gives rise to  $\cos((2\pi k/N)t)$  and  $\sin((2\pi k/N)t)$  by Properties 8.1.7, and so we can expect the correspondence between the real and complex version of Discrete Fourier Transform.

Now, the matrix  $G$  will take the form of

$$G = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & \exp\left(i\frac{2\pi}{N}\right) & \exp\left(i\frac{2\pi(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})}{N}\right) & \cdots & \exp\left(-i\frac{2\pi}{N}\right) \\ 1 & \exp\left(i\frac{2\pi(2)}{N}\right) & \exp\left(i\frac{2\pi(2)(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)(2)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})(2)}{N}\right) & \cdots & \exp\left(-i\frac{2\pi(2)}{N}\right) \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \exp\left(i\frac{2\pi(N-1)}{N}\right) & \exp\left(i\frac{2\pi(2)(N-1)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)(N-1)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})(N-1)}{N}\right) & \cdots & \exp\left(-i\frac{2\pi(N-1)}{N}\right) \end{bmatrix}$$

for even  $N$ , where each column representing frequencies that have  $k = 0, 1, \dots, \frac{N}{2} - 1, -\frac{N}{2}, \dots, -1$ . This is a common convention where we starts from  $k = 0$ , and increases to the largest positive  $k = \frac{N}{2} - 1$ , then flips the sign and resumes from the most negative  $k = -\frac{N}{2}$ , goes all the way back to  $k = -1$ . For odd  $N$ , the matrix  $G$  is essentially the same, except the  $k$  is replaced by the appropriate range of integers and in particular the flipping leads to  $k = -\frac{N-1}{2}$ .

The entries of  $G^*G$  in the formula from Theorem 13.1.3 are then the complex dot products between the series of complex exponentials appearing as the column vectors of  $G$ . The orthogonality relation between different column vectors of  $G$  is very easy to find. The procedure is similar to what we have done when proving the orthogonality for the real case, but even less tedious. The key point is to write the complex dot product over any pair of two columns as a geometric sequence, that when the values of  $k$  are different, will be evaluated to zero. The readers are invited to verify this result, in addition to the fact that the complex dot product between any such a column vector and itself is  $N$ .<sup>7</sup> Thus, the expression  $G^*G$  is just  $N$  times the identity  $I$ , and  $(G^*G)^{-1} = \frac{1}{N}I$ .

<sup>7</sup>When  $k \neq l$ , we have

$$\sum_{t=0}^{N-1} \exp(i(2\pi k/N)t) \overline{\exp(i(2\pi l/N)t)} = \sum_{t=0}^{N-1} \exp(i(2\pi k/N)t) \exp(-i(2\pi l/N)t)$$

Now we denote the new, complex best-fit parameters, or coefficients by  $C_k$ . Subsequently,

$$\vec{\beta}_f = (G^* G)^{-1} G^* \vec{d} = \left(\frac{1}{N} I\right) G^* \vec{d}$$

$$\begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_{N/2-1} \\ C_{-N/2} \\ \vdots \\ C_{-1} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \dots \\ 1 & \exp\left(-i\frac{2\pi}{N}\right) & \exp\left(-i\frac{2\pi(2)}{N}\right) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \exp\left(-i\frac{2\pi(N/2-1)}{N}\right) & \exp\left(-i\frac{2\pi(N/2-1)(2)}{N}\right) & \dots \\ 1 & \exp\left(-i\frac{2\pi(-N/2)}{N}\right) & \exp\left(-i\frac{2\pi(-N/2)(2)}{N}\right) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \exp\left(-i\frac{2\pi(-1)}{N}\right) & \exp\left(-i\frac{2\pi(-1)(2)}{N}\right) & \dots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix}$$

Again this is for even  $N$ , and we ought to replace the indices for coefficients and complex exponentials appropriately for odd  $N$ . Again, the  $1/N$  factor will be ignored. We now conclude the method of (**complex**) *Discrete Fourier Transform* in a compact way as follows.

**Definition 14.1.2.** The coefficients, or amplitudes, of the DFT in complex form are computed by

$$C_k = \sum_{t=0}^{N-1} f(t) \exp\left(-i\frac{2\pi k}{N} t\right)$$

$$\begin{aligned} &= \sum_{t=0}^{N-1} \exp(i(2\pi(k-l)/N)t) \\ &= \frac{1 - \exp(i(2\pi(k-l)/N))^N}{1 - \exp(i(2\pi(k-l)/N))} \\ &= \frac{1 - \exp(i(2\pi(k-l)))}{1 - \exp(i(2\pi(k-l)/N))} \\ &= \frac{1 - 1}{1 - \exp(i(2\pi(k-l)/N))} = 0 \end{aligned}$$

If  $k = l$ , then  $\exp(i(2\pi k/N)t)\overline{\exp(i(2\pi l/N)t)} = \exp(i(2\pi k/N)t) \exp(-i(2\pi k/N)t) = 1$ , and the sum will be  $N$ .

where  $N$  is the number of data.  $k$  are all integers ranging from  $[-\frac{N}{2}, \frac{N}{2} - 1]$  for even  $N$ , and  $[-\frac{N-1}{2}, \frac{N-1}{2}]$  for odd  $N$ . For a time step  $\Delta t$  different from 1, the appropriate expression is

$$C_k = \sum_{s=0}^{N-1} f(s\Delta t) \exp\left(-i \frac{2\pi k}{n\Delta t} s\right)$$

We will sometimes denote the series of  $C_k$  by  $F(k)$  or  $\hat{f}(k)$ .

The negative sign inside the complex exponentials in the formula comes from the conjugate transpose required to produce  $G^*$ . The relation between  $C_k$  and the parameters  $A_k, B_k$  in the real counterpart are inferred from Properties 8.1.7 and comparing the expression for the real case<sup>8</sup>, that is

$$\begin{aligned} A_k &= \frac{C_k + C_{-k}}{N} \\ B_k &= -\frac{(C_k - C_{-k})}{Ni} \end{aligned}$$

for  $k \neq 0$ . If  $N$  is even, then we define  $C_N = 0$  for convenience. Moreover, since sine is an odd function, and cosine is an even function,  $\text{Re}(C_k) = \text{Re}(C_{-k})$ , and  $\text{Im}(C_k) = -\text{Im}(C_{-k})$  if the input signal  $f(t)$  is real-valued, or in other words,  $C_k$  and  $C_{-k}$  are a pair of complex conjugates. And so an alternative relationship between the real and complex DFT is

<sup>8</sup>

$$\begin{aligned} \frac{C_k + C_{-k}}{N} &= \frac{1}{N} \left( \sum_{t=0}^{N-1} f(t) \exp\left(-i \frac{2\pi k}{N} t\right) + \sum_{t=0}^{N-1} f(t) \exp\left(-i \frac{2\pi(-k)}{N} t\right) \right) \\ &= \frac{1}{N} \left( \sum_{t=0}^{N-1} f(t) \left( \exp\left(-i \frac{2\pi k}{N} t\right) + \exp\left(i \frac{2\pi k}{N} t\right) \right) \right) \\ &= \frac{2}{N} \sum_{t=0}^{N-1} f(t) \cos\left(\frac{2\pi k}{N} t\right) \quad (\text{Properties 8.1.7}) \end{aligned}$$

which is just equal to  $A_k$  in Equation (14.11). The derivation is similar for  $B_k$ .

**Properties 14.1.3.** Given a real time-series, the amplitudes  $A_k$ ,  $B_k$  in real DFT, and  $C_k$  in complex DFT satisfy the relations

$$A_k = 2 \operatorname{Re}(C_k)/N$$

$$B_k = -2 \operatorname{Im}(C_k)/N$$

for any  $0 \neq |k| < N$ . Meanwhile,  $C_0$  is the same in both types of DFT and when  $N$  is even,  $A_N = C_N/N$ . The  $1/N$  factor is optional and depends on if any convention is used.

**Example 14.1.3.** Find the complex DFT for the time-series in example 14.1.1.

*Solution.* Using the formula in Definition 14.1.2, we have

$$C_0 = 4 + 1 + 2 + 3 + 1 = 11$$

$$C_1 = 4 + 1 \exp\left(-i\frac{2\pi}{5}\right) + 2 \exp\left(-i\frac{2\pi(2)}{5}\right) + \\ 3 \exp\left(-i\frac{2\pi(3)}{5}\right) + 1 \exp\left(-i\frac{2\pi(4)}{5}\right) = 0.573 + 0.588i$$

$$C_2 = 4 + 1 \exp\left(-i\frac{2\pi(2)}{5}\right) + 2 \exp\left(-i\frac{2\pi(2)(2)}{5}\right) \\ + 3 \exp\left(-i\frac{2\pi(2)(3)}{5}\right) + 1 \exp\left(-i\frac{2\pi(2)(4)}{5}\right) = 3.927 - 0.951i$$

Either by the aforementioned property or a direct computation,  $C_{-1}$  and  $C_{-2}$  are seen to be the complex conjugates of  $C_1$  and  $C_2$  respectively.  $\square$

Short Exercise: Check if Properties 14.1.3 is true in this example.<sup>9</sup>

---

<sup>9</sup>We will only check  $A_1$  and  $B_1$  and leave  $A_2$  and  $B_2$  to the readers.  $2 \operatorname{Re}(C_1)/N = 2(0.573)/5 \approx 0.229 = A_1$ ,  $-2 \operatorname{Im}(C_1)/N = -2(0.588)/5 \approx -0.235 = B_1$ .

#### 14.1.4 Inverse DFT

During the derivation of complex DFT, we have found that  $G^*G = I$  (omitting the factor  $N$ ) and thus the square matrix  $G$  is unitary by Definition 10.4.1. This further implies that  $G^* = G^{-1}$  (and thus  $(G^*)^{-1} = G$ ) is invertible. And since the complex DFT coefficients are given by  $C_k = G^*\vec{d}$ , we can undo the DFT and recover the original time-series by multiplying to the left with the inverse  $(G^*)^{-1}$  so that  $\vec{d} = (G^*)^{-1}C_k = GC_k$ . This means that

$$\begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & \exp\left(i \frac{2\pi}{N}\right) & \cdots & \exp\left(i \frac{2\pi(N/2-1)}{N}\right) & \exp\left(i \frac{2\pi(-N/2)}{N}\right) & \cdots & \exp\left(i \frac{2\pi(-1)}{N}\right) \\ 1 & \exp\left(i \frac{2\pi(2)}{N}\right) & \cdots & \exp\left(i \frac{2\pi(N/2-1)(2)}{N}\right) & \exp\left(i \frac{2\pi(-N/2)(2)}{N}\right) & \cdots & \exp\left(i \frac{2\pi(-1)(2)}{N}\right) \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & & & & & \vdots \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_{N/2-1} \\ C_{-N/2} \\ \vdots \\ C_{-1} \end{bmatrix}$$

for even  $N$ , and can be adapted for odd  $N$  with small tweaks similar to those before. It can also be written in a summation form just like the one in Definition 14.1.2 as below.

**Definition 14.1.4.** The inverse Discrete Fourier Transform is computed by

$$f(t) = \frac{1}{N} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) \exp\left(i \frac{2\pi k}{N} t\right)$$

Sometimes we use the operator symbol  $F^{-1}$  to denote the inverse DFT operation.

This can also be verified by a direct substitution. Plugging in the formula from Definition 14.1.2 (with a dummy variable  $t'$ ) into R.H.S., we have

$$= \frac{1}{N} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \left( \sum_{t'=0}^{N-1} f(t') \exp\left(-i \frac{2\pi k}{N} t'\right) \right) \exp\left(i \frac{2\pi k}{n} t\right)$$

where

$$\sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \exp\left(i \frac{2\pi k}{N} (t - t')\right) = \begin{cases} N & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \quad (14.13)$$

The first case should be obvious, while the second case is derived in the footnote.<sup>10</sup> Hence

$$\begin{aligned} \frac{1}{N} \sum_{t'=0}^{N-1} f(t') \left( \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \exp\left(i \frac{2\pi k}{N} (t - t')\right) \right) &= \frac{1}{N} ((0) + \cdots + f(t)(N) + \cdots + (0)) \\ &= f(t) \end{aligned}$$

so the R.H.S. indeed reproduces the original function on L.H.S.

**Example 14.1.4.** Apply the inverse DFT on  $C_{-3} = 1, C_{-2} = 1 + i, C_{-1} = 2 - i, C_0 = 3, C_1 = 2 + i, C_2 = 1 - i$  to retrieve the physical time-series.

*Solution.* The time series  $f(t)$  will have a period of 6 where  $t = 0, 1, 2, 3, 4, 5$ . By Definition 14.1.4, we have

$$f(0) = \frac{1}{N} \sum_{k=-3}^2 F(k) \exp\left(i \frac{2\pi k}{N}(0)\right)$$

---

<sup>10</sup>For an integer  $t - t' = \Delta t \neq 0$ , we have

$$\begin{aligned} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \exp\left(i \frac{2\pi k}{N} \Delta t\right) &= \exp\left(i \frac{2\pi(-\lfloor \frac{N}{2} \rfloor)}{N} (\Delta t)\right) \frac{1 - \exp\left(i \frac{2\pi \Delta t}{N}\right)^N}{1 - \exp\left(i \frac{2\pi \Delta t}{N}\right)} \\ &= \exp\left(i \frac{-2\pi \lfloor \frac{N}{2} \rfloor}{N} (\Delta t)\right) \frac{1 - \exp(i 2\pi \Delta t)}{1 - \exp\left(i \frac{2\pi \Delta t}{N}\right)} \\ &= \exp\left(i \frac{-2\pi \lfloor \frac{N}{2} \rfloor}{N} (\Delta t)\right) \frac{1 - 1}{1 - \exp\left(i \frac{2\pi \Delta t}{N}\right)} = 0 \end{aligned}$$

as a geometric sum that evaluates to zero.  $\Delta t$  cannot be an integer multiple of  $N$  as well but the range of summation prevents this.

$$\begin{aligned}
 &= \frac{1}{6} [C_{-3} \exp\left(i \frac{2\pi(-3)}{6}(0)\right) + C_{-2} \exp\left(i \frac{2\pi(-2)}{6}(0)\right) \\
 &\quad + C_{-1} \exp\left(i \frac{2\pi(-1)}{6}(0)\right) + C_0 \exp\left(i \frac{2\pi(0)}{6}(0)\right) \\
 &\quad + C_1 \exp\left(i \frac{2\pi(1)}{6}(0)\right) + C_2 \exp\left(i \frac{2\pi(2)}{6}(0)\right)] \\
 &= \frac{1}{6} [(1) + (1+i) + (2-i) + 3 + (2+i) + (1-i)] = \frac{10}{6}
 \end{aligned}$$

$$\begin{aligned}
 f(1) &= \frac{1}{N} \sum_{k=-3}^2 F(k) \exp\left(i \frac{2\pi k}{N}(1)\right) \\
 &= \frac{1}{6} [C_{-3} \exp\left(i \frac{2\pi(-3)}{6}(1)\right) + C_{-2} \exp\left(i \frac{2\pi(-2)}{6}(1)\right) \\
 &\quad + C_{-1} \exp\left(i \frac{2\pi(-1)}{6}(1)\right) + C_0 \exp\left(i \frac{2\pi(0)}{6}(1)\right) \\
 &\quad + C_1 \exp\left(i \frac{2\pi(1)}{6}(1)\right) + C_2 \exp\left(i \frac{2\pi(2)}{6}(1)\right)] \\
 &= \frac{1}{6} [(1)(-1) + (1+i)e^{-i \frac{2\pi}{3}} + (2-i)e^{-i \frac{\pi}{3}} + 3 + (2+i)e^{i \frac{\pi}{3}} + (1-i)e^{i \frac{2\pi}{3}}] \\
 &= \frac{1}{2}
 \end{aligned}$$

$$\begin{aligned}
 f(2) &= \frac{1}{N} \sum_{k=-3}^2 F(k) \exp\left(i \frac{2\pi k}{N}(2)\right) \\
 &= \frac{1}{6} [C_{-3} \exp\left(i \frac{2\pi(-3)}{6}(2)\right) + C_{-2} \exp\left(i \frac{2\pi(-2)}{6}(2)\right) \\
 &\quad + C_{-1} \exp\left(i \frac{2\pi(-1)}{6}(2)\right) + C_0 \exp\left(i \frac{2\pi(0)}{6}(2)\right) \\
 &\quad + C_1 \exp\left(i \frac{2\pi(1)}{6}(2)\right) + C_2 \exp\left(i \frac{2\pi(2)}{6}(2)\right)] \\
 &= \frac{1}{6} [(1)(1) + (1+i)e^{-i \frac{4\pi}{3}} + (2-i)e^{-i \frac{2\pi}{3}} + 3 + (2+i)e^{i \frac{2\pi}{3}} + (1-i)e^{i \frac{4\pi}{3}}] \\
 &\approx 0.4107
 \end{aligned}$$

We leave the calculations for the remaining three data points to the readers. They are  $f(3) = 0$ ,  $f(4) \approx 0.7440$ , and  $f(5) = \frac{1}{2}$ .  $\square$

## 14.2 Properties of DFT

### 14.2.1 Power Spectrum and Parseval's Theorem

The complex DFT coefficients actually store the relevant information of sinusoidal signals in the corresponding frequencies, namely the *phase* and *amplitude*. To see this, go back to Properties 14.1.3, where

$$A_k = 2 \operatorname{Re}(C_k)/N$$

$$B_k = -2 \operatorname{Im}(C_k)/N$$

but also, the wave signal of that particular frequency takes the form of

$$A_k \cos\left(\frac{2\pi k}{N}t\right) + B_k \sin\left(\frac{2\pi k}{N}t\right)$$

Plugging in gives

$$\begin{aligned} & \frac{2}{N} (\operatorname{Re}(C_k) \cos\left(\frac{2\pi k}{N}t\right) - \operatorname{Im}(C_k) \sin\left(\frac{2\pi k}{N}t\right)) \\ &= \frac{2}{N} (\hat{A}_k \cos \phi_k \cos\left(\frac{2\pi k}{N}t\right) - \hat{A}_k \sin \phi_k \sin\left(\frac{2\pi k}{N}t\right)) \\ &= \frac{2}{N} \hat{A}_k \cos\left(\frac{2\pi k}{N}t + \phi_k\right) \end{aligned}$$

where we let  $-\pi \leq \phi_k < \pi$  and  $\hat{A}_k \geq 0$  as real-valued quantities, in a way such that  $\operatorname{Re}(C_k) = \hat{A}_k \cos \phi_k$  and  $\operatorname{Im}(C_k) = \hat{A}_k \sin \phi_k$ , and then apply the trigonometric identity  $\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta$ .  $\phi_k$  and  $\hat{A}_k$  will be the phase (relative to a cosine wave) and the amplitude of the signal (again

putting the  $\frac{2}{N}$  factor aside) at the  $k$ -th frequency respectively. The required values of  $\phi_k$  and  $\hat{A}_k$  are then derived according to

$$\begin{aligned}\frac{\operatorname{Re}(C_k)}{\operatorname{Im}(C_k)} &= \frac{\sin \phi_k}{\cos \phi_k} = \tan \phi_k \\ \operatorname{Re}(C_k)^2 + \operatorname{Im}(C_k)^2 &= \hat{A}_k^2 \cos^2 \phi_k + \hat{A}_k^2 \sin^2 \phi_k = \hat{A}_k^2\end{aligned}$$

Thus  $\phi_k = \arctan\left(\frac{\operatorname{Im}(C_k)}{\operatorname{Re}(C_k)}\right)$  and  $\hat{A}_k = \sqrt{\operatorname{Re}(C_k)^2 + \operatorname{Im}(C_k)^2}$ . Recall that from Section 8.1.3 these are exactly the argument and modulus of  $C_k$ . Therefore, by simply looking at the complex DFT coefficient  $C_k$  we can readily extract the phase and amplitude of the corresponding DFT signal. However, in the area of signal processing, we often report the *power* of the signal instead of its amplitude for convenience, which is just the square of the amplitude. It can be easily obtained from

$$|C_k|^2 = C_k \overline{C_k} = C_k C_{-k}$$

due to Equation (8.1) and the fact that  $C_k$  has  $C_{-k}$  as its complex conjugate (if the input time-series is real-valued). The powers over all frequencies are collectively referred to as the ***power spectrum***.

**Example 14.2.1.** Find the phase and power at each frequency bin for the complex DFT computed in Example 14.1.3.

*Solution.* The phase of the zeroth frequency (i.e. constant term) signal is trivially zero, and its power is simply  $C_0^2 = (11)^2 = 121$ . For the first (base) frequency, the phase is

$$\phi_1 = \arctan\left(\frac{\operatorname{Im}(C_1)}{\operatorname{Re}(C_1)}\right) = \arctan\left(\frac{0.588}{0.573}\right) = 0.80 \text{ rad}$$

and the power is

$$\begin{aligned}|C_1|^2 &= |C_{-1}|^2 = C_1 C_{-1} = (0.573 + 0.588i)(0.573 - 0.588i) \\ &= 0.67\end{aligned}$$

Similar for the second frequency, the phase is

$$\phi_2 = \arctan\left(\frac{\text{Im}(C_2)}{\text{Re}(C_2)}\right) = \arctan\left(\frac{-0.951}{3.927}\right) = -0.24 \text{ rad}$$

and the power is

$$\begin{aligned} |C_2|^2 &= |C_{-2}|^2 = C_2 C_{-2} = (3.927 - 0.951i)(3.927 + 0.951i) \\ &= 16.33 \end{aligned}$$

□

The full power spectrum across all frequencies are related to the initial time-series according to the **Parseval's Theorem**.

**Theorem 14.2.1** (Parseval's Theorem). The sum of powers for each DFT coefficients divided by the length  $N$  is equal to the sum of all data values in the (real) time-series squared, i.e.

$$\sum_{t=0}^{N-1} f(t)^2 = \frac{1}{N} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} |F(k)|^2$$

*Proof.* By substituting the form of inverse DFT given in Definition 14.1.4, the L.H.S. becomes

$$\begin{aligned} \sum_{t=0}^{N-1} f(t)^2 &= \sum_{t=0}^{N-1} \left[ \frac{1}{N} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) \exp\left(i \frac{2\pi k}{N} t\right) \left( \frac{1}{N} \sum_{l=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(l) \exp\left(i \frac{2\pi l}{N} t\right) \right) \right] \\ &= \frac{1}{N^2} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) \left( \sum_{t=0}^{N-1} \exp\left(i \frac{2\pi k}{N} t\right) \left( \sum_{l=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \overline{F(l)} \exp\left(-i \frac{2\pi l}{N} t\right) \right) \right) \\ &= \frac{1}{N^2} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) \left( \sum_{l=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \sum_{t=0}^{N-1} \overline{F(l)} \exp\left(i \frac{2\pi(k-l)}{N} t\right) \right) \end{aligned}$$

Similar to (14.13), we have

$$\sum_{t=0}^{N-1} \exp\left(i \frac{2\pi(k-l)}{N} t\right) = \begin{cases} N & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}$$

Thus

$$\begin{aligned} & \frac{1}{N^2} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) \left( \sum_{l=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} \sum_{t=0}^{N-1} \overline{F(l)} \exp\left(i \frac{2\pi(k-l)}{N} t\right) \right) \\ &= \frac{1}{N^2} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} F(k) [(0) + \cdots + \overline{(F(k))}(N) + \cdots + (0)] \\ &= \frac{1}{N^2} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} N F(k) \overline{F(k)} \\ &= \frac{1}{N} \sum_{k=-\lfloor \frac{N}{2} \rfloor}^{\lfloor \frac{N-1}{2} \rfloor} |F(k)|^2 \end{aligned}$$

is equal to the R.H.S.  $\square$

**Example 14.2.2.** Verify the Parseval's Theorem for Examples 14.1.1 and 14.1.3.

*Solution.* The L.H.S. of the formula in Theorem 14.2.1 applied to the time-series of Example 14.1.1, is simply

$$\begin{aligned} \sum_{t=0}^4 f(t)^2 &= f(0)^2 + f(1)^2 + f(2)^2 + f(3)^2 + f(4)^2 \\ &= (4)^2 + (1)^2 + (2)^2 + (3)^2 + (1)^2 = 31 \end{aligned}$$

and from Example 14.1.3, the R.H.S. is

$$\frac{1}{5} \sum_{k=-2}^2 |F(k)|^2 = \frac{1}{5} (|C_{-2}|^2 + |C_{-1}|^2 + |C_0|^2 + |C_1|^2 + |C_2|^2)$$

$$\begin{aligned}
 &= \frac{1}{5}(16.33 + 0.67 + 121 + 0.67 + 16.33) \\
 &= \frac{1}{5}(155) = 31
 \end{aligned}$$

which are indeed equal.  $\square$

### 14.2.2 Convolution Theorem

**Convolution** often appears along with Fourier Transform due to an elegant theorem that will be introduced soon. In the area of Earth Science, as well as Physics and Statistics, convolution commonly has a place in describing the solution of problems, e.g. via *Green's Function*. Now we are going to introduce how convolution in a discrete sense is defined first, and a daily example will be provided as an illustration. It always involves two time-series or functions.

**Definition 14.2.2** (Discrete Convolution). The convolution  $h(t)$  between two discrete time-series  $f(t)$  and  $g(t)$  is written as  $f(t) * g(t)$ , defined by

$$h(t) = f(t) * g(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(t - \tau)$$

if  $\tau$  takes a value so that the index in either  $f$  or  $g$  is out of their range, then the term is treated as zero.

A schematic diagram of convolution is provided as Figure 14.1. Note that the formula for convolution is symmetric such that we can also define it as  $\sum_{\tau=-\infty}^{\infty} f(t - \tau)g(\tau)$ . Now, for instance, if  $f(t)$  is defined from  $t = 0$  and  $t = 4$  and  $g(t)$  is defined from  $t = 0$  and  $t = 6$  so that they have a length of 5 and 7 respectively, then

$$\begin{aligned}
 h(0) &= f(0)g(0) \\
 h(1) &= f(0)g(1 - 0) + f(1)g(1 - 1) = f(0)g(1) + f(1)g(0) \\
 h(4) &= f(0)g(4 - 0) + f(1)g(4 - 1) + f(2)g(4 - 2) \\
 &\quad + f(3)g(4 - 3) + f(4)g(4 - 4)
 \end{aligned}$$

$$= f(0)g(4) + f(1)g(3) + f(2)g(2) + f(3)g(1) + f(4)g(0)$$

and

$$\begin{aligned} h(6) &= f(0)g(6-0) + f(1)g(6-1) + f(2)g(6-2) \\ &\quad + f(3)g(6-3) + f(4)g(6-4) \\ &= f(0)g(6) + f(1)g(5) + f(2)g(4) + f(3)g(3) + f(4)g(2) \\ h(9) &= f(3)g(9-0) + f(4)g(9-1) = f(3)g(6) + f(4)g(5) \\ h(10) &= f(4)g(6) \end{aligned}$$

Moreover,  $h(t) = 0$  for  $t \geq 11$  or  $t < 0$  so the effective length of  $h(t)$  is 11. It is not hard to deduce that the resulting convolution will have a length of  $m+n-1$  if the two input time-series have a length of  $m$  and  $n$ .

**Example 14.2.3.** The probability of Mary getting married within some years can be described by the time-series  $q(t)$ , where

$$q(t) = \begin{cases} 0.2 & 0 \leq t \leq 2 \\ 0.1 & 3 \leq t \leq 6 \\ 0 & t \geq 7 \end{cases}$$

The probability  $r(t)$  that Mary gives birth to a baby some years  $t$  after marriage is

$$r(t) = \begin{cases} 0 & t = 0 \\ 0.15 & 1 \leq t \leq 4 \\ 0.08 & 5 \leq t \leq 9 \\ 0 & t \geq 10 \end{cases}$$

Find the net probability of Mary getting a baby at some years  $t$  from now on, assuming she will only get pregnant after married.

*Solution.* The required probability  $p(t) = q(t) * r(t)$  is actually given by the convolution between  $q(t)$  and  $r(t)$ , which effectively have a length of 7 and 9.

Particularly, taking  $t = 5$  as an example, then we have

$$\begin{aligned}
 p(t=5) &= P(\text{Birth 5 yrs later}|\text{Married now})P(\text{Married now}) \\
 &\quad + P(\text{Birth 4 yrs later}|\text{Married 1 yr later})P(\text{Married 1 yr later}) \\
 &\quad + \dots \\
 &\quad + P(\text{Birth 1 yr later}|\text{Married 4 yrs later})P(\text{Married 4 yrs later}) \\
 &\quad + P(\text{Birth now}|\text{Married 5 yrs later})P(\text{Married 5 yrs later}) \\
 &= q(0)r(5) + q(1)r(4) + q(2)r(3) \\
 &\quad + q(3)r(2) + q(4)r(1) + q(5)r(0) \\
 &= (0.2)(0.08) + (0.2)(0.15) + (0.2)(0.15) \\
 &\quad + (0.1)(0.15) + (0.1)(0.15) + (0.1)(0) \\
 &= 0.106
 \end{aligned}$$

where  $P(A|B)$  is the conditional probability of  $A$  occurring when  $B$  has happened.

□

Short Exercise: Find the chance of Mary getting a baby at 8 years later by convolution.<sup>11</sup>

Since Fourier analysis essentially treats the functions or time-series as periodic, it happens that it is more useful to define the **circular convolution** for two time-series of the same length, where we wrap either one of the input time-series in a cyclic manner (see Figure 14.2).

**Definition 14.2.3** (Circular Convolution). The circular convolution  $h_c(t)$  between two discrete time-series  $f(t)$  and  $g(t)$  that are of the same length  $N$  is another time-series also has a length of  $N$ , defined as

$$h_c(t) = f(t) \circledast g(t) = \sum_{\tau=0}^{N-1} f(\tau)g((t - \tau) \bmod N)$$

---

<sup>11</sup>It is  $q(0)r(8) + q(1)r(7) + q(2)r(6) + q(3)r(5) + q(4)r(4) + q(5)r(3) + q(6)r(2) = (0.2)(0.08) + (0.2)(0.08) + (0.2)(0.08) + (0.1)(0.08) + (0.1)(0.15) + (0.1)(0.15) + (0.1)(0.15) = 0.101$ .

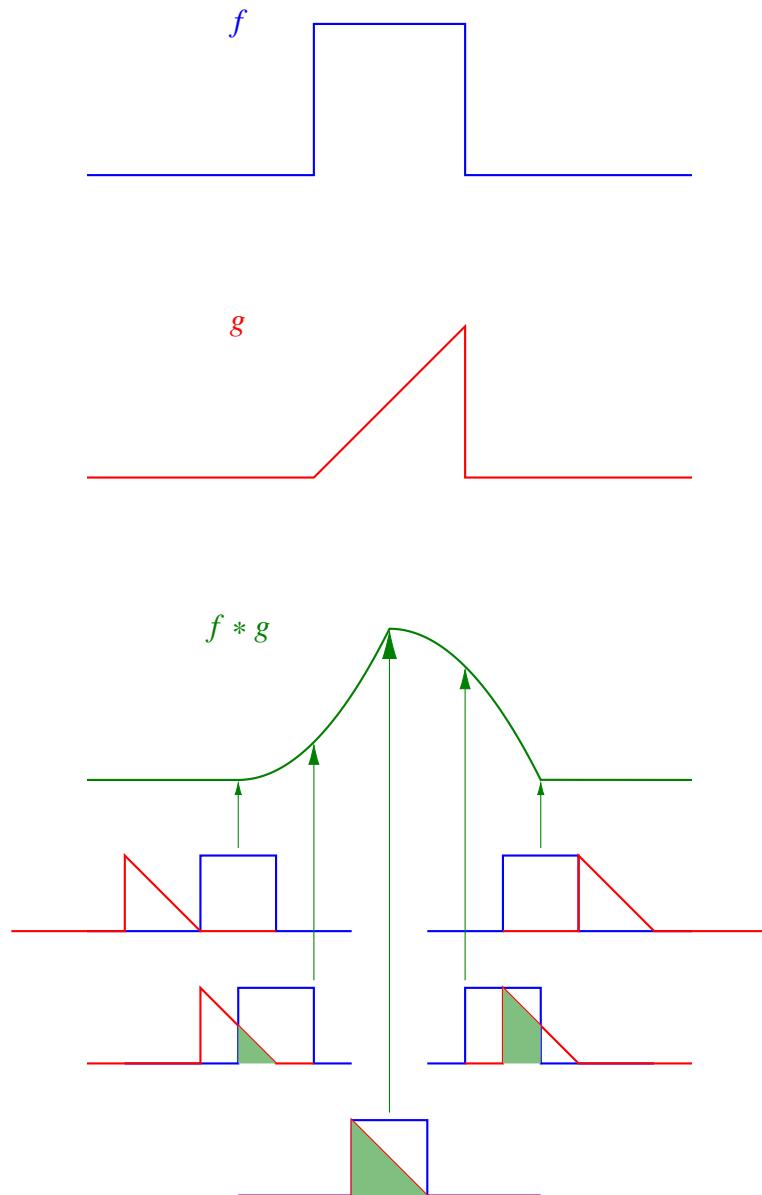


Figure 14.1: An example schematic of convolution.

where  $m \bmod N$  is the *modulo* operation that returns an integer  $0 \leq n < N$  between 0 and  $N - 1$  such that  $m + qN = n$  for some integer  $q$ . In this context

$$(t - \tau) \bmod N = \begin{cases} t - \tau & \text{if } N > t \geq \tau \\ t + N - \tau & \text{if } 0 \leq t < \tau \end{cases}$$

Alternatively, one can manipulate the periodic extension of  $f(t)$  and  $g(t)$  by repeating them indefinitely to produce  $f_x(t)$  and  $g_x(t)$  such that

$$h_c(t) = f(t) \circledast g(t) = \sum_{\tau=-\infty}^{\infty} f_x(\tau)g_x(t - \tau)$$

so that it coincides with Definition 14.2.2 and we only take the values from  $h_c(t)$  within  $0 \leq t < N$ .

Circular convolution is also symmetric so that  $\sum_{\tau=0}^{N-1} f((t - \tau) \bmod N)g(\tau)$  works fine as well. With circular convolution properly defined, we can go ahead to derive the main result, the (***Circular Convolution Theorem***).

**Theorem 14.2.4** (Circular Convolution Theorem). For two time-series  $f(t)$  and  $g(t)$  of the same length  $N$ , denote their DFT by  $\hat{f}(k)$  and  $\hat{g}(k)$ . Then

$$F[f \circledast g](k) = \hat{f}(k)\hat{g}(k)$$

This means that convolution in the physical/time domain is equivalent to component-wise multiplication in the frequency domain.

*Proof.* We start from the L.H.S. and show that it is the same as R.H.S.

$$\begin{aligned} & F[f \circledast g](k) \\ &= \sum_{t=0}^{N-1} \left( \sum_{\tau=0}^{N-1} f(\tau)g((t - \tau) \bmod N) \right) \exp\left(-i\frac{2\pi k}{N}t\right) \\ &= \sum_{\tau=0}^{N-1} \sum_{t=0}^{N-1} \left[ f(\tau)g((t - \tau) \bmod N) \exp\left(-i\frac{2\pi k}{N}(t - \tau)\right) \exp\left(-i\frac{2\pi k}{N}\tau\right) \right] \end{aligned}$$

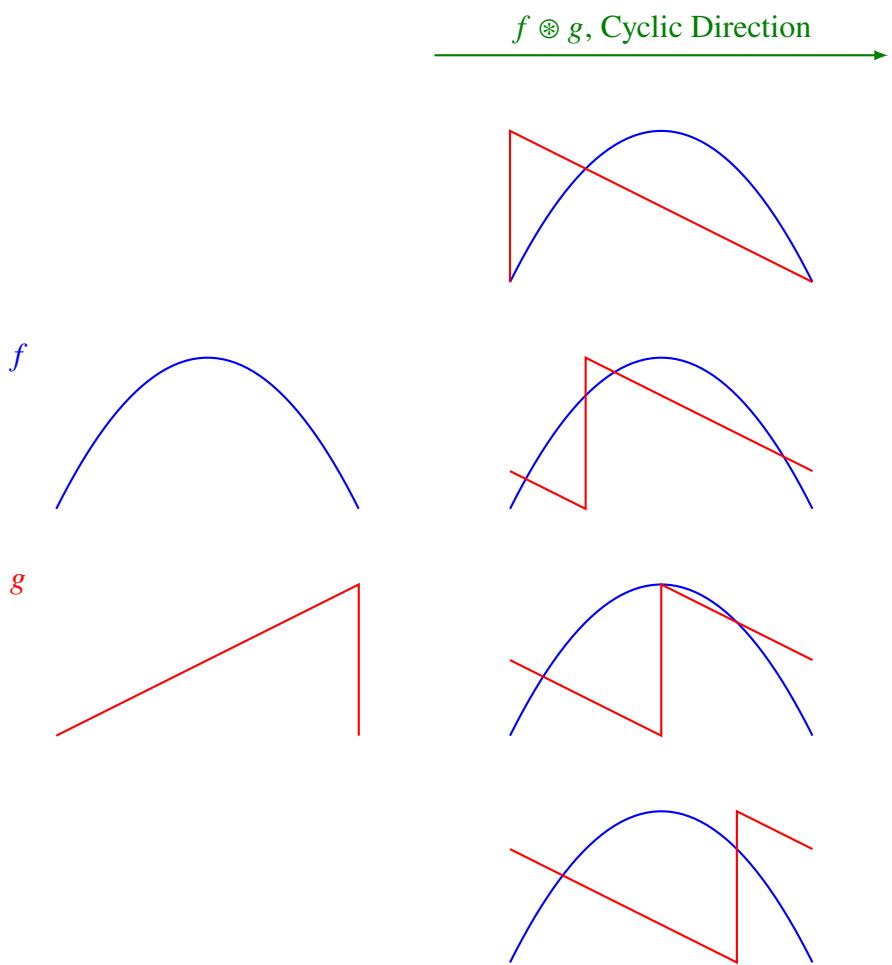


Figure 14.2: An illustration for circular convolution.

$$\begin{aligned}
 &= \sum_{\tau=0}^{N-1} \sum_{t'=-\tau}^{N-1-\tau} f(\tau)g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) \exp\left(-i\frac{2\pi k}{N}\tau\right) \\
 &= \sum_{\tau=0}^{N-1} f(\tau) \left( \sum_{t'=-\tau}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) \right) \exp\left(-i\frac{2\pi k}{N}\tau\right)
 \end{aligned}$$

where we have made a change of variable from  $t$  to  $t' = t - \tau$  with  $\tau$  fixed. Note that the bracket term

$$\begin{aligned}
 &\sum_{t'=-\tau}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) \\
 &= \sum_{t'=-\tau}^{-1} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) + \sum_{t'=0}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) \\
 &= \sum_{t'=N-\tau}^{N-1} g((t' - N) \bmod N) \exp\left(-i\frac{2\pi k}{N}(t' - N)\right) \\
 &\quad + \sum_{t'=0}^{N-\tau-1} g(t') \exp\left(-i\frac{2\pi k}{N}t'\right) \\
 &= \sum_{t'=N-\tau}^{N-1} g(t') \exp\left(-i\frac{2\pi k}{N}t'\right) + \sum_{t'=0}^{N-\tau-1} g(t') \exp\left(-i\frac{2\pi k}{N}t'\right) \\
 &= \sum_{t'=0}^{N-1} g(t') \exp\left(-i\frac{2\pi k}{N}t'\right) = \hat{g}(k)
 \end{aligned}$$

where from the second line to third line we replace  $t'$  by  $t' - N$  in the first term and note that  $(t' - N) \bmod N = t'$ ,  $\exp\left(-i\frac{2\pi k}{N}(t' - N)\right) = \exp\left(-i\left(\frac{2\pi k}{N}t' - 2\pi k\right)\right) = \exp\left(-i\frac{2\pi k}{N}t'\right)$ . In the end we simply invoke Definition 14.1.2. Subsequently,

$$\begin{aligned}
 &\sum_{\tau=0}^{N-1} f(\tau) \left( \sum_{t'=-\tau}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right) \right) \exp\left(-i\frac{2\pi k}{N}\tau\right) \\
 &= \sum_{\tau=0}^{N-1} f(\tau) \hat{g}(k) \exp\left(-i\frac{2\pi k}{N}\tau\right)
 \end{aligned}$$

$$= \left( \sum_{\tau=0}^{N-1} f(\tau) \exp\left(-i\frac{2\pi k}{N}\tau\right) \right) \hat{g}(k) = \hat{f}(k)\hat{g}(k)$$

where we use Definition 14.1.2 again and the desired equality is established.  $\square$

**Example 14.2.4.** Let  $f(t)$  be the time-series in Example 14.1.1, and  $g(t)$  be another time series with  $g(0) = 1, g(1) = 4, g(2) = 0, g(3) = 2, g(4) = -1$ . Verify the Circular Convolution Theorem for these two time-series.

*Solution.*  $\hat{f}(k)$  has been computed in Example 14.1.3 as  $(11, 0.573 + 0.588i, 3.927 - 0.951i, 3.927 + 0.951, 0.573 - 0.588i)$ . Meanwhile,

$$\begin{aligned}\hat{g}(0) &= 1 + 4 + 0 + 2 + (-1) = 6 \\ \hat{g}(1) &= 1 + 4 \exp\left(-i\frac{2\pi}{5}\right) + 0 \exp\left(-i\frac{2\pi(2)}{5}\right) \\ &\quad + 2 \exp\left(-i\frac{2\pi(3)}{5}\right) + (-1) \exp\left(-i\frac{2\pi(4)}{5}\right) = 0.309 - 3.580i \\ \hat{g}(2) &= 1 + 4 \exp\left(-i\frac{2\pi(2)}{5}\right) + 0 \exp\left(-i\frac{2\pi(2)(2)}{5}\right) \\ &\quad + 2 \exp\left(-i\frac{2\pi(2)(3)}{5}\right) + (-1) \exp\left(-i\frac{2\pi(2)(4)}{5}\right) = -0.809 - 4.841i\end{aligned}$$

and hence  $\hat{g}(k) = (6, 0.309 - 3.580i, -0.809 - 4.841i, -0.809 + 4.841i, 0.309 + 3.580i)$ . The circular convolution  $f(t) \circledast g(t)$  is found by

$$\begin{aligned}(f \circledast g)(0) &= f(0)g(0) + f(1)g(4) + f(2)g(3) + f(3)g(2) + f(4)g(1) \\ &= (4)(1) + (1)(-1) + (2)(2) + (3)(0) + (1)(4) = 11 \\ (f \circledast g)(1) &= f(0)g(1) + f(1)g(0) + f(2)g(4) + f(3)g(3) + f(4)g(2) \\ &= (4)(4) + (1)(1) + (2)(-1) + (3)(2) + (1)(0) = 21 \\ (f \circledast g)(2) &= f(0)g(2) + f(1)g(1) + f(2)g(0) + f(3)g(4) + f(4)g(3) \\ &= (4)(0) + (1)(4) + (2)(1) + (3)(-1) + (1)(2) = 5\end{aligned}$$

We leave to the readers to obtain  $(f \otimes g)(3) = 18$  and  $(f \otimes g)(4) = 11$ . Thus  $(f \otimes g)(t) = (11, 21, 5, 18, 11)$  and by Definition 14.1.2 its DFT is

$$\begin{aligned} (\widehat{f \otimes g})(0) &= 11 + 21 + 5 + 18 + 11 = 66 \\ (\widehat{f \otimes g})(1) &= 11 + 21 \exp\left(-i\frac{2\pi}{5}\right) + 5 \exp\left(-i\frac{2\pi(2)}{5}\right) \\ &\quad + 18 \exp\left(-i\frac{2\pi(3)}{5}\right) + 11 \exp\left(-i\frac{2\pi(4)}{5}\right) = 2.281 - 1.869i \\ (\widehat{f \otimes g})(2) &= 11 + 21 \exp\left(-i\frac{2\pi(2)}{5}\right) + 5 \exp\left(-i\frac{2\pi(2)(2)}{5}\right) \\ &\quad + 18 \exp\left(-i\frac{2\pi(2)(3)}{5}\right) + 11 \exp\left(-i\frac{2\pi(2)(4)}{5}\right) \\ &= -7.781 - 18.242i \end{aligned}$$

Therefore, we can verify that

$$\begin{aligned} F[f \otimes g](k) &= (66, 2.281 - 1.869i, -7.781 - 18.242i, -7.781 + 18.242i, 2.281 + 1.869i) \\ &= (11, 0.573 + 0.588i, 3.927 - 0.951i, 3.927 + 0.951i, 0.573 - 0.588i) \\ &\odot (6, 0.309 - 3.580i, -0.809 - 4.841i, -0.809 + 4.841i, 0.309 + 3.580i) \\ &= \hat{f}(k)\hat{g}(k) \end{aligned}$$

Theorem 14.2.4 holds, where  $\odot$  represents component-wise multiplication.  $\square$

### 14.3 Fast Fourier Transform (FFT)

The naive way to compute DFT following Definition 14.1.2 is essentially evaluating the matrix product  $G^* \vec{d}$  aforementioned. If there are  $N$  data points then  $G^*$  will be an  $n \times n$  matrix (and  $\vec{d}$  will be a vector of length  $N$ ) and the calculation will involve  $N^2$  multiplications and  $N(N - 1)$  additions, ultimately leading to a *time complexity* of  $O(N^2)$ . It means that as the number of data  $N$  goes up, the amount of computation required increases quadratically, which is

quite inefficient. Therefore, people have been developing alternative methods to compute DFT more efficiently, and these algorithms are collectively known as ***Fast Fourier Transform***. The first and the most famous one among them is the ***Radix-2 Algorithm***, which utilizes the strategy of *bisection* to cut the calculation into halves recursively. To understand how it works, first notice that we can rewrite the matrix  $G^*$  into a nicer form as

$$G^* = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_N & \omega_N^2 & \omega_N^3 & \cdots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \omega_N^6 & \cdots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \omega_N^{3(N-1)} & \cdots & \omega_N^{(N-1)^2} \end{bmatrix}$$

where  $\omega_N = \exp\left(-i\frac{2\pi}{N}\right)$  is the *base/fundamental frequency* (and see Footnote 12 below). The algorithm works most effectively when  $N$  is the power of 2. For easier explanation, we consider a smaller value of  $N = 2^3 = 8$ . Then we denote

$$G^* = F_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8 & \omega_8^2 & \omega_8^3 & \omega_8^4 & \omega_8^5 & \omega_8^6 & \omega_8^7 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8^8 & \omega_8^{10} & \omega_8^{12} & \omega_8^{14} \\ 1 & \omega_8^3 & \omega_8^6 & \omega_8^9 & \omega_8^{12} & \omega_8^{15} & \omega_8^{18} & \omega_8^{21} \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^{16} & \omega_8^{20} & \omega_8^{24} & \omega_8^{28} \\ 1 & \omega_8^5 & \omega_8^{10} & \omega_8^{15} & \omega_8^{20} & \omega_8^{25} & \omega_8^{30} & \omega_8^{35} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^{24} & \omega_8^{30} & \omega_8^{36} & \omega_8^{42} \\ 1 & \omega_8^7 & \omega_8^{14} & \omega_8^{21} & \omega_8^{28} & \omega_8^{35} & \omega_8^{42} & \omega_8^{49} \end{bmatrix}$$

and the DFT is given by  $\hat{\mathbf{f}} = F_8 \mathbf{f}$ . However, we can shuffle  $\mathbf{f}$  such that the interleaving even and odd indices are split into two groups, and it can be shown that

$$\hat{\mathbf{f}} = \begin{bmatrix} I_4 & D_{4 \rightarrow 8} \\ I_4 & -D_{4 \rightarrow 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} \begin{bmatrix} \mathbf{f}_{\text{even}} \\ \mathbf{f}_{\text{odd}} \end{bmatrix}$$

where

$$D_{4 \rightarrow 8} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \omega_8 & 0 & 0 \\ 0 & 0 & \omega_8^2 & 0 \\ 0 & 0 & 0 & \omega_8^3 \end{bmatrix}$$

and

$$F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_4 & \omega_4^2 & \omega_4^3 \\ 1 & \omega_4^2 & \omega_4^4 & \omega_4^6 \\ 1 & \omega_4^3 & \omega_4^6 & \omega_4^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} \end{bmatrix}$$

where  $\omega_4 = \exp\left(-i\frac{2\pi}{4}\right) = \exp\left(-i\frac{2\pi}{8}(2)\right) = \omega_8^2$ . Subsequently,

$$\begin{aligned} & \begin{bmatrix} I_4 & D_{4 \rightarrow 8} \\ I_4 & -D_{4 \rightarrow 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} \\ &= \begin{bmatrix} I_4 F_4 & D_{4 \rightarrow 8} F_4 \\ I_4 F_4 & -D_{4 \rightarrow 8} F_4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8 & \omega_8^3 & \omega_8^5 & \omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^2 & \omega_8^6 & \omega_8^{10} & \omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^3 & \omega_8^9 & \omega_8^{15} & \omega_8^{21} \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & -\omega_8 & -\omega_8^3 & -\omega_8^5 & -\omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & -\omega_8^2 & -\omega_8^6 & -\omega_8^{10} & -\omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & -\omega_8^3 & -\omega_8^9 & -\omega_8^{15} & -\omega_8^{21} \end{bmatrix} \end{aligned}$$

Note that  $\omega_N^{k+qN} = \omega_N^k$ <sup>12</sup> for any integer  $q$  and  $-1 = \omega_N^{N/2}$ , hence the matrix can be further rewritten into

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8 & \omega_8^3 & \omega_8^5 & \omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^2 & \omega_8^6 & \omega_8^{10} & \omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^3 & \omega_8^9 & \omega_8^{15} & \omega_8^{21} \\ 1 & \omega_8^8 & \omega_8^{16} & \omega_8^{24} & \omega_8^4 & \omega_8^{12} & \omega_8^{20} & \omega_8^{28} \\ 1 & \omega_8^{10} & \omega_8^{20} & \omega_8^{30} & \omega_8^5 & \omega_8^{15} & \omega_8^{25} & \omega_8^{35} \\ 1 & \omega_8^{12} & \omega_8^{24} & \omega_8^{36} & \omega_8^6 & \omega_8^{18} & \omega_8^{30} & \omega_8^{42} \\ 1 & \omega_8^{14} & \omega_8^{28} & \omega_8^{42} & \omega_8^7 & \omega_8^{21} & \omega_8^{35} & \omega_8^{49} \end{bmatrix}$$

which is essentially the same as  $F_8$  except the columns have been arranged into even and odd just like what we have done to  $\mathbf{f}$ . Hence the FFT formulation

$$\hat{\mathbf{f}} = \begin{bmatrix} I_4 & D_{4 \rightarrow 8} \\ I_4 & -D_{4 \rightarrow 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} \begin{bmatrix} \mathbf{f}_{\text{even}} \\ \mathbf{f}_{\text{odd}} \end{bmatrix}$$

will give the same DFT result as the direct formula  $\hat{\mathbf{f}} = F_8\mathbf{f}$ . We can repeat the same procedure to similarly split  $F_4$  into two  $F_2$  blocks and shuffle the indices again. A schematic flowchart is shown as Figure 14.3 below. For a bigger  $N$ , e.g.  $N = 2^{10} = 1024$ , the method proceeds in the same principle such that  $F_{1024} \rightarrow F_{512} \rightarrow F_{256} \rightarrow \dots$ . The FFT time complexity is log-linear  $O(n \log n)$ , where the  $\log n$  factor has replaced  $n$  thanks to its bisective nature. In fact, it requires  $\frac{N}{2} \log_2(N)$  multiplications and  $N \log_2(N)$  additions. This is much faster than  $O(n^2)$  of the naive DFT when  $n$  becomes very large. For instance, the naive DFT of size  $N = 2^{12} = 4096$  will require 33550336 operations but the Radix-2 FFT algorithm only needs 73728 which is around 450 times more efficient. For this reason, any actual computer implementation of DFT will always use FFT underneath.

---

<sup>12</sup> $\omega_N^{k+qN} = \exp\left(-i\frac{2\pi}{N}(k + qN)\right) = \exp\left(-i\left(\frac{2\pi}{N}k + 2\pi q\right)\right) = \exp\left(-i\frac{2\pi}{N}k\right).$

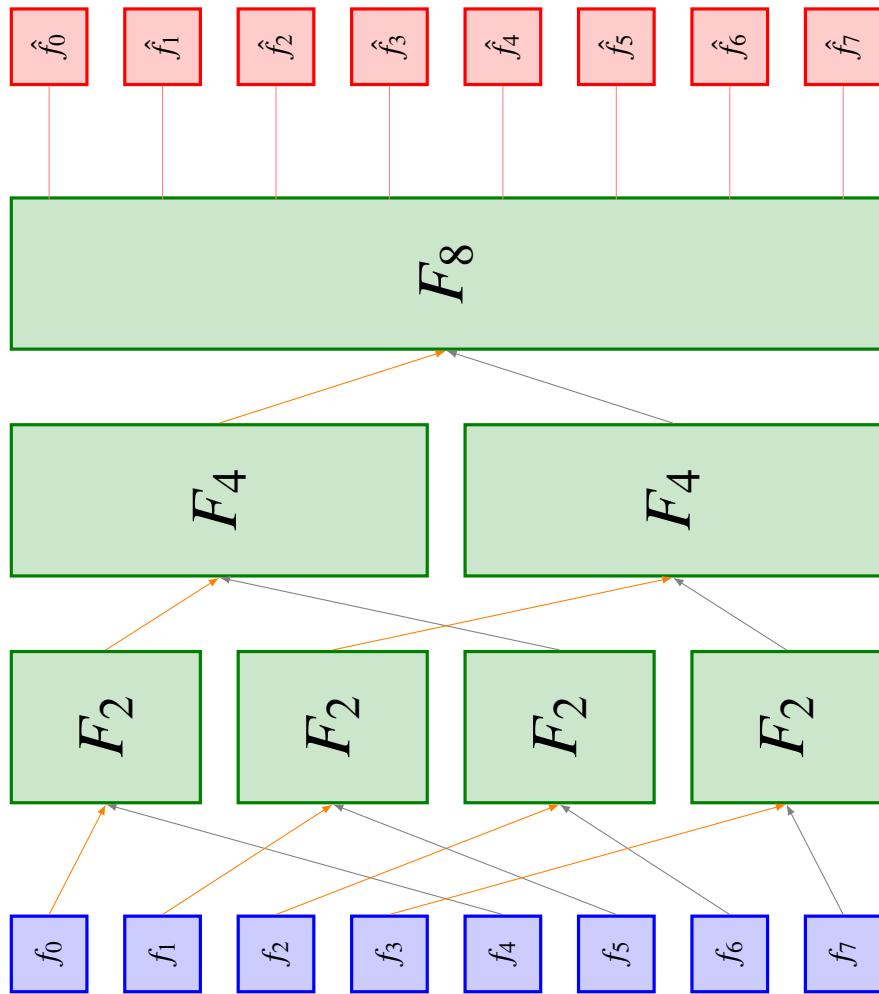


Figure 14.3: A schematic diagram outlining the principle of FFT.

## 14.4 Python Programming and Earth System Applications

Since in practice the time-series to be processed are often very long, it is impossible to do DFT/FFT manually and we will combine the application part with programming tutorial into one single section. We will use the Niño 3.4 SST Index which is an indicator of the ENSO phenomenon discussed in Section 11.4 for demonstration. Now, download the data file from <https://psl.noaa.gov/data/timeseries/month/data/nino34.long.anom.csv> and use the following code to read the time-series.

```
import numpy as np
import pandas as pd

Nino34 = pd.read_csv("nino34.long.anom.csv", header=0, names=[ "Date", "Nino34"])
print(Nino34)
```

Next, import the required functions from the `scipy.fft` module and apply FFT on the Niño 3.4 time-series over the 120 years time period of 1901-2020.

```
from scipy.fft import fft, fftfreq, fftshift

Nino34_120yrs = Nino34[(Nino34["Date"] >= "1901-01-01") & ( Nino34["Date"] <= "2020-12-31")]
print(Nino34_120yrs)

Nino34_fft = fft(Nino34_120yrs["Nino34"].values)
```

Compute the power spectrum by multiplying the transformed data by its complex conjugate.

```
Nino34_power = np.real(Nino34_fft*np.conjugate(Nino34_fft)) # Call the function np.real to remove negligible imaginary parts due to round-off error.
```

Use the `fftfreq` function to produce the frequency and period bins.

```
Nino34_freq = fftfreq(len(Nino34_power), 1/12) # 1 month = 1/12 yrs
Nino34_period = 1/Nino34_freq
```

## *Chapter 14 Discrete Fourier Transform (DFT)*

---

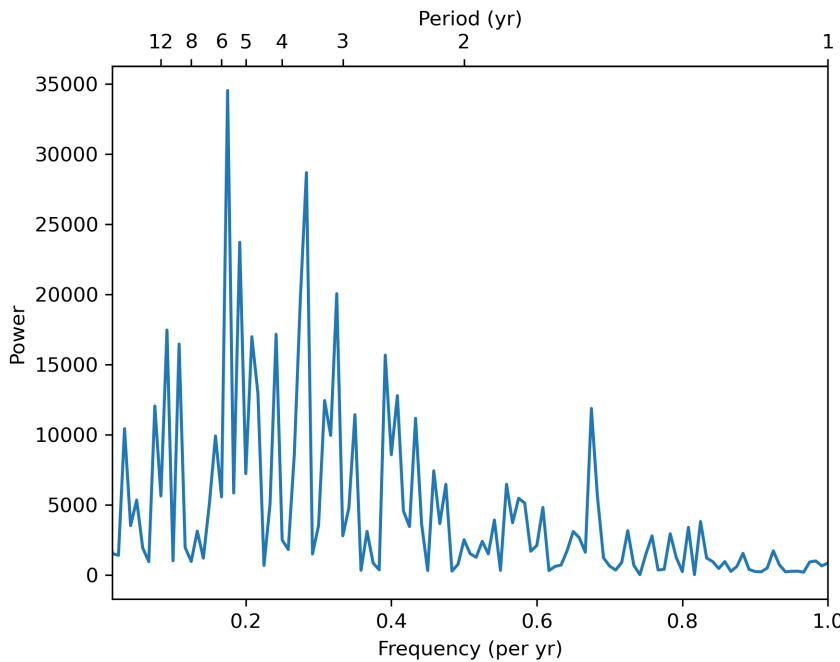
Finally, we can plot the power spectrum as follows.

```
import matplotlib.pyplot as plt

def reciprocal(x): # Function to transform between frequency
    and period.
    return(1/x)

plt.plot(Nino34_freq[:len(Nino34_power)//2], Nino34_power[:len
    (Nino34_power)//2])
plt.xlabel("Frequency (per yr)")
period_ax = plt.gca().secondary_xaxis('top', functions=(

    reciprocal, reciprocal))
plt.xlim([1/60,1])
period_ax.set_ticks([1,2,3,4,5,6,8,12])
period_ax.set_xlabel("Period (yr)")
plt.savefig("NinoFFT")
```



It can be seen that the strongest signals are located over the periods from 2.5

to 7 years, which coincides with the typical time scale of ENSO. We can carry out a simple filtering to extract the Niño 3.4 signals corresponding to ENSO by zeroing out the FFT array at all other frequencies and then apply an inverse FFT:

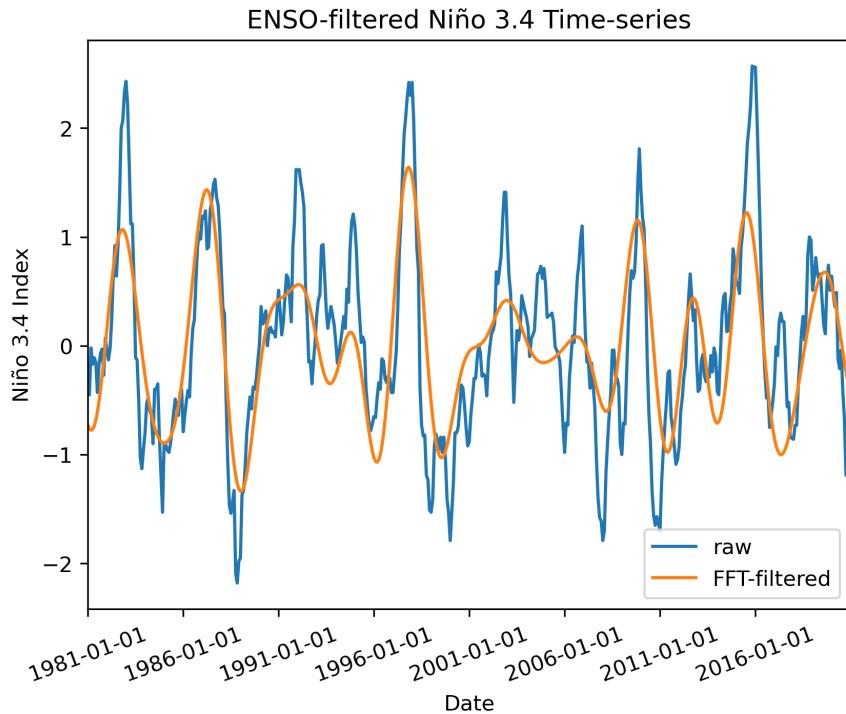
```
from scipy.fft import ifft

Nino34_fft_ENSO = np.copy(Nino34_fft)
Nino34_fft_ENSO[~((2.5 <= np.abs(Nino34_period)) & (np.abs(
    Nino34_period) <= 7))] = 0
Nino34_ENSO = np.real(ifft(Nino34_fft_ENSO))
```

Let's make a plot to compare the filtered time-series with the original one.

```
plt.plot(Nino34_120yrs["Date"], Nino34_120yrs["Nino34"].values
        , label="raw")
plt.plot(Nino34_120yrs["Date"], Nino34_ENSO, label="FFT-
        filtered")
plt.xticks(np.arange(0, len(Nino34_120yrs["Date"])), 60,
           rotation=20)
plt.xlim(["1981-01-01", "2020-12-31"])
plt.xlabel("Date")
plt.ylabel("Niño 3.4 Index")
plt.legend()
plt.title("ENSO-filtered Niño 3.4 Time-series")
plt.savefig("NinoENSOFilter")
```

However, note that this "zeroing-out" filtering method is not a very good idea to be implemented in practice and we do it here only due to heuristic purpose. For more information, search about the "*Gibbs Phenomenon*" and also read the comprehensive discussion in [this DSP StackExchange post](#) (6220).



## 14.5 Exercise

**Exercise 14.1** Compute the Discrete Fourier Transform for the following data.

unit time	0	1	2	3	4
f(t)	4.5	6.2	7.8	1.1	3.4
unit time	5	6	7	8	9
f(t)	2.5	3.6	5.9	2.9	6.0

Find the amplitude/power and phase of the sinusoidal wave signal corresponding to the third frequency bin, i.e. with an angular frequency of  $\omega = 2\pi(\frac{3}{10})$ .

**Exercise 14.2** Download an [ERA5 Temperature dataset](#) over any time period of 15 years. Select any location as you like, and extract the temperature time-series there. Apply DFT on the time series, and identify any dominant frequency or period with a large power magnitude. Explain the peaks with Earth Science knowledge.

**Exercise 14.3** Perform DFT on the time-series for the two MJO EOF modes derived in Exercise ([11.7](#)) and deduce the characterisitc time scale of MJO by plotting the power spectrum against periods.

**Exercise 14.4** Find the circular convolution of two time-series  $f(t) = (1, 4, 2, 4, 3, 0, -1, 2)$  and  $g(t) = (2, 3, -2, 1, -1, 0, 4, 3)$  by definition, as well as via the Convolution Theorem to check the consistency.

**Exercise 14.5** Write your own FFT function in Python and compare with the one in the `scipy.fft` library by testing them on any time-series.



# Answers to Exercises

---

## Exercise 1.1

(a)  $\begin{bmatrix} -3 & 5 \\ 3 & 6 \end{bmatrix}$

(b)  $\begin{bmatrix} 8 & -\frac{1}{2} \\ 13 & -\frac{25}{2} \end{bmatrix}$

(c)  $\begin{bmatrix} -8 & 17 \\ -18 & 8 \end{bmatrix}$

(d)  $\begin{bmatrix} 11 & -11 \\ 33 & -11 \end{bmatrix}$

## Exercise 1.2

(a)  $\begin{bmatrix} -2 & 1 & 3 \\ -1 & -1 & -9 \\ -8 & 2 & -2 \end{bmatrix}$

(b)  $\begin{bmatrix} -8 & -5 \\ 15 & 3 \end{bmatrix}$

## Exercise 1.3

(a)  $\begin{bmatrix} 42 & 72 & 0 \\ 32 & 51 & -1 \end{bmatrix}$

(b) Same as above

## *Answer to Exercises*

---

(c)  $\begin{bmatrix} 90 & 162 & 2 \\ 51 & 99 & 3 \end{bmatrix}$

(d) Same as above

### **Exercise 1.4**

(a)  $\begin{bmatrix} 16 & 23 & 129 \\ 133 & 33 & 102 \\ 27 & 9 & 128 \end{bmatrix}$

(b)  $\begin{bmatrix} -\frac{233}{4} & -\frac{19}{4} & \frac{69}{2} \\ -\frac{339}{4} & -16 & 31 \\ \frac{109}{4} & \frac{33}{4} & -\frac{289}{4} \end{bmatrix}$

### **Exercise 1.5**

(a)  $\begin{bmatrix} 16 & 6 & 3 \\ 34 & 13 & 12 \\ 9 & 2 & 27 \end{bmatrix}$

(b)  $\begin{bmatrix} 27 & 15 & 69 \\ 37 & 12 & 85 \\ 36 & 12 & 69 \end{bmatrix}$

(c)  $\begin{bmatrix} 14 & 3 & 26 \\ 29 & 9 & 60 \\ 12 & 21 & 41 \end{bmatrix}$

(d)  $\begin{bmatrix} 33 & 13 & 24 \\ 47 & 19 & 21 \\ 39 & 14 & 12 \end{bmatrix}$

### **Exercise 1.6**

$$\begin{bmatrix} 0 & 3 & -4 \\ 5 & -1 & 2 \\ 6 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 13 \\ 8 \end{bmatrix}$$

or

$$\left[ \begin{array}{ccc|c} 0 & 3 & -4 & 6 \\ 5 & -1 & 2 & 13 \\ 6 & 0 & 1 & 8 \end{array} \right]$$

### Exercise 1.7

(a)  $\left[ \begin{array}{cccc} 2 & 3 & 5 & 7 \\ 1 & 2 & 4 & 8 \\ 1 & 4 & 8 & 12 \end{array} \right]$

(b) (WIP)

**Exercise 1.8** The air temperature/dew point at any height  $z$  before saturation is  $T_a = T_{a,ini} - (\Gamma_{dry})z$  and  $T_{dew} = T_{dew,ini} - (\Gamma_{dew})z$  respectively. At the condensation level  $z = z_{cd}$ , the air temperature equals to the dew point temperature  $T_a = T_{dew} = T_{cd}$ , and hence we have

$$T_{a,ini} - \Gamma_{dry}(z_{cd}) = T_{dew,ini} - \Gamma_{dew}(z_{cd}) = T_{cd}$$

which can be separated into two equations

$$\begin{cases} T_{a,ini} - \Gamma_{dry}(z_{cd}) = T_{cd} \\ T_{dew,ini} - \Gamma_{dew}(z_{cd}) = T_{cd} \end{cases}$$

Rearranging to put the unknowns  $z_{cd}$  and  $T_{cd}$  to L.H.S., we obtain

$$\begin{cases} T_{cd} + \Gamma_{dry}(z_{cd}) = T_{a,ini} \\ T_{cd} + \Gamma_{dew}(z_{cd}) = T_{dew,ini} \end{cases}$$

or, in matrix form

$$\begin{bmatrix} 1 & \Gamma_{dry} \\ 1 & \Gamma_{dew} \end{bmatrix} \begin{bmatrix} T_{cd} \\ z_{cd} \end{bmatrix} = \begin{bmatrix} T_{a,ini} \\ T_{dew,ini} \end{bmatrix}$$

Plugging in the lapse rates, we have

$$\begin{bmatrix} 1 & 9.8 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} T_{cd} \\ z_{cd} \end{bmatrix} = \begin{bmatrix} 25.4 \\ 17.8 \end{bmatrix}$$

**Exercise 1.9** Obviously, there are 35 chickens and rabbits in total, and  $x + y = 35$ . Considering the total amount of legs, we also have  $2x + 4y = 94$ . Hence the required linear system is

$$\begin{cases} x + y = 35 \\ 2x + 4y = 94 \end{cases}$$

In matrix form, it is

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 35 \\ 94 \end{bmatrix}$$

**Exercise 2.1** (Applying cofactor expansion along the leftmost column recursively) The determinant is just the product of the diagonal elements =  $(1)(6)(10)(13)(15) = 11700$ .

### Exercise 2.2

$$(a) \begin{bmatrix} 8 & 20 \\ 15 & 37 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 6 & 1 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 3 & 7 \end{bmatrix}$$

$$(b) \begin{bmatrix} -\frac{37}{4} & \frac{15}{4} \\ 5 & -2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & -\frac{3}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -7 & 3 \\ 5 & -2 \end{bmatrix}$$

$$(c) \begin{vmatrix} 8 & 15 \\ 20 & 37 \end{vmatrix} = -4 = (-1)(4) = \begin{vmatrix} 2 & 3 \\ 5 & 7 \end{vmatrix} \begin{vmatrix} 4 & 6 \\ 0 & 1 \end{vmatrix}$$

### Exercise 2.3

(a)

$$\begin{array}{l}
 \left[ \begin{array}{ccc|ccc} 3 & 2 & 9 & 1 & 0 & 0 \\ 1 & 2 & 3 & 0 & 1 & 0 \\ 4 & 0 & 4 & 0 & 0 & 1 \end{array} \right] \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 3 & 2 & 9 & 1 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 1 \end{array} \right] \quad R_1 \leftrightarrow R_2 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & -4 & 0 & 1 & -3 & 0 \\ 0 & -8 & -8 & 0 & -4 & 1 \end{array} \right] \quad R_2 - 3R_1 \rightarrow R_2, R_3 - 4R_1 \rightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 1 & 1 & 0 & \frac{1}{2} & -\frac{1}{8} \end{array} \right] \quad -\frac{1}{4}R_2 \rightarrow R_2, -\frac{1}{8}R_3 \rightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{array} \right] \quad R_3 - R_2 \rightarrow R_3 \\
 \rightarrow \left[ \begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{4} & \frac{1}{4} & \frac{3}{8} \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{array} \right] \quad R_1 - 3R_3 - 2R_2 \rightarrow R_1
 \end{array}$$

 (b)  $\det(A) = -32$  and

$$\begin{aligned}
 \text{adj}(A) &= \begin{bmatrix} \left| \begin{array}{cc} 2 & 3 \\ 0 & 4 \end{array} \right| & -\left| \begin{array}{cc} 1 & 3 \\ 4 & 4 \end{array} \right| & \left| \begin{array}{cc} 1 & 2 \\ 4 & 0 \end{array} \right| \\ -\left| \begin{array}{cc} 2 & 9 \\ 0 & 4 \end{array} \right| & \left| \begin{array}{cc} 3 & 9 \\ 4 & 4 \end{array} \right| & -\left| \begin{array}{cc} 3 & 2 \\ 4 & 0 \end{array} \right| \\ \left| \begin{array}{cc} 2 & 9 \\ 2 & 3 \end{array} \right| & -\left| \begin{array}{cc} 3 & 9 \\ 1 & 3 \end{array} \right| & \left| \begin{array}{cc} 3 & 2 \\ 1 & 2 \end{array} \right| \end{bmatrix}^T \\
 &= \begin{bmatrix} 8 & 8 & -8 \\ -8 & -24 & 8 \\ -12 & 0 & 4 \end{bmatrix}^T
 \end{aligned}$$

$$= \begin{bmatrix} 8 & -8 & -12 \\ 8 & -24 & 0 \\ -8 & 8 & 4 \end{bmatrix}$$

Hence

$$\begin{aligned} A^{-1} &= \frac{1}{\det(A)} \text{adj}(A) \\ &= -\frac{1}{32} \begin{bmatrix} 8 & -8 & -12 \\ 8 & -24 & 0 \\ -8 & 8 & 4 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{3}{8} \\ -\frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{bmatrix} \end{aligned}$$

#### Exercise 2.4

$$(a) \begin{bmatrix} 19 & 35 & 9 \\ 33 & 61 & 16 \\ 52 & 96 & 24 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 3 \\ 3 & 4 & 5 \\ 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 2 & 4 & 2 \\ 5 & 9 & 1 \end{bmatrix}$$

$$(b) \begin{bmatrix} 18 & -10 & 1 \\ -6 & 3 & 1 \\ -\frac{11}{4} & \frac{7}{4} & -1 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 \\ 1 & 2 & -2 \\ -1 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 7 & -4 & 1 \\ -\frac{9}{2} & \frac{5}{2} & 0 \\ 2 & -1 & 0 \end{bmatrix}$$

$$(c) \begin{vmatrix} 19 & 33 & 52 \\ 35 & 61 & 96 \\ 9 & 16 & 24 \end{vmatrix} = -4 = (-2)(2) = \begin{vmatrix} 0 & 2 & 5 \\ 0 & 4 & 9 \\ 1 & 2 & 1 \end{vmatrix} \begin{vmatrix} 2 & 3 & 4 \\ 2 & 4 & 6 \\ 3 & 5 & 8 \end{vmatrix}$$

**Exercise 2.5** Either by evaluating the determinant to show that  $|A| = 0$ , or find its reduced row echelon form which is

$$\begin{bmatrix} 1 & 0 & -\frac{1}{3} \\ 0 & 1 & \frac{5}{3} \\ 0 & 0 & 0 \end{bmatrix}$$

and not equal to the identity.

### **Exercise 2.6**

$$\det(A) = -42$$

$$\det(A^{-1}) = -\frac{1}{42}$$

$$A^{-1} = \begin{bmatrix} \frac{9}{7} & -\frac{3}{14} & -\frac{2}{7} & -\frac{6}{7} \\ -\frac{1}{21} & -\frac{1}{21} & \frac{1}{21} & \frac{1}{7} \\ -\frac{11}{7} & -\frac{15}{14} & \frac{11}{7} & \frac{5}{7} \\ \frac{3}{7} & \frac{3}{7} & -\frac{3}{7} & -\frac{2}{7} \end{bmatrix}$$

**Exercise 2.7** By cofactor expansion along the first column, we can obtain the determinant of  $A$  as

$$|A| = 2p^2 + 4p - 16$$

which has two roots,  $p = -4$  and  $p = 2$  such that  $|A| = 0$  and  $A$  is not invertible. All values of  $p$  other than  $p = -4$  and  $p = 2$  make  $A$  invertible.

**Exercise 2.8**  $(A + A^T)^T = A^T + (A^T)^T = A^T + A = A + A^T$ , and  $(A - A^T)^T = A^T - (A^T)^T = A^T - A = -(A - A^T)$ . We can split  $A$  into

$$\begin{aligned} A &= A + \frac{1}{2}(A^T - A^T) \\ &= \frac{1}{2}A + \frac{1}{2}A + \frac{1}{2}A^T - \frac{1}{2}A^T \\ &= \frac{1}{2}A + \frac{1}{2}A^T + \frac{1}{2}A - \frac{1}{2}A^T \\ &= \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T) \end{aligned}$$

where the first term is symmetric and the second term is skew-symmetric.

### Exercise 2.9

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$$

$$\det(A^{-1}) = \det\left(\frac{1}{\det(A)} \text{adj}(A)\right) \quad (\text{Notice that } \frac{1}{\det(A)} \text{ is now a scalar})$$

$$\frac{1}{\det(A)} = \left(\frac{1}{\det(A)}\right)^n \det(\text{adj}(A))$$

$$\det(\text{adj}(A)) = (\det(A))^{n-1}$$

### Exercise 3.1

$$A^{-1} = \begin{bmatrix} \frac{1}{21} & \frac{5}{21} & \frac{2}{21} \\ \frac{11}{42} & -\frac{29}{42} & \frac{1}{42} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \end{bmatrix}$$

$$\vec{x} = A^{-1}\vec{h} = \begin{bmatrix} \frac{1}{21} & \frac{5}{21} & \frac{2}{21} \\ \frac{11}{42} & -\frac{29}{42} & \frac{1}{42} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \end{bmatrix} \begin{bmatrix} 6 \\ 7 \\ -\frac{13}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{3}{2} \end{bmatrix}$$

or

$$\left[ \begin{array}{ccc|c} 5 & 1 & 3 & 6 \\ 2 & -1 & 1 & \frac{7}{2} \\ 3 & 2 & -4 & -\frac{13}{2} \end{array} \right]$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 2 & -1 & 1 & \frac{7}{2} \\ 3 & 2 & -4 & -\frac{13}{2} \end{array} \right] \quad \frac{1}{5}R_1 \rightarrow R_1$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & -\frac{7}{5} & -\frac{1}{5} & \frac{11}{10} \\ 0 & \frac{7}{5} & -\frac{29}{5} & -\frac{101}{10} \end{array} \right] \quad R_2 - 2R_1 \rightarrow R_2, R_3 - 3R_1 \rightarrow R_3$$

$$\rightarrow \left[ \begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & -\frac{7}{5} & -\frac{1}{5} & \frac{11}{10} \\ 0 & 0 & -6 & -9 \end{array} \right] \quad R_3 + R_2 \rightarrow R_3$$

$$\begin{aligned}
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & 1 & \frac{1}{7} & -\frac{11}{14} \\ 0 & 0 & 1 & \frac{3}{2} \end{array} \right] \quad -\frac{5}{7}R_2 \rightarrow R_2, -\frac{1}{6}R_3 \rightarrow R_3 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & \frac{1}{5} & 0 & \frac{3}{10} \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & \frac{3}{2} \end{array} \right] \quad R_1 - \frac{3}{5}R_3 \rightarrow R_1, R_2 - \frac{1}{7}R_3 \rightarrow R_2 \\
 & \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & \frac{3}{2} \end{array} \right] \quad R_1 - \frac{1}{5}R_2 \rightarrow R_1
 \end{aligned}$$

**Exercise 3.2**

$$\begin{aligned}
 A^{-1} &= \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix} \\
 \vec{x}_1 = A^{-1}\vec{h}_1 & \qquad \qquad \qquad \vec{x}_2 = A^{-1}\vec{h}_2 \\
 &= \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix} \begin{bmatrix} -1 \\ 5 \\ 1 \end{bmatrix} \quad &= \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix} \begin{bmatrix} \frac{19}{4} \\ 1 \\ \frac{5}{4} \end{bmatrix} \\
 \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix} \qquad \qquad \qquad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{4} \end{bmatrix}
 \end{aligned}$$

**Exercise 3.3**

$$\begin{aligned}
 \left[ \begin{array}{ccc|c} 3 & 0 & 4 & 2 \\ 1 & 1 & 2 & -1 \\ 1 & -2 & 0 & 0 \end{array} \right] &\rightarrow \left[ \begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 1 & 1 & 2 & -1 \\ 3 & 0 & 4 & 2 \end{array} \right] \quad R_1 \leftrightarrow R_3 \\
 &\rightarrow \left[ \begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 3 & 2 & -1 \\ 0 & 6 & 4 & 2 \end{array} \right] \quad R_2 - R_1 \rightarrow R_2, R_3 - 3R_1 \rightarrow R_3
 \end{aligned}$$

$$\begin{aligned} & \rightarrow \left[ \begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 6 & 4 & 2 \end{array} \right] \qquad \qquad \qquad \frac{1}{3}R_2 \rightarrow R_2 \\ & \rightarrow \left[ \begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 0 & 0 & 4 \end{array} \right] \qquad \qquad \qquad R_3 - 6R_2 \rightarrow R_3 \end{aligned}$$

The last row is inconsistent and the system has no solution.

Note: You may get, to the right of the last row, some number other than 4, but this is possible and not wrong. (Why?)

### Exercise 3.4

$$\begin{aligned} & \left[ \begin{array}{cccc|c} 1 & 1 & -1 & -3 & 2 \\ 1 & 0 & 0 & -1 & 5 \\ 3 & 2 & -2 & -7 & 9 \end{array} \right] \\ & \rightarrow \left[ \begin{array}{cccc|c} 1 & 0 & 0 & -1 & 5 \\ 1 & 1 & -1 & -3 & 2 \\ 3 & 2 & -2 & -7 & 9 \end{array} \right] \qquad \qquad \qquad R_1 \leftrightarrow R_2 \\ & \rightarrow \left[ \begin{array}{cccc|c} 1 & 0 & 0 & -1 & 5 \\ 0 & 1 & -1 & -2 & -3 \\ 0 & 2 & -2 & -4 & -6 \end{array} \right] \qquad \qquad \qquad R_2 - R_1 \rightarrow R_2, R_3 - 3R_1 \rightarrow R_3 \\ & \rightarrow \left[ \begin{array}{cccc|c} 1 & 0 & 0 & -1 & 5 \\ 0 & 1 & -1 & -2 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right] \qquad \qquad \qquad R_3 - 2R_2 \rightarrow R_3 \end{aligned}$$

Let  $p = s$ ,  $q = t$  as the two free variables. Substituting them back into the equations, we have  $m - t = 5$  and  $n - s - 2t = -3$ , hence  $m = 5 + t$  and  $n = -3 + s + 2t$ , and

$$\begin{bmatrix} m \\ n \\ p \\ q \end{bmatrix} = \begin{bmatrix} 5+t \\ -3+s+2t \\ s \\ t \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

**Exercise 3.5** The determinant of the coefficient matrix can be found to be

$$\begin{vmatrix} 1 & 0 & \alpha \\ 0 & \alpha & 0 \\ \alpha & 0 & 1 \end{vmatrix} = -\alpha^3 + \alpha \\ = -\alpha(\alpha - 1)(\alpha + 1)$$

The system will have no solution or infinitely many of them only when the determinant equals to zero, which gives us three possible values of  $\alpha = -1, 0, 1$ . When  $\alpha = -1$ , the system is

$$\left[ \begin{array}{ccc|c} 1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 1 & -1 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 \end{array} \right] \quad R_3 + R_1 \rightarrow R_3$$

where the last row is inconsistent and there is no solution. When  $\alpha = 0$ , it becomes

$$\left[ \begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array} \right]$$

It is obvious that  $x = z = 0$ , and  $y = t$  is a free variable, so the solution is infinitely many and is in the form of

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{0} + t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The last case,  $\alpha = 1$ , gives rise to the system of

$$\left[ \begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|c} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \quad R_3 - R_1 \rightarrow R_3$$

such that  $y = 0$  and  $z = t$  can be set to be a free variable and there are infinitely many solutions in the form of

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

**Exercise 3.6** The first two equations below come from the left inner loop and right inner loop, but one of them can be replaced by the outer loop as well.

$$\begin{aligned} -4I_1 + 6I_2 &= 6 \\ -6I_2 + 9I_3 &= -12 \\ I_1 + I_2 + I_3 &= 0 \end{aligned}$$

and the solution is  $I_1 = -\frac{3}{19}$ ,  $I_2 = \frac{17}{19}$ ,  $I_3 = -\frac{14}{19}$  (in Amperes).

**Exercise 3.7** Substituting the given wave solution forms into the equation, we have

$$\begin{aligned} \omega\tilde{\eta} \sin(kx + ly - \omega t) + H(-k\tilde{U} \sin(kx + ly - \omega t) \\ - l\tilde{V} \sin(kx + ly - \omega t)) &= 0 \\ \omega\tilde{U} \sin(kx + ly - \omega t) &= gk\tilde{\eta} \sin(kx + ly - \omega t) \\ \omega\tilde{V} \sin(kx + ly - \omega t) &= gl\tilde{\eta} \sin(kx + ly - \omega t) \end{aligned}$$

Cancelling out all the sine factors, we arrive at the linear system displayed in the question

$$\begin{cases} \omega\tilde{\eta} - kH\tilde{U} - lH\tilde{V} &= 0 \\ \omega\tilde{U} - kg\tilde{\eta} &= 0 \\ \omega\tilde{V} - lg\tilde{\eta} &= 0 \end{cases}$$

For  $\tilde{U}$ ,  $\tilde{V}$ ,  $\tilde{\eta}$  to have a non-trivial solution other than all zeros, we require the determinant of the corresponding coefficient matrix to be zero according to Theorem 3.1.2, which leads to

$$\begin{vmatrix} \omega & -kH & -lH \\ -kg & \omega & 0 \\ -lg & 0 & \omega \end{vmatrix} = 0$$

$$\omega^3 - gHk^2\omega - gHl^2\omega = 0$$

$$\omega^2 - gH(k^2 + l^2) = 0$$

as the dispersion relation of gravity wave.

**Exercise 3.8**  $T_{cd} \approx 15.9^\circ\text{C}$ ,  $z_{cd} \approx 0.97 \text{ km}$ .

**Exercise 3.9**  $x = 23$ ,  $y = 12$ . For the extra part, the new system of equations become (denote the number of third species as  $z$ )

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 48 \\ 122 \end{bmatrix}$$

By Gaussian Elimination, we have

$$\begin{array}{l} \left[ \begin{array}{ccc|cc} 1 & 1 & 1 & 48 \\ 2 & 4 & 3 & 122 \end{array} \right] \rightarrow \left[ \begin{array}{ccc|cc} 1 & 1 & 1 & 48 \\ 0 & 2 & 1 & 26 \end{array} \right] \quad R_2 - 2R_1 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{ccc|cc} 1 & 1 & 1 & 48 \\ 0 & 1 & \frac{1}{2} & 13 \end{array} \right] \quad \frac{1}{2}R_2 \rightarrow R_2 \\ \rightarrow \left[ \begin{array}{ccc|cc} 1 & 0 & \frac{1}{2} & 35 \\ 0 & 1 & \frac{1}{2} & 13 \end{array} \right] \quad R_1 - R_2 \rightarrow R_1 \end{array}$$

Let  $z = t$  as the free variable, then we have  $y = 13 - \frac{1}{2}t$  and  $x = 35 - \frac{1}{2}t$ , and hence

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 35 - \frac{1}{2}t \\ 13 - \frac{1}{2}t \\ t \end{bmatrix} = \begin{bmatrix} 35 \\ 13 \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{bmatrix}$$

Since the numbers of species must be a non-negative integer, the solution can be expressed in a more good-looking form of

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 35 \\ 13 \\ 0 \end{bmatrix} + s \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$$

where  $s = \frac{t}{2}$ , and the range of  $s$  is  $0, 1, \dots, 13$  (when  $s$  reaches 13 there is no chicken remained).

**Exercise 4.1**

- (a)  $(2, 5, 5, 12)^T$   
 (b)  $(1, \frac{7}{2}, \frac{7}{2}, 8)^T$   
 (c)  $(1)(1) + (3)(2) + (3)(2) + (7)(5) = 48$   
 (d)  $(1)(1) + (2)(3) + (2)(3) + (5)(7) = 48$   
 (e)  $\vec{u} - 2\vec{v} = (-1, -1, -1, -3)^T, 2\vec{u} + \vec{v} = (3, 8, 8, 19)^T, (\vec{u} - 2\vec{v}) \cdot (2\vec{u} + \vec{v}) = (-1)(3) + (-1)(8) + (-1)(8) + (-3)(19) = -76$

### Exercise 4.2

(a)

$$\vec{u} \times \vec{v} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 7 & 4 & 1 \\ 8 & 1 & 1 \end{vmatrix} = 3\hat{i} + \hat{j} - 25\hat{k} = (3, 1, -25)^T$$

$$\vec{v} \times \vec{u} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 8 & 1 & 1 \\ 7 & 4 & 1 \end{vmatrix} = -3\hat{i} - \hat{j} + 25\hat{k} = (-3, -1, 25)^T$$

(b)

$$A\vec{v} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 8 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 2 \\ 1 \end{bmatrix}$$

$$\vec{u} \cdot (A\vec{v}) = (7, 4, 1)^T \cdot (10, 2, 1)^T$$

$$= (7)(10) + (4)(2) + (1)(1)$$

$$= 79$$

$$A^T \vec{u} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 11 \\ 12 \end{bmatrix}$$

$$(A^T \vec{u}) \cdot \vec{v} = (7, 11, 12)^T \cdot (8, 1, 1)^T$$

$$= (7)(8) + (11)(1) + (12)(1)$$

$$= 79$$

(c) By (a),  $\vec{u} \times \vec{v} = (3, 1, -25)^T$  and  $(3\vec{u} - 4\vec{v}) = (-11, 8, -1)^T$ , then

$$\begin{aligned}(3\vec{u} - 4\vec{v}) \cdot (\vec{u} \times \vec{v}) &= (-11, 8, -1)^T \cdot (3, 1, -25)^T \\ &= (-11)(3) + (8)(1) + (-1)(-25) = 0\end{aligned}$$

This makes sense as we have shown that  $\vec{u} \cdot (\vec{u} \times \vec{v}) = \vec{v} \cdot (\vec{u} \times \vec{v}) = 0$ , and therefore by distributive property  $(\alpha\vec{u} + \beta\vec{v}) \cdot (\vec{u} \times \vec{v}) = 0$  for any  $\alpha$  and  $\beta$ .

### Exercise 4.3

(a)

$$\|\vec{u}\| = \sqrt{1^2 + (-3)^2 + 9^2} = \sqrt{91}$$

$$\hat{u} = \left( \frac{1}{\sqrt{91}}, -\frac{3}{\sqrt{91}}, \frac{9}{\sqrt{91}} \right)^T$$

$$\|\vec{v}\| = \sqrt{1^2 + (-2)^2 + 4^2} = \sqrt{21}$$

$$\hat{v} = \left( \frac{1}{\sqrt{21}}, -\frac{2}{\sqrt{21}}, \frac{4}{\sqrt{21}} \right)^T$$

(b)

$$\vec{u} \cdot \vec{v} = (1)(1) + (-3)(-2) + (9)(4) = 43$$

$$\cos \theta = \frac{43}{\sqrt{21}\sqrt{91}} \approx 0.9836$$

$$\theta \approx 0.181 \text{ rad}$$

$$(c) \vec{u} \times \vec{v} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 1 & -3 & 9 \\ 1 & -2 & 4 \end{vmatrix} = 6\hat{i} + 5\hat{j} + \hat{k} = (6, 5, 1)^T$$

## *Answer to Exercises*

---

$$(d) \vec{u} \cdot (\vec{u} \times \vec{v}) = (1, -3, 9)^T \cdot (6, 5, 1)^T = (1)(6) + (-3)(5) + (9)(1) = 0,$$

$$\vec{v} \cdot (\vec{u} \times \vec{v}) = (1)(6) + (-2)(5) + (4)(1) = 0$$

### Exercise 4.4

Typhoon Name	Time	Speed	Direction	Vector Form
Nuri	2008/08/22, 08:00	13 km h <sup>-1</sup>	315°	(-9.192, 9.192)
Vicente	2012/07/24, 02:00	18 km h <sup>-1</sup>	299°	(-15.743, 8.727)
Linfa	2015/07/09, 23:00	15 km h <sup>-1</sup>	245°	(-13.595, -6.339)
Mangkhut	2018/09/16, 22:00	25 km h <sup>-1</sup>	288°	(-23.776, 7.725)

### Exercise 4.5

$$\begin{aligned}\|\vec{u} + \vec{v}\|^2 &= (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) \\ &= \|\vec{u}\|^2 + 2(\vec{u} \cdot \vec{v}) + \|\vec{v}\|^2 \\ &\leq \|\vec{u}\|^2 + 2\|\vec{u}\|\|\vec{v}\| + \|\vec{v}\|^2 \\ &= (\|\vec{u}\| + \|\vec{v}\|)^2\end{aligned}$$

### Exercise 4.6

$$\begin{aligned}\|\vec{u} + \vec{v}\|^2 + \|\vec{u} - \vec{v}\|^2 &= (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) + (\vec{u} - \vec{v}) \cdot (\vec{u} - \vec{v}) \\ &= (\|\vec{u}\|^2 + 2(\vec{u} \cdot \vec{v}) + \|\vec{v}\|^2) + (\|\vec{u}\|^2 - 2(\vec{u} \cdot \vec{v}) + \|\vec{v}\|^2) \\ &= 2\|\vec{u}\|^2 + 2\|\vec{v}\|^2\end{aligned}$$

**Exercise 4.7** In Example 4.3.1, we have

$$\overrightarrow{F_{\text{cor}}} = (2\Omega(v \sin \varphi - w \cos \varphi))\hat{i} + (-2\Omega u \sin \varphi)\hat{j} + (2\Omega u \cos \varphi)\hat{k}$$

and hence the rate of work done is

$$\overrightarrow{F_{\text{cor}}} \cdot \vec{v}$$

$$\begin{aligned}
 &= [(2\Omega(v \sin \varphi - w \cos \varphi))\hat{i} + (-2\Omega u \sin \varphi)\hat{j} + (2\Omega u \cos \varphi)\hat{k}] \cdot (u\hat{i} + v\hat{j} + w\hat{k}) \\
 &= (2\Omega(v \sin \varphi - w \cos \varphi))u + (-2\Omega u \sin \varphi)v + (2\Omega u \cos \varphi)w \\
 &= 2\Omega uv \sin \varphi - 2\Omega uw \sin \varphi - 2\Omega uv \sin \varphi + 2\Omega uw \sin \varphi = 0
 \end{aligned}$$

Alternatively, note that  $\overrightarrow{F_{\text{cor}}} = -2\vec{\Omega} \times \vec{v}$  and  $(\vec{\Omega} \times \vec{v}) \cdot \vec{v} = 0$  always holds.

### Exercise 5.1

$$(a) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{4}{3} \\ 1 \end{bmatrix}$$

$$(b) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 7 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} -9 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -9 \\ 0 \\ 1 \end{bmatrix}$$

$$(c) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$(d) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{9}{2} \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix}$$

### Exercise 5.2

$$(a) \text{Normal vector to the line is } \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

$$\text{Equation: } \begin{bmatrix} 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 9 \end{bmatrix} \rightarrow x - y = -7$$

$$(b) \text{Normal vector to the plane is } \begin{bmatrix} 7 \\ 4 \\ 1 \end{bmatrix} \times \begin{bmatrix} 8 \\ 0 \\ 5 \end{bmatrix} = \begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix}.$$

$$\text{Equation: } \begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix} \cdot \begin{bmatrix} 6 \\ 3 \\ 2 \end{bmatrix} \rightarrow 20x - 27y - 32z = -25$$

**Exercise 5.3** Part 1: Choose  $(0, 0, 2)^T$  as a reference point on the plane.

Projection of the vector from  $(0, 0, 2)^T$  to  $(3, 2, 9)^T$ :  $(3-0)\hat{i} + (2-0)\hat{j} + (9-2)\hat{k} = 3\hat{i} + 2\hat{j} + 7\hat{k}$  onto the normal vector  $\hat{i} + 2\hat{j} + 5\hat{k}$  of the plane is

$$\frac{(3)(1) + (2)(2) + (7)(5)}{\sqrt{1^2 + 2^2 + 5^2}} = \frac{42}{\sqrt{30}}$$

which is the required distance.

Part 2: Choose  $(0, 1, 2)^T$  as a reference point along the line. Find the projection of  $(3, 2, 9)^T - (0, 1, 2)^T = 3\hat{i} + 1\hat{j} + 7\hat{k}$  onto the direction vector  $\hat{j} + 2\hat{k}$ , which is

$$\frac{(3)(0) + (1)(1) + (7)(2)}{0^2 + 1^2 + 2^2}(\hat{j} + 2\hat{k}) = 3(\hat{j} + 2\hat{k}) = 3\hat{j} + 6\hat{k}$$

The displacement vector between the point and line (which is orthogonal to the line) is then  $(3\hat{i} + 1\hat{j} + 7\hat{k}) - (3\hat{j} + 6\hat{k}) = 3\hat{i} - 2\hat{j} + \hat{k}$  and the required distance equals to  $\sqrt{3^2 + (-2)^2 + 1^2} = \sqrt{14}$ .

**Exercise 5.4** Using the hints, we have the distance as

$$\begin{aligned} \frac{(\vec{v} - \vec{u}) \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|} &= \frac{[(\vec{b} + \hat{m}t) - (\vec{a} + \hat{l}s)] \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|} \\ &= \frac{(\vec{b} - \vec{a}) \cdot (\hat{l} \times \hat{m}) + [\hat{m} \cdot (\hat{l} \times \hat{m})]t - [\hat{l} \cdot (\hat{l} \times \hat{m})]s}{\|\hat{l} \times \hat{m}\|} \end{aligned}$$

Notice that  $\hat{l} \times \hat{m}$  is orthogonal to both  $\hat{l}$  and  $\hat{m}$ , and thus  $\hat{l} \cdot (\hat{l} \times \hat{m}) = \hat{m} \cdot (\hat{l} \times \hat{m}) = 0$  both vanish. Therefore we are left with

$$\frac{(\vec{b} - \vec{a}) \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|}$$

If  $\vec{a} \cdot (\hat{l} \times \hat{m}) = \vec{b} \cdot (\hat{l} \times \hat{m})$ , then the numerator  $(\vec{b} - \vec{a}) \cdot (\hat{l} \times \hat{m}) = 0$  equals to zero such that the two lines intersect. In this case, the values of  $s$  or  $t$  at the point of intersection ( $\vec{u} = \vec{v}$ ) can be found by applying a cross product with  $\hat{m}$  on  $\vec{u} = \vec{a} + \hat{l}s = \vec{b} + \hat{m}s = \vec{v}$  and note that  $\hat{m} \times \hat{m} = \vec{0}$ , and hence

$$(\vec{a} + \hat{l}s) \times \hat{m} = (\vec{b} + \hat{m}s) \times \hat{m}$$

$$\begin{aligned}\vec{a} \times \hat{m} + s(\hat{l} \times \hat{m}) &= \vec{b} \times \hat{m} + s(\hat{m} \times \hat{m}) = \vec{b} \times \hat{m} + s\vec{0} \\ s(\hat{l} \times \hat{m}) &= (\vec{b} - \vec{a}) \times \hat{m}\end{aligned}$$

$s$  is then inferred from the scaling ratio of  $(\vec{b} - \vec{a}) \times \hat{m}$  to  $(\hat{l} \times \hat{m})$ .  $t$  is found similarly.

### Exercise 5.5

$$\begin{aligned}\frac{1}{2} \|\vec{a} \times \vec{b}\| &= \frac{1}{2} \|\vec{b} \times \vec{c}\| = \frac{1}{2} \|\vec{c} \times \vec{a}\| \\ \rightarrow \frac{1}{2} \|\vec{a}\| \|\vec{b}\| \sin C &= \frac{1}{2} \|\vec{b}\| \|\vec{c}\| \sin A = \frac{1}{2} \|\vec{c}\| \|\vec{a}\| \sin B \\ \rightarrow \frac{\sin A}{a} &= \frac{\sin B}{b} = \frac{\sin C}{c}\end{aligned}$$

where we divide the entire equality by  $abc = \|\vec{a}\| \|\vec{b}\| \|\vec{c}\|$ .

**Exercise 5.6** It is just  $\frac{1}{6}|(\vec{u} \times \vec{v}) \cdot \vec{w}|$ .

### Exercise 5.7

$$(a) \vec{u} \times \vec{v} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ 1 & 2 & 3 \\ 2 & 1 & 5 \end{vmatrix} = 7\hat{i} + \hat{j} - 3\hat{k}$$

$$\text{Area} = \sqrt{7^2 + 1^2 + (-3)^2} = \sqrt{59}$$

$$(b) \text{Volume is the absolute value of } |\vec{u} \times \vec{v}| \cdot \vec{w} = |(7\hat{i} + \hat{j} - 3\hat{k}) \cdot (\hat{i} + 4\hat{j})| = |(7)(1) + (1)(4) + (-3)(0)| = 11$$

$$(c) \text{Volume} = \text{abs} \begin{vmatrix} 1 & 2 & 3 \\ 2 & 1 & 5 \\ 1 & 5 & 4 \end{vmatrix} = 0.$$

So the three vectors are co-planar.

### Exercise 5.8

- (a) The solution refers to the point  $(1, 1, 0)$ .  
(b) By Gaussian Elimination, one possible form of general solution is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ -\frac{5}{3} \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{3} \\ -\frac{5}{3} \\ 1 \end{bmatrix}$$

Therefore, the solution space is a line parallel to  $-\frac{1}{3}\hat{i} - \frac{5}{3}\hat{j} + \hat{k}$  and passing through the point  $(\frac{2}{3}, -\frac{5}{3}, 0)^T$ .

# Index

---

- Active Transformation, 239  
Adjugate, 66  
Algebraic Multiplicity, 299  
Argand Plane, 259  
Argument, 259  
Automorphism, 242  
  
Bijective, 230  
  
Canonical Quadratic Form, 389  
Cartesian Coordinate System, 103  
Cauchy–Schwarz Inequality, 114  
Cayley–Hamilton Theorem, 307  
Central Conics, 393  
Characteristic Equation, 298  
Characteristic Polynomial, 299  
Circular Convolution, 524  
Circular Convolution Theorem, 526  
Co-planar, 141  
Cofactor, 59  
Cofactor Expansion, 59  
Column Space, 166  
Column-row Factorization, 175  
Complement, 191  
Complementary Solution, 82  
  
Complex  $n$ -space, 269  
Complex Conjugate, 256  
Complex Discrete Fourier Transform, 512  
Complex Number, 255  
Complex Plane, 259  
Components, 101  
Congruent, 387  
Conjugate Transpose, 265  
Consistent, 78  
Convolution, 522  
Convolution Theorem, 526  
Coordinate Basis, 182  
Correlation, 405  
Cosine Law, 115  
Cosine Law for Spherical Trigonometry, 145  
Covariance, 403  
Covariance Matrix, 407  
CR Factorization, 175  
Cross Product, 116  
  
De Moivre’s Formula, 261  
Definiteness, 384  
Dependence Relation, 170  
Determinant, 57  
Diagonal Matrix, 309

## *Index*

---

- Diagonalization, 309  
Dimension, 188  
Direct Sum, 190  
    Matrix Direct Sum, 285  
Dot Product, 108
- El Niño–Southern Oscillation (ENSO), 423  
Elementary Matrix, 53  
Elementary Row Operations, 24  
Empirical Orthogonal Functions, 409  
Endomorphism, 241  
Euclidean  $n$ -space, 429  
Euclidean Norm, 106  
Euler’s Formula, 259  
Expected Value, 401  
Explained Variance, 413
- Fast Fourier Transform, 531  
Finite-dimensional, 188  
First-order Linear ODE, 323  
Fourier Basis, 443  
Fourier Coefficients, 454  
Fourier Series, 453  
Free Variable, 81
- Gaussian Elimination, 48  
    Gauss-Jordan Elimination, 48  
General Solution, 82  
Generating Set, 166  
Geometric Multiplicity, 299  
Gram-Schmidt Orthogonalization, 243
- Hermitian, 439  
Hermitian Matrix, 265  
Hermitian Transpose, 265  
Hilbert Space, 433  
Hyperplane, 135
- Identity Mapping, 229  
Identity Matrix, 39  
Identity Transformation, 229  
Imaginary Axis, 259  
Imaginary Number, 255  
Imaginary Part, 256  
Inconsistent, 78  
Infinite-dimensional, 188  
Injective, 221  
Inner Product, 428  
Inner Product Space, 428  
Invariant Subspaces, 306  
Inverse, 43  
Invertible, 43  
Isomorphic, 230  
Isomorphism, 234
- Kernel, 226
- Lagrange Multiplier, 410  
Laplace Expansion, 59  
Least-square Approximation, 473  
Left Null Space, 206  
Length, 106  
Linear Combination, 161  
Linear Equation, 21  
    Homogeneous Linear Equation, 22
- Linear Operator, 241  
Linear Regression, 479

- Linear System of Equations, 21  
Homogeneous Linear System of Equations, 23
- Linear Transformation, 213  
Linear Mapping, 213  
Matrix Representation of Linear Transformation, 215
- Linearly Dependent, 167
- Linearly Independent, 167
- Magnitude, 106
- Matrix, 13  
Augmented Matrix, 24  
Coefficient Matrix, 23  
Square Matrix, 14
- Matrix Invariants, 315
- Matrix Product (Matrix Multiplication), 16
- Mean, 401
- Minimal Generating Set, 175
- Minor, 59
- Modulus, 259
- Negative-Definite, 384
- Norm, 430
- Normal, 369
- Normal Equation, 474
- Normal Vector, 132
- Null Space, 202
- Nullity, 202
- Nyquist Frequency, 508
- Nyquist Sampling Theorem, 508
- One-to-one, 221
- Onto, 222
- Ordinary Differential Equation (ODE), 323
- Orthogonal, 114
- Orthogonal Complement, 207
- Orthogonal Diagonalization, 351
- Orthogonal Matrix, 342
- Orthogonal Projection, 359
- Orthonormal, 245
- Orthonormal Matrix, 342
- Overdetermined, 84
- Partial Differential Equation (PDE), 456
- Particular Solution, 82
- Passive Transformation, 239
- Pivot, 49  
Pivoting, 49
- Population Variance, 401
- Positive-Definite, 384
- Power Spectrum, 519
- Principal Axes, 412
- Principal Axes Theorem, 399
- Principal Component Analysis, 409
- Principal Components, 413
- Principal Directions, 412
- Pythagoras' Theorem, 106
- QR Decomposition, 247
- Quadratic Form, 381
- Radix-2 Algorithm, 531
- Range, 226
- Rank, 198
- Rank-nullity Theorem, 203
- Real  $n$ -space, 103

## *Index*

---

- Real Axis, 259  
Real Discrete Fourier Transform, 509  
Real Part, 255  
Resolution of the Identity, 367  
Restriction of a Linear Transformation, 280  
Right Hand Rule, 116  
Row Echelon Form, 46  
Reduced Row Echelon Form, 46  
Row Equivalent, 51  
Row Space, 197  
Sarrus' Rule, 57  
Scalar Multiplication, 15  
Scalar Product, 108  
Scalar Projection, 136  
Scalar Triple Product, 139  
Schur complement, 276  
Self-adjoint, 438  
Semidefiniteness, 384  
Signature, 391  
Similar, 242  
Singular, 43  
Skew-symmetric Matrix, 42  
Spanning Set, 166  
Special Polynomials, 461  
Spectral Decomposition, 367  
Spectral Theorem, 365  
Spectrum, 367  
Standard Basis, 182  
Standard Inner Product, 429  
Standard Unit Vector, 103  
Sturm-Liouville Equation, 456  
Sturm–Liouville operator, 456  
Subspace, 157  
Improper Subspace, 159  
Proper Subspace, 159  
Zero Subspace, 159  
Subspace Sum, 189  
Surjective, 222  
Sylvester's Law of Inertia, 391  
Symmetric Matrix, 42  
System of ODEs, 324  
The Four Fundamental Subspaces, 207  
Transformation, 213  
Mapping, 213  
Transpose, 41  
Underdetermined, 84  
Unitary Diagonalization, 368  
Unitary Matrix, 368  
Vandermonde Matrix, 484  
Variance, 400  
Vector, 101  
Column Vector, 102  
Geometric Vector, 101  
Row Vector, 102  
Unit Vector, 107  
Vector Product, 116  
Vector Projection, 136  
Vector Space, 155  
Real Vector Space, 155  
Vector Triple Product, 143

# Bibliography

---

- [1] H. Anton and C. Rorres. *Elementary Linear Algebra: Applications Version, 11th Edition*. John Wiley & Sons Incorporated, 2013. ISBN: 9781118879160.
- [2] S.H. Friedberg, A.J. Insel, and L.E. Spence. *Linear Algebra, 5th Edition*. Pearson Education, 2018. ISBN: 9780136745495.
- [3] K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*. Cambridge University Press, 2006. ISBN: 9781139450997.
- [4] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2022. ISBN: 9781733146678.