Introduction to Linear Algebra

with Earth Science Applications

First Edition

C.L. Loi

A student from
CUHK-EESC/NTU-AS

Introduction to Linear Algebra with Earth Science Applications

Copyright ©, C.L. Loi, 2025

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

First Edition, e-book, Self-published (May 2025)

Preface

This is a Linear Algebra textbook specifically designed for students who are majoring in any Earth Science-related subjects like Geophysics and Atmospheric Sciences, and are also interested in Mathematics. With these target readers in mind, we set out to provide an adequate treatment of 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.

C.L. Loi

Acknowledgement: This work is also partially supported by the grants from NSTC (112-2123-M-002-002, 113-2639-M-002-008-ASP) as the author is working at TDRC, NTU-AS.

The tex/pdf files and Python scripts can be found on the GitHub repository: https://github.com/BenjaminGor/Intro_to_LinAlg_Earth.

Forewords

It is my pleasure to recommend this special textbook to you. I appreciate this book from both a personal level and from a professional perspective.

C. L. Loi, the author of this book, was my Master's degree graduate student and is currently a member of the research staff in my research group, Typhoon Dynamics Research Center. In addition to his excellent research work on using machine learning to enhance the understanding of tropical cyclogenesis and to investigate the predictability of tropical cyclone density in the subseasonal-to-seasonal reforecast, C. L. dedicated tremendous time and efforts in the preparation and writing of this textbook over the past two years. I am truly impressed with his perseverance and the quality of his work.

Linear Algebra plays a crucial role in the study of Earth Science as a mathematical tool to analyze and model natural and simulated phenomena, from geophysical and geomechanics studies, remote sensing to climate modeling and beyond. I believe that readers, especially students interested in Earth Science problems, would find this textbook a clear and in-depth guide that is highly applicable to their studies. Below are some of the values that I find in this book and trust would benefit you as a reader.

- A very solid and comprehensive review of Linear Algebra,
- Close examination of the core concepts and detailed proofs,
- Connections to Earth Science applications, such as

- Empirical Orthogonal Functions to decompose El Niño-Southern Oscillation (ENSO) modes;
- Special polynomial solutions for equatorial waves;
- Fourier analysis of Sea Surface Temperature (SST) time series;
- Singular Value Decomposition (SVD) for data compression;
- Investigation of dynamical systems like the Lorenz-63 model; as well as
- Continuum Mechanics and Data Assimilation.

Overall, this is a high-quality, uncommon effort from a very hard-working researcher, intended for Earth Science students but can also benefit those who are looking to learn the fundamental concepts and applications of Linear Algebra. I highly recommend this book and hope you enjoy the reading and grow in your learning!

Chun-Chieh Wu
Chair Professor, Department of Atmospheric Sciences,
National Taiwan University

This book, compiled by our alumni Benjamin Loi, delves into the application of Linear Algebra in the field of Earth Sciences and provides a solid mathematical foundation along with practical insights. As one of the core areas of mathematics, Linear Algebra plays a crucial role in various scientific disciplines, particularly in Earth and Atmospheric Sciences. The purpose of this book is to help readers establish a strong foundation in Linear Algebra and guide them on how to effectively apply these mathematical concepts to real-world problems in Earth Sciences.

In addition to a thorough exploration of the mathematical theories, the author also includes many Python programming tutorials, which are immensely helpful for both academic research and practical applications. Whether you are a course instructor or an undergraduate student, this book offers valuable resources to help enhance your skills in both academic and practical settings.

The content of this book can also serve as a textbook or tutorial material for course design or teaching support. It is not only valuable to scholars in Earth Sciences but also to students interested in mathematics and programming, offering a challenging and inspiring learning resource.

We hope that readers will find inspiration in this book and apply the theories and methods of Linear Algebra to real-world problems. We believe this will help them achieve greater success in their academic or professional careers.

Man Nin Chan
Associate Professor, Department of Earth and Environmental Sciences,
The Chinese University of Hong Kong

Preface (the Heartfelt Version)

I have toiled for almost one-and-a-half years to deliver this work. This project started from a small Linear Algebra workshop I gave at CUHK-EESC (previously ESSC) in 2021. The initial draft was only 200 pages long and very immature (same as me back then, lol!). After finishing my Master's degree in 2023 at NTU-AS, I decided to resume the writing of this book. Little did I know at that moment I had taken on a very long and epic journey. In the early beginning, I advanced very quickly, but soon got stuck in the "Saggy Middle". It was absolutely no easy task to produce a Mathematical textbook with 18 chapters, and merely looking at this goal could be disheartening and overwhelming. Sometimes, I struggled to produce even only one page a day. More often than not, I considered the option of giving up, wondering if I really could see the finishing of this work. However, thanks to my stubbornness, I was too proud to kneel: I had a burning desire to finish what I started and show it to the world. Perhaps deep in my heart, I really wanted to get praised for my effort and intelligence, to prove to others that I was a force to be reckoned with. Don't get me wrong, I am still aiming for these rewards. However, along my writer's journey, I have also earned and learned far more precious things and important lessons.

I have become much more resilient and persevering. I disciplined myself to write (almost) every day despite all the ups and downs happening in my life. Many less-than-happy and discouraging events occurred to me, e.g. relationships and personal well-being, but I have weathered the storm and come back stronger every time.

I have also become more grateful for the supportive people around me. Without their help, this book (as well as me) would never have survived and got to this stage. I guess one of the reasons why I persisted in completing this book was that I did not want to let them down and must return their favors.

Finally, I have realized my true potential and determination. I am much more capable than I ever imagined. For a long time, I was pathetic, but now I have found my passion and dream. I confronted my fear whenever I wrote. I am no longer a little shepherd: I am now a mighty, wise ruler who has managed to overcome Goliath! (As an afterthought, all of these are fully due to God's grace.)

I wholeheartedly hope that as you read this book, apart from learning the knowledge of Linear Algebra, you will feel my soul and excitement, appreciate the enormous efforts that were put into it, and have a glimpse of my growth and transformation. So, let me conclude this with a quote.

"Every secret of a writer's soul, every experience of his life, every quality of his mind is written large in his works."
– Virginia Woolf

Sincerely,
Benjamin

Dedications

Dear Mom and Dad:

Thank you for bringing me to life and nurturing me.

To Prof. Chun-Chieh Wu:

Thank you for bestowing an infinite amount of patience on me.

To Tonybill Chan:

You are the messenger of God and
I appreciate your wisdom.

To King Lau, Paul Yeung, and Felix Chiang:

Thank you for the companionship as always.

To my Counselor M.H.:

Thank you for supporting me whenever things happen.
I hope that you will be happy about my growth.

To Prof. Kai-Chih Tseng:

You are the best mentor in the world.
I hope that I can proudly stand alongside you
in the coming future.

Contents

1. Introduction to Matrices and Linear Systems **27**
1.1. Definition and Operations of Matrices 28
1.1.1. Basic Structure of Matrices 28
1.1.2. Matrix Operations 29
1.2. Definition of Linear Systems of Equations 36
1.3. Elementary Row Operations 39
1.4. Earth Science Applications 41
1.5. Python Programming 48
1.6. Exercises 51

2. Inverses and Determinants **55**
2.1. Identity Matrices and Transpose 55
2.1.1. Identity Matrices 55
2.1.2. Transpose 57
2.1.3. Symmetric Matrices 59
2.2. Inverses 60
2.2.1. Definition and Properties of Inverses 60
2.2.2. (Reduced) Row Echelon Form 64
2.2.3. Finding Inverses by Gaussian Elimination 70
2.3. Determinants 75
2.3.1. Computing Determinants 75
2.3.2. Properties of Determinants 81
2.3.3. Finding Inverses by Adjugate 85
2.4. Python Programming 88

2.5. Exercises 92

3. Solutions for Linear Systems **95**
3.1. Number of Solutions for Linear Systems 95
3.2. Solving Linear Systems 99
3.2.1. Solving Linear Systems by Gaussian Elimination . . . 99
3.2.2. Solving Linear Systems by Inverse 109
3.3. Earth Science Applications 114
3.4. Python Programming 120
3.5. Exercises 122

4. Introduction to Vectors **127**
4.1. Definition and Operations of Geometric Vectors 127
4.1.1. Basic Structure of Vectors in the Real n-space $\mathbb{R}^n$ . . . 127
4.1.2. Fundamental Vector Operations 131
4.2. Special Vector Operations 135
4.2.1. Dot Product 135
4.2.2. Cross Product 144
4.3. Earth Science Applications 149
4.4. Python Programming 151
4.5. Exercises 153

5. Vector Geometry **157**
5.1. Lines and Planes 157
5.1.1. Translating Equation Form to Vector Form 158
5.1.2. Recovering Equation Form from Vector Form 160
5.1.3. Generalizing to Higher Dimensions 161
5.2. Further Geometric Applications of Dot Product 164
5.2.1. Projection 164
5.2.2. Distance 165
5.3. Further Geometric Applications of Cross Product 167
5.3.1. Area 167

5.3.2. Volume 168
5.4. Useful Vector Identities 172
5.5. Earth Science Applications 176
5.6. Python Programming 179
5.7. Exercises 181

6. Vector Spaces and Coordinate Bases 185
6.1. Making of the Real n-space $\mathbb{R}^n$ 186
6.1.1. $\mathbb{R}^n$ as a Vector Space 186
6.1.2. Subspaces of $\mathbb{R}^n$ 187
6.1.3. Span by Linear Combinations of Vectors 191
6.1.4. Linear Independence, CR Factorization 198
6.2. Coordinate Bases for $\mathbb{R}^n$ and its Subspaces 212
6.2.1. Coordinate Bases for $\mathbb{R}^n$ 212
6.2.2. Coordinate Bases for Subspaces of $\mathbb{R}^n$ 216
6.2.3. Direct Sum Representation 221
6.3. The Four Fundamental Subspaces Induced by Matrices 230
6.3.1. Row Space, Column Space 230
6.3.2. Null Space, Rank-Nullity Theorem 235
6.4. Python Programming 242
6.5. Exercises 244

7. More on Coordinate Bases, Linear Transformations 247
7.1. Ideas of Linear Transformations 247
7.1.1. Linear Transformations between Vector Spaces 247
7.1.2. One-to-one and Onto, Kernel and Range 256
7.1.3. Composition of Linear Transformations 264
7.1.4. Vector Space Isomorphism to $\mathbb{R}^n$ 266
7.2. Additional Discussions about Coordinate Bases 273
7.2.1. Linear Change of Coordinates 273
7.2.2. Gram-Schmidt Orthogonalization, QR Decomposition 280
7.3. Python Programming 288

7.4. Exercises 290

8. Complex Vectors/Matrices and Block Matrices 293

8.1. Definition and Operations of Complex Numbers 294

8.1.1. Basic Structure of Complex Numbers 294

8.1.2. Complex Number Operations 295

8.1.3. Geometric Meaning of Complex Numbers 297

8.2. Complex Vectors and Complex Matrices 301

8.2.1. Operations and Properties of Complex Vectors 301

8.2.2. Operations and Properties of Complex Matrices 303

8.2.3. The Complex n-space $\mathbb{C}^n$ 309

8.3. Manipulating Block Matrices 310

8.3.1. Block Matrix Multiplication 311

8.3.2. Inverse and Determinant of a Block Matrix 314

8.3.3. Restriction of a Linear Transformation, Direct Sum of a Matrix 320

8.4. Python Programming 329

8.5. Exercises 332

9. Eigenvalues and Eigenvectors 335

9.1. Eigenvalues and Eigenvectors of a Square Matrix 335

9.1.1. Definition of Eigenvalues and Eigenvectors 335

9.1.2. Finding Eigenvalues and Eigenvectors with Characteristic Polynomials 338

9.1.3. Eigenspace as an Invariant Subspace 347

9.1.4. Cayley-Hamilton Theorem 348

9.2. Diagonalization 350

9.2.1. Mathematical Ideas of Diagonalization 350

9.2.2. Diagonalization for Real Eigenvalues 352

9.2.3. Properties of Diagonalization 356

9.2.4. Diagonalization for Complex Eigenvalues 360

9.3. System of Ordinary Differential Equations 366

9.4. Earth Science Applications 376
9.5. Python Programming 380
9.6. Exercises 383

10. Orthogonal and Normal Matrices **387**
10.1. Orthogonal Matrices 388
10.1.1. Definition of Orthogonal Matrices 388
10.1.2. Geometric Implications of Orthogonal Matrices 391
10.2. Orthogonal Diagonalization 398
10.3. Orthogonal Projections and Spectral Theorem 405
10.3.1. Projections onto a Subspace 405
10.3.2. Orthogonal Projections 407
10.3.3. Spectral Theorem 415
10.4. Normal Matrices and Unitary Diagonalization 417
10.4.1. Unitary and Normal Matrices 417
10.4.2. Unitary Diagonalization 419
10.5. Python Programming 426
10.6. Exercises 428

11. Quadratic Forms **433**
11.1. Mathematical and Geometric Ideas of Quadratic Forms 433
11.1.1. Definition of Quadratic Forms 433
11.1.2. (Semi)Definiteness and Congruence 437
11.1.3. Conic Sections 445
11.1.4. Hermitian Forms 453
11.2. Statistics with Quadratic Form 455
11.2.1. Variance and Covariance 455
11.2.2. Principal Component Analysis (PCA) 464
11.3. Python Programming 472
11.4. Earth Science Applications: Empirical Orthogonal Functions (EOFs) 477
11.5. Exercises 480

12. Inner Product Spaces **483**
12.1. Definition and Properties of Inner Product Spaces 484
12.1.1. Requirements of Inner Products 484
12.1.2. Generalization of Length and Orthogonality via Inner Products 486
12.1.3. Infinite-dimensional Inner Product Spaces 488
12.2. Adjoints and Hermitian/Unitary Operators 492
12.2.1. Definition of Adjoints 492
12.2.2. Hermitian Operators 496
12.2.3. Unitary Operators 501
12.3. Revisiting Orthogonal Projections 503
12.3.1. Orthogonal Projections for an Inner Product Space . . 503
12.3.2. Revisiting Gram-Schmidt Orthogonalization 506
12.3.3. Spectral Theorem for Hermitian Operators 510
12.4. Special Polynomials 515
12.4.1. Sturm-Liouville Equations 515
12.4.2. Generating Special Polynomials by Gram-Schmidt Orthogonalization 521
12.5. Earth Science Applications 525
12.6. Python Programming 529
12.7. Exercises 531

13. Least-Square Approximation **535**
13.1. Mathematical Ideas of Least-Square Approximation 535
13.2. Linear Regression 542
13.2.1. Linear Regression for One Predictor Variable 542
13.2.2. Linear Regression for Multiple Predictor Variables . . 545
13.2.3. Properties of Linear Regression 548
13.3. Earth Science Applications 551
13.4. Python Programming 552
13.5. Exercise 553

14. Discrete Fourier Transform (DFT) **557**
14.1. Mathematical Ideas of DFT 558
14.1.1. From Fourier Series to a Prototype of DFT 558
14.1.2. Nyquist Frequency and Real DFT 571
14.1.3. Complex DFT 574
14.1.4. Inverse DFT 579
14.2. Properties of DFT 583
14.2.1. Power Spectrum and Parseval's Theorem 583
14.2.2. Convolution Theorem 588
14.3. Fast Fourier Transform (FFT) 596
14.4. Python Programming and Earth System Applications 601
14.5. Exercise 604

15. Markov Chains **607**
15.1. Statistical Prerequisites for Markov Chains 607
15.1.1. Lagged Auto-correlation 607
15.1.2. Conditional Probabilities, Stochastic Matrices 610
15.2. Construction of and Prediction by Markov Chains 613
15.2.1. State Vector, Steady State 613
15.2.2. Expected Value Problem 620
15.2.3. Lagged Auto-correlation Predicted by Markov Chains 623
15.3. Python Programming and Earth System Applications 628
15.4. Exercise 630

16. Matrix Factorization Methods **633**
16.1. Square Matrix Factorization 634
16.1.1. Cholesky Factorization 634
16.1.2. LU/LDU Factorization 637
16.1.3. Solving Linear Systems with LU Decomposition . . . 643
16.2. Singular Value Decomposition (SVD) 646
16.2.1. Mathematical Ideas of SVD and its Computation . . . 646
16.2.2. Geometric Perspective of SVD 655

16.2.3. Compact SVD 655
16.2.4. Relation between SVD and PCA 660
16.3. Pseudoinverses 662
16.3.1. Definition of Pseudoinverses and Inference from SVD 662
16.3.2. Minimal Solution 665
16.4. Generalization of SVD to Alternative Inner Products 670
16.5. Earth Science Applications 675
16.6. Python Programming 677
16.7. Exercises 680

17. Dynamical Systems **683**
17.1. Linear Dynamical Systems 684
17.1.1. Types of Equilibrium Points in Two-dimensional Linear Dynamical Systems 684
17.1.2. Generalization to Time-independent Forcings and Higher-dimensional Linear Dynamical Systems 690
17.2. Non-linear Dynamical Systems 694
17.2.1. Heuristic Understandings of Stability and Attractiveness 694
17.2.2. Linearization and Hartman-Grobman Theorem 696
17.3. Earth Science Applications: the Lorenz-63 Model 700
17.3.1. Derivation of the Lorenz-63 Model 701
17.3.2. Stability Analysis of Each Equilibrium Point 709
17.4. Python Programming 714
17.5. Exercises 718

18. Index Notation and Introduction to Tensors **721**
18.1. Index Notation and Tensors 722
18.1.1. Rules of Index Notation 722
18.1.2. Kronecker Delta and Epsilon Symbols 724
18.1.3. Transformation and Quotient Laws for Tensors 730
18.1.4. Isotropic Tensors 734

18.2. Matrix/Tensor Calculus 737
18.2.1. Differentiating Tensors 737
18.2.2. Tensor Integral Theorem 740
18.3. Earth System Applications: Continuum Mechanics 741
18.3.1. Stress Tensor and Mohr's Circle 741
18.3.2. Eulerian and Lagrangian Views, Spatial and Material Derivatives 751
18.3.3. Field Equations 756
18.4. Earth System Applications: Data Assimilation 767
18.4.1. Basic Ideas behind Data Assimilation 767
18.4.2. Optimal Interpolation 771
18.4.3. Kalman Filter 780
18.4.4. Python Programming: Demonstration of DA using the Lorenz-63 Model 785
18.5. Exercise 788

Answers to Exercises **791**

A. Supplementary Information for the Main Text **861**
A.1. Chapter 2 861
A.2. Chapter 6 863
A.3. Chapter 14 867
A.4. Chapter 15 870
A.5. Chapter 18 872

B. Cyclic Subspaces and Jordan Normal Forms **875**
B.1. From Cyclic Subspaces to Cayley-Hamilton Theorem 875
B.2. Jordan Normal Forms 879
B.2.1. Generalized Eigenvectors, Existence of Jordan Normal Forms 879
B.2.2. Methods to Compute Jordan Normal Forms 884
B.3. Matrix Exponentials 892

Index **899**

Bibliography **909**

Chapter 1

Introduction to Matrices and Linear Systems

Although the Earth System is well-known to be dominated by 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 several scenarios. For instance, in Atmosphere Sciences, we often consider the so-called *perturbation equations*, which assume that deviations from the mean state are small enough to neglect quadratic terms. In general, Linear Algebra enables us to compute useful statistics and make complicated dynamics appearing in Earth Science tractable, e.g. the simplification of the convection problem that leads to the three-variable *Lorenz-63 Model* in Chapter 17. The most fundamental usage of Linear Algebra in Applied Sciences is to formulate, analyze, and solve *linear systems of equations*. Some examples in Earth Science 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 describe such linear systems, and tactful matrix operations can help us derive their solution, in addition to being the machinery used to acquire many theoretical results later on. Therefore, given their importance, 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. A matrix with m rows and n columns is called an $m \times n$ matrix. Matrices that happen to have the same number of rows and columns, i.e. $m = n$, are known as ***square matrices***. Below are some examples of matrices.

$$\begin{bmatrix} 1.17 & 3 & -2.15 & 5 \\ 1.44 & 4 & 2.88 & -3 \end{bmatrix}$$

A 2×4 *real matrix.*

$$\begin{bmatrix} 1 - \sqrt{2}i \\ \frac{7}{8}i \\ 5 + 4i \\ \sqrt{3} \end{bmatrix}$$

A 4×1 *complex matrix.*

$$\begin{bmatrix} 3 & \frac{2}{7} & 9 \\ 0 & -4 & \frac{1}{10} \\ 5 & 2 & -1 \end{bmatrix}$$

A 3×3 *real, square matrix.*

For now, we will work with the simpler case of real matrices first. Given any matrix A, its entry at row i and column j will be denoted as A_{ij}. For example,

$$A = \begin{array}{c} \text{Col 1} \\ \begin{bmatrix} 2 & 1 & 7 & \frac{8}{9} \\ 5 & -\frac{1}{3} & 5 & 0 \\ -3 & \frac{4}{11} & 6 & -\frac{1}{6} \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{Row 2} \\ \\ \end{array} \qquad A_{21} = 5$$

► Short Exercise: Find A_{13}, A_{22}, A_{34} and A_{42}.[1]

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 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

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

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 $A - B$.[2]

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

[1] $A_{13} = 7$, $A_{22} = -\frac{1}{3}$, $A_{34} = -\frac{1}{6}$, A_{42} does not exist.

below.

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

► Short Exercise: Find $\frac{1}{4}A$.[3]

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 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 from the first/second input matrix. The following two examples explain this requirement.

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

---

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

Since the shapes of A and B are 2×3 and 3×1, 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 the shape 2×1. On the other hand, BA is not defined if we reverse the order of the matrix product. Meanwhile, for

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

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 valid, and what will be its shape?[4]) Now we are ready to see exactly how the entries in a matrix product are 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, which will have the shape of $m \times n$. To calculate any entry in AB at row i and column j, we select the i-th row from the first matrix A and j-th column from the second matrix B. Subsequently, take the products between the r pairs of numbers from that row and column, and their sum will then be the required value, i.e.

$$(AB)_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i3}B_{3j} + \ldots + A_{ir}B_{rj} \tag{1.1a}$$

$$= \sum_{k=1}^{r} A_{ik}B_{kj} \tag{1.1b}$$

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

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

[4] DC will be a 3×4 matrix.

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

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

Solution. The output will be a 2×2 matrix. Using Definition 1.1.1 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 in the second row can be found in a similar way, and the readers are 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, and 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.[5]

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

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

Solution.

$$\begin{aligned} -2A + 3B &= -2\begin{bmatrix} 1 & 6 & 9 \\ 4 & 4 & 6 \end{bmatrix} + 3\begin{bmatrix} 4 & 8 & 6 \\ -5 & 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} -2 & -12 & -18 \\ -8 & -8 & -12 \end{bmatrix} + \begin{bmatrix} 12 & 24 & 18 \\ -15 & 0 & 9 \end{bmatrix} \\ &= \begin{bmatrix} 10 & 12 & 0 \\ -23 & -8 & -3 \end{bmatrix} \end{aligned}$$

□

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

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

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

$$(A + 3B)(2A - B) = A(2A - B) + (3B)(2A - B)$$

[5] A trivial exception is that $A = B$.

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

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

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

Hence, the final answer will be

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

$$= \begin{bmatrix} -16 & 2 \\ 64 & 82 \end{bmatrix}$$

□

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 in an equation requires the same matrix to be inserted to the left (or right) on both sides, respecting the order. So, for a matrix equation that looks like (assuming the shapes of matrices are compatible)

$$AB - C = DE + F \tag{1.2}$$

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

$$G(AB - C) \neq (DE + F)G$$
$$(AB - C)G \neq G(DE + F)$$

Doing successive matrix multiplications follows the same principle, step by step. Using the same example of Equation (1.2), given another matrix H, we note some valid outcomes:

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

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

However, be careful that cancellation on both sides may not be correct. If $AB = AC$, then we cannot conclude that $B = C$ for sure. 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 $x_1, x_2, \ldots$ or $x, y, \ldots$). 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, $y^{-5/2}$, or non-linear functions, including $\sin x$, e^y. For the case of 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

$$\sum_{j=1}^{n} a_j x_j = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n = h \tag{1.3}$$

where $x_1, x_2, \ldots, x_n$ are the unknowns/variables, while $a_1, a_2, \ldots, a_n$ 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.[6]

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 is 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 the shape $m \times n$, i.e. m linear equations in n unknowns $(x_1, x_2, \ldots, x_n)$, has the form of

$$\begin{cases} \sum_{j=1}^{n} a_j^{(1)} x_j = a_1^{(1)} x_1 + a_2^{(1)} x_2 + a_3^{(1)} x_3 + \cdots + a_n^{(1)} x_n & = h^{(1)} \\ \sum_{j=1}^{n} a_j^{(2)} x_j = a_1^{(2)} x_1 + a_2^{(2)} x_2 + a_3^{(2)} x_3 + \cdots + a_n^{(2)} x_n & = h^{(2)} \\ \qquad \vdots \\ \sum_{j=1}^{n} a_j^{(m)} x_j = a_1^{(m)} x_1 + a_2^{(m)} x_2 + a_3^{(m)} x_3 + \cdots + a_n^{(m)} x_n & = h^{(m)} \end{cases} \tag{1.4}$$

If $h^{(1)}, h^{(2)}, \ldots, h^{(m)}$ on R.H.S. are all zeros, i.e. all the equations are homogeneous, 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 $j = 1, 2 \ldots, n$, or expressed as $\vec{x} = \mathbf{0}$. However,

such a trivial solution may not be the only solution to the system, as we shall see in Chapter 3. Below are some examples of linear systems.

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

A 2×2 *linear system with two equations and two unknowns.*

$$\begin{cases} x + 2y - 4z = 3 \\ x - y + 3z \quad = -4 \end{cases} \tag{1.5}$$

A 2×3 *linear system with two equations and three unknowns.*

$$\begin{cases} x + 2.2y + 3z = 0 \\ 2x + 3z \qquad = 0 \\ 4x - 5.6y \qquad = 0 \end{cases} \tag{1.6}$$

A 3×3 *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 the 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 unknowns/variables in (1.4). 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.

[6] Linear/Inhomogeneous, Non-linear, Linear/Inhomogeneous, Non-linear, Linear/Homogeneous, Non-linear.

Properties 1.2.3. For a linear system defined like (1.4) in Definition 1.2.2, it can be rewritten as

$$A\vec{x} = \vec{h} \tag{1.7}$$

where $A_{ij} = a_j^{(i)}$, $\vec{x} = x_j$, and $\vec{h} = h^{(i)}$.

Using the second example (Equation (1.5)) 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 to concatenate 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 (1.6).[7]

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.

Definition 1.3.1 (Elementary Row Operations)**.** Denote the p-th row of a matrix by 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 will bookmark these three kinds of operations using the following notations.

1. $cR_p \to R_p$;
2. $R_p + cR_q \to R_p$;
3. $R_p \leftrightarrow R_q$

For example, the matrix A

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

if we apply the elementary row operation of subtracting $2R_1$ from R_2 (i.e. $R_2 - 2R_1 \to R_2$) on it, can be transformed to a new matrix A'

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

► 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.[8]

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

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

solving a 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 modeled 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 $\theta_1 = 45°/\theta_2 = 60°$ to the vertical at the top/bottom layer. (θ_1 can be found by analyzing the seismic waveform, and then θ_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) at which the wave reaches the interface relative to the source.

Solution. We can deduce two equations from the given information. Considering 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

$$\frac{x}{h-y} = \tan\theta_2$$

[8] $\begin{bmatrix} 0 & 3 & 4 \\ 3 & 3 & 5 \end{bmatrix}$

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

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

The corresponding linear system is

$$\begin{cases} x + (\tan\theta_1)y = d \\ x + (\tan\theta_2)y = (\tan\theta_2)h \end{cases} \tag{1.8}$$

where x and y are the unknowns to be solved. d, h, θ_1 and θ_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 ($\tan\theta_1 = \tan(45°) = 1$, $\tan\theta_2 = \tan(60°) = \sqrt{3}$), 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

$$E_j = \sigma T_j^4 \tag{1.9}$$

according to the *Stefan-Boltzmann Law*, with $\sigma = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$. The overall scenario can be seen in Figure 1.2. Formulate a linear system that represents the energy equilibrium (incoming radiation balancing outgoing radiation) of all layers and the surface, with E_j being the unknowns, over $j = 1, 2, \ldots, N, N+1$.

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

$$E_{j-1} - 2E_j + E_{j+1} = 0 \tag{1.10}$$

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

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

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.*

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

$$\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{N-1} \\ E_N \\ E_{N+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ -E_{\text{in}} \end{bmatrix} \tag{1.11}$$

Particularly, for $N = 4$, it is

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

□

Example 1.4.3. The seawater in oceans contains a variety of dissolved salts in the form of ions. Most of them are sodium (Na^+), magnesium (Mg^{2+}), chlorine (Cl^-) and sulphate (SO_4^{2-}). Consider a sample of seawater and assume the concentrations of other ions are negligible. There are two major constraints over the individual concentrations of each type of ions ($n = [Na^+]$, $m = [Mg^{2+}]$, $c = [Cl^-]$, $s = [SO_4^{2-}]$). 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 g kg^{-1} which is the unit preferred in oceanography), and their molar masses are Na^+: 23.0 g mol^{-1}, Mg^{2+}: 24.3 g mol^{-1}, Cl^-: 35.5 g mol^{-1}, SO_4^{2-}: 96.1 g mol^{-1}. Write down the corresponding linear system that consists of two equations for this situation.

Solution. The first constraint, charge neutrality, can be simply translated to

$$n + 2m - c - 2s = 0$$

While for the second constraint about salinity, we need to convert the molar concentration (*molarity*) of each ion species to mass concentration through multiplying it by their respective molar masses $(23.0, 24.3, 35.5, 96.1)$. This will lead to

$$23.0n + 24.3m + 35.5c + 96.1s = 34.0$$

Thus, the linear system will be just

$$\begin{cases} n + 2m - c - 2s & = 0 \\ 23.0n + 24.3m + 35.5c + 96.1s & = 34.0 \end{cases} \tag{1.12}$$

and its matrix form is

$$\begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \begin{bmatrix} n \\ m \\ c \\ s \end{bmatrix} = \begin{bmatrix} 0 \\ 34.0 \end{bmatrix}$$

□

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 air temperatures of the four stations at some time are $27.1\,°\mathrm{C}, 27.3\,°\mathrm{C}, 27.4\,°\mathrm{C}, 26.9\,°\mathrm{C}$. 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 Δx and Δy are the x/y-distances relative to the station at the origin. Substituting the provided data, we get

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

Reorganizing them gives

$$\begin{cases} 10\dfrac{\partial T}{\partial x} + 20\dfrac{\partial T}{\partial y} = 0.2 \\ 25\dfrac{\partial T}{\partial x} + 15\dfrac{\partial T}{\partial y} = 0.3 \\ -10\dfrac{\partial T}{\partial x} + 5\dfrac{\partial T}{\partial y} = -0.2 \end{cases} \tag{1.13}$$

The matrix form of this linear system will then be

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

□

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. The version of Python used throughout this book is `3.12.4` for reference. 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 straightforward.

```
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 the 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 first row of `myMatrix5` (which is a copy of `myMatrix4`) by 3, and give

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

Meanwhile, these two 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],:]
```

The list `[0, 1]` locates the first row followed by the second row in that order, and `[1, 0]` means that the rows are now selected, arranged from the second to first. This effectively swaps the first and second rows. (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 this 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_fl = np.array([[1, 4], [5, 3]], dtype=float)
```

when printed out via `print(myMatrix1_fl)` 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_fl = np.array([[1., 4.], [5., 3.]])
```

1.6. Exercises

Exercise 1.1 Let

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

Find:

(a) $A + B$;

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

(c) AB;

(d) BA.

Exercise 1.2 Let

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

Find:

(a) AB, and

(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

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

Find:

(a) $(A + B)(2A - B)$, and

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

Exercise 1.5 Let

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

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 to it:

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

(b) Adding the first row by 3 times the third row, and then interchanging the first and second row, and finally subtracting 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 $\Gamma_{dry} = 9.8\,°\mathrm{C\,km^{-1}}$. When the temperature of the air parcel falls below the *dew point*, the air saturates and condensation occurs. Typically, the dew point temperature of an air parcel will decrease at a rate of roughly $\Gamma_{dew} = 2\,°\mathrm{C\,km^{-1}}$. Now, an air parcel with an initial air temperature/dew point temperature of $T_{a,ini} = 25.4\,°\mathrm{C}$ / $T_{dew,ini} = 17.8\,°\mathrm{C}$ at the ground starts to rise. Let z_{cd} and T_{cd} be the height above the ground (in km) and temperature (in °C) of the air parcel when condensation occurs. Construct a linear system with z_{cd} and T_{cd} as the two 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: their *inverses* and *determinants*. They will appear from time to time in the remaining parts of this book. Inverses mainly arise when we have to "undo" the effect of multiplying a matrix, or solve linear systems of equations by inverting the coefficient matrix. In Earth Science, inverses, as the name suggests, appear frequently in the so-called *inversion problems* that require us to deduce unknown information from given known data in a backward manner. Meanwhile, determinants have the geometric meaning of being the magnification/shrinking factor in a transformation that we shall see in later chapters. To derive them, we first need to introduce some prerequisite ideas, 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 the row index equals column index) being 1 and other off-diagonal entries being 0. Usually, they are denoted by I_n, or simply by 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 2×2 *and* 3×3 *with the main diagonal* 1*s highlighted.*

Definition 2.1.1 (Identity Matrix)**.** An identity matrix I_n of the square shape $n \times n$ 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.[1]

One important property of identity matrices is

Properties 2.1.2. The matrix product between any matrix A with an identity matrix I always returns A whenever the matrix product is well-defined. If A is of the shape $m \times n$, then

$$AI_n = I_m A = A \tag{2.1}$$

If A is now a square matrix such that $m = n$ (and $I_m = I_n = I$), then we simply have

$$AI = IA = A \tag{2.2}$$

In other words, the identity I can be regarded as the equivalent of "1" in the world of matrices. This is one of the cases where $AB = BA$ commutes (if both

of them are square and either one 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 computed AI_3 and I_2A to see if the results are A itself.[2]

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 entries about the main diagonal.

Definition 2.1.3 (Transpose)**.** The transpose of an $m \times n$ matrix A, denoted as A^T, is constructed according to the relation

$$[A^T]_{pq} = A_{qp} \tag{2.3}$$

---

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

[2]

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

The calculation of $I_2A = A$ is similar.

where $1 \leq p \leq n$, $1 \leq q \leq m$, 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.

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

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

Particularly, in the second example, we have outlined the main diagonal (red dashed line) of B (as well as B^T) and how the entries are flipped about it (green dashed arrows) when a transpose is carried out. Some useful properties of 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 (notice the reversed order);
5. $A_{kk} = A^T_{kk}$ 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.[3]

2.1.3. Symmetric Matrices

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

Definition 2.1.5 (Symmetric Matrix)**.** If an $n \times n$ square matrix A and its transpose A^T are equal, i.e.

$$A_{pq} = [A^T]_{pq} = A_{qp} \tag{2.4a}$$

for all $1 \leq p, q \leq n$, or simply

$$A = A^T \tag{2.4b}$$

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 3×3 symmetric matrix.

► Short Exercise: Show that $Y = XX^T$ and $Z = X^T X$ are symmetric for any matrix X.[4]

In contrast, we also have ***skew-symmetric matrices*** such that

$$A^T = -A \tag{2.5}$$

[3]By (4), $(ABC)^T = ((AB)(C))^T = C^T(AB)^T = C^T B^T A^T$.

This automatically forces the components along the main diagonal in the matrix to be all zeros.

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

A 3×3 *skew-symmetric matrix.*

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

$$AB = BA = I_n \tag{2.6}$$

This inverse matrix is denoted as $B = A^{-1}$, and the relation becomes

$$AA^{-1} = A^{-1}A = I \tag{2.7}$$

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.

[4]By (4) of Properties 2.1.4, $Y^T = (XX^T)^T = (X^T)^T(X)^T = XX^T = Y$, the similar goes for $Z = X^TX$.

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.[5]

In the last chapter, we only define addition, subtraction, and multiplication for matrices, omitting division like 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 that the relevant inverses exist. Remember, in the last chapter, we mentioned that cancellation on both sides may not work for some equations like $AB = AC$. But if the inverse A^{-1} exists, then by multiplying it to the left on both sides of the equation, we can effectively "divide by A" to get

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

so 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:

$$AG = H$$

[5] $AA^{-1} = I$ implies $A^{-1}A = I$ and vice versa. However, while appearing 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 $A^{-1}A = I$. In fact, this is guaranteed if A is indeed invertible.

$$A^{-1}AG = A^{-1}H$$

$$((A^{-1}A)G = IG =)\ G = A^{-1}H \qquad \text{(Properties 1.1.2 and 2.1.2, Definition 2.2.1)}$$

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. □

Example 2.2.1. Let

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

Show that A and B are the inverse of each other.

Solution.

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

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

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

The following 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, again notice the reversed order);
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 the above equation respectively yields

$$ABB^{-1}A^{-1} = AIA^{-1}$$

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

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. □

► Short Exercise: Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ if A, B and C are invertible and conformable.[6]

(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.[7] 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) 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 consisting of all zeros is placed at the bottom;

4. If additionally, every column containing a leading 1 (sometimes called a *pivotal column*) has 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

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

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

Meanwhile, examples of *reduced* row echelon form are

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

The following matrices are *not* in row echelon form. (why?)[8]

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

[6] By (4), $(ABC)^{-1} = ((AB)(C))^{-1} = C^{-1}(AB)^{-1} = C^{-1}B^{-1}A^{-1}$.

[7] 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.

[8] P violates (2) and Q does not satisfy (1) and (3) of Definition 2.2.5.

► Short Exercise: Decide if the following matrices are in (reduced) row echelon form or not.[9]

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

We have studied 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 converting it to reduced row echelon form. The first phase is also named ***Gaussian Elimination***, and together they are called ***Gauss-Jordan Elimination***[10]. 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.

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

Solution. At each step of the forward phase, the strategy is to look at the leftmost column that has at least one non-zero entry (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 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.

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

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 entries below the leading 1 along that *pivotal column* become zero, by adding the top row (which now holds the leading 1), times $-a_i$ (where a_i is the corresponding leading entry in the i-th row) to the other rows. The zeros produced in this way will be highlighted in blue.

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

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

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

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

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

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

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.

[9] Yes, Yes (reduced), No.

[10] Often we just write Gaussian Elimination in place of Gauss-Jordan Elimination.

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 entries above the leading 1 along every pivotal column into zeros. Non-pivotal columns (the last column here) are ignored.

$$\begin{aligned}
\begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & -\frac{10}{3} & -6 \\ 0 & 0 & 1 & \frac{5}{3} \end{bmatrix} &\to \begin{bmatrix} 1 & \frac{1}{2} & 2 & 3 \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{bmatrix} && R_2 + \frac{10}{3}R_3 \to R_2 \\
&\to \begin{bmatrix} 1 & \frac{1}{2} & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{bmatrix} && R_2 - 2R_3 \to R_1 \\
&\to \begin{bmatrix} 1 & 0 & 0 & -\frac{1}{9} \\ 0 & 1 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & \frac{5}{3} \end{bmatrix} && R_1 - \frac{1}{2}R_2 \to R_1
\end{aligned}$$

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 3 here. □

► Short Exercise: Repeat the example above but start by interchanging R_1 and R_3.[11]

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)**.** The reduced row echelon form (RREF) of a matrix is unique.

We shall omit the proof here. The following property further reveals how elementary row operations are relevant to reduced row echelon forms.

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 of 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

$$\begin{bmatrix} 2 & 2 & 1 \\ 6 & 4 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{bmatrix} \to \begin{bmatrix} 6 & 4 & 1 \\ 2 & 2 & 1 \\ 2 & 3 & 2 \\ 2 & 1 & 0 \end{bmatrix} \qquad R_1 \leftrightarrow R_2$$

[11] For checking, after the first iteration, it will be

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -3 & -8 & -12 \\ 0 & -3 & -2 & -2 \end{bmatrix}$$

and the result will be the same.

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

$$\rightarrow \begin{bmatrix} 1 & \frac{2}{3} & \frac{1}{6} \\ 0 & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{5}{3} & \frac{5}{3} \\ 0 & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} \qquad \begin{aligned} R_2 - 2R_1 &\rightarrow R_2 \\ R_3 - 2R_1 &\rightarrow R_3 \\ R_4 - 2R_1 &\rightarrow R_4 \end{aligned}$$

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

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

The backward elimination is straightforward.

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

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 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 on the left, originally as A, is reduced to the identity matrix I by Gaussian Elimination. The identity matrix on the right will then be transformed into the desired inverse by the same set of elementary row operations, and the concatenated matrix will now 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 a 3×3 identity matrix to the right, we have

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

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

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

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

$$\to \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -1 & -2 \\ 0 & 1 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{array}\right] \quad R_1 - 4R_2 \to R_1$$

Hence the required inverse is

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

□

► Short Exercise: Verify that 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***.

Definition 2.2.9 (Elementary Matrices)**.** Any elementary row operation on an $m \times n$ matrix can be represented by multiplying it with a suitable *elementary matrix* to the left. Such a matrix is essentially the one that appears 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. $cR_p \to R_p$, $c \neq 0$;
2. $R_p + cR_q \to R_p$;
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.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{Multiplying } R_2 \text{ by a factor of 2: } 2R_2 \to R_2$$

$$\begin{bmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{Adding 3 times } R_2 \text{ to } R_1\text{: } R_1 + 3R_2 \to R_1$$

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \text{Swapping } R_1 \text{ and } R_3\text{: } R_1 \leftrightarrow R_3$$

The third type of elementary matrices listed above is known as *permutation matrices*. 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.[13]

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, $R_3 - R_2 \to R_3$ can be expressed as

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

[12] You should be able to retrieve the matrix A back. The first column of A^{-1} already contains a leading 1, and the entries below it are zeros. A possible next step is to multiply R_2 by -1 and then subtract R_3 by R_2.

► Short Exercise: Find out the 3×3 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?[14]

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 A to I (i.e. the RREF 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 Definition 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 $i = 1, 2, \ldots, n$ 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

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

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, $i = 1, 2, \ldots, n$, reduce A to I as we demand in the assumption. With

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

$A^{-1} = E_n \cdots E_3 E_2 E_1 I$, 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. □

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 Definition 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 backward from n to 1. This leads to

$$\begin{aligned} E_1^{-1} E_2^{-1} E_3^{-1} \cdots E_n^{-1} E_n \cdots E_3 E_2 E_1 A &= E_1^{-1} E_2^{-1} E_3^{-1} \cdots E_n^{-1} I \\ A &= E_1' E_2' E_3' \cdots E_n' \end{aligned} \tag{2.8}$$

as each of the pairs $E_n^{-1} E_n$, $E_{n-1}^{-1} E_{n-1}$, ..., $E_2^{-1} E_2$, $E_1^{-1} E_1$ on L.H.S. 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 intrinsic geometrical properties of the matrix which can help us to find its inverse (Determinant of non-square matrices is undefined). The determinant of a 1×1 matrix is equal to the matrix's only entry. Meanwhile, determinants of 2×2 and 3×3 matrices can be calculated by a trick called ***Sarrus' Rule***.

---

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

Sarrus' Rule

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

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

Its determinant is computed by

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$|A| = a_{11}a_{22} - a_{21}a_{12} \tag{2.9}$$

which is the product of entries crossed by the red arrow, minus the blue one. Similarly, for a 3×3 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

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

$$\begin{aligned} |A| = & (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) \\ & - (a_{31}a_{22}a_{13} + a_{32}a_{23}a_{11} + a_{33}a_{21}a_{12}) \end{aligned} \tag{2.10}$$

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

$$C_{ij} = (-1)^{i+j} \det\left(M_{ij}\right) \tag{2.11}$$

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

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

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.[15]

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 the matrix.

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

$$|A| = A_{i1}C_{i1} + A_{i2}C_{i2} + \cdots + A_{in}C_{in} \tag{2.12a}$$

$$= \sum_{k=1}^{n} A_{ik} C_{ik} \tag{2.12b}$$

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

$$|A| = A_{1j}C_{1j} + A_{2j}C_{2j} + \cdots + A_{nj}C_{nj} \tag{2.13a}$$

$$= \sum_{k=1}^{n} A_{kj} C_{kj} \tag{2.13b}$$

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.[16]

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 (2.12) in Properties 2.3.3, if we choose the first row to be expanded, its determinant will be

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

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

[16] This is due to the uniqueness of the determinant as an "alternating k-linear form" which is an advanced topic out of our scope.

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

□

► Short Exercise: Confirm the answer by carrying out cofactor expansion on another row or column.[17]

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 4×4 matrix and we have to apply cofactor expansion. We can choose any 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

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

By Sarrus' Rule (Properties 2.3.1), we can calculate each of the four 3×3 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.

Properties 2.3.4. If a matrix has 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 of determinants. First of all, it is very easy to see that the determinant for any $n \times n$ identity matrix I_n is just 1. Second, there is a close relationship 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. $cR_p \longrightarrow R_p$, $c \neq 0$;
2. $R_p + cR_q \longrightarrow R_p$;
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 Definition 2.2.9 that correspond to the elementary row operations in Definition 1.3.1,

1. $E_{kk} = 1$ for any k, except $E_{pp} = c$ ($cR_p \longrightarrow R_p$, $c \neq 0$);
2. $E_{kk} = 1$ for all k, with $E_{pq} = c$ ($R_p + cR_q \longrightarrow R_p$);

[17] You should able to get $|A| = 0$, no matter which row/column is selected.

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.

We will prove the above properties for the second kind of elementary row operations/elementary matrices (corresponding to addition/subtraction) in Appendix A.1. The properties for the two other types of elementary matrices are easy to show and we will take them for granted. With these, we can easily establish the second case in Properties 2.3.4 by subtraction between proportional rows to reduce one of them to a row with full zeros.

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

Properties 2.3.7. For any elementary matrix E and a square matrix A, we have

$$\det(EA) = \det(E)\det(A) \tag{2.14}$$

This property 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 the second row and subtract 3 times the third row from the fourth row. By Properties 2.3.5, we have

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

The new determinant 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 Properties 2.3.7, we can unearth the relationship between the 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 $E_1, E_2, \ldots, E_{n-1}, E_n$, 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 Properties 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 for reference.[18] □

Other properties of determinants include:

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

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

By extension of (3), $\det(A_1 A_2 \cdots A_n) = \det(A_1)\det(A_2)\cdots\det(A_n)$.

For instance, if

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

then

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

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

[18] 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, they will have a determinant of zero.

$$
\begin{aligned}
&= \begin{bmatrix} (2)(4)+(3)(1) & (2)(5)+(3)(0) \\ (5)(4)+(9)(1) & (5)(5)+(9)(0) \end{bmatrix} \\
&= \begin{bmatrix} 11 & 10 \\ 29 & 25 \end{bmatrix} \\
|AB| &= (11)(25) - (10)(29) \\
&= -15 = (3)(-5) = |A||B|
\end{aligned}
$$

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.[19]

► Short Exercise: Prove (4) of Properties 2.3.9.[20]

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 with it.

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

$$[\operatorname{adj}(A)]_{pq} = (C_{pq})^T = C_{qp} \tag{2.15}$$

[19]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 $\det(AB)$ 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 $A = E_1E_2\cdots E_{n-1}E_n$ as a sequence of elementary matrices. By using Properties 2.3.7 back and forth, we have

$$
\begin{aligned}
\det(AB) &= \det(E_1E_2\cdots E_{n-1}E_nB) \\
&= \det(E_1)\det(E_2)\cdots\det(E_{n-1})\det(E_n)\det(B) && \text{(Properties 2.3.7)} \\
&= (\det(E_1)\det(E_2)\cdots\det(E_{n-1})\det(E_n))\det(B) \\
&= \det(E_1E_2\cdots E_{n-1}E_n)\det(B) && \text{(Properties 2.3.7)} \\
&= \det(A)\det(B)
\end{aligned}
$$

So the equality is true in both cases.

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 by

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A) \tag{2.16}$$

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

Example 2.3.5. For a 2×2 matrix

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

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

$$\operatorname{adj}(A) = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}^T = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

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

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{2.17}$$

[20]Consider $A^{-1}A = I$, and take determinant on both sides. By (3), we have

$$\det\left(A^{-1}A\right) = \det(I)$$

$$\det\left(A^{-1}\right)\det(A) = 1 \qquad \text{(The identity always has a determinant of 1)}$$

$$\det\left(A^{-1}\right) = \frac{1}{\det(A)}$$

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)

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

The adjugate matrix is

$$\begin{aligned} \operatorname{adj}(A) &= \begin{bmatrix} \begin{vmatrix} 3 & 5 \\ 4 & 11 \end{vmatrix} & -\begin{vmatrix} 1 & 5 \\ 1 & 11 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 1 & 4 \end{vmatrix} \\ -\begin{vmatrix} 2 & 3 \\ 4 & 11 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 1 & 11 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 1 & 4 \end{vmatrix} \\ \begin{vmatrix} 2 & 3 \\ 3 & 5 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 1 & 5 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 1 & 3 \end{vmatrix} \end{bmatrix}^T \\ &= \begin{bmatrix} 13 & -6 & 1 \\ -10 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix}^T = \begin{bmatrix} 13 & -10 & 1 \\ -6 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix} \end{aligned}$$

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

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

$$= \frac{1}{4}\begin{bmatrix} 13 & -10 & 1 \\ -6 & 8 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{13}{4} & -\frac{5}{2} & \frac{1}{4} \\ -\frac{3}{2} & 2 & -\frac{1}{2} \\ \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

□

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 the identity 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 `1.11022302e-16` (this number may be different in your machine). 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 compute its inverse by `print(linalg.inv(myMatrix3))`, we may obtain an absurd output like

```
[[ 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 magnifying 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* ϵ (the order of relative round-off error) of `float` and we want the absolute value of the determinant to be larger than that. Subsequently, calling `safe_inv(myMatrix3)` will print a warning. However, this behavior was observed in older versions and has been fixed in newer versions, which instead produces

```
LinAlgError: singular matrix
```

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 item is a tuple that 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.[21]

Exercise 2.2 Let

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

Verify:

(a) $(AB)^T = B^T A^T$,

(b) $(AB)^{-1} = B^{-1} A^{-1}$, and

[21] An upper[lower]-triangular matrix is a matrix whose elements below [above] the main diagonal are all zeros.

(c) $\det(AB) = \det(A)\det(B)$.

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

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

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 that any square matrix A can be written as the sum of a symmetric matrix and a skew-symmetric matrix with an explicit formula.

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

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

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 those methods (Gaussian Elimination and Inverse) under suitable circumstances. Remember, in the first chapter, we formulated some problems about linear systems of equations appearing in Earth Science, such as seismic tomography and radiative transfer in the atmosphere, and they will be solved accordingly.

3.1. Number of Solutions for Linear Systems

Before tackling any linear system, we may like to know how many solutions there are. 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. A 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, shown in Figure 3.1 below.

$$\begin{cases} a_1x + b_1y = h_1 \\ a_2x + b_2y = h_2 \end{cases} \tag{3.1}$$

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 in Figure 3.2. The readers can try to imagine and draw out other possibilities.

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

In fact, this theorem about the existence of solutions is true for any number of variables and equations.[1] If there is any solution, then the system is called ***consistent***. Otherwise, if no solution exists, then it is known as ***inconsistent***.

(b) *No solution: Two parallel lines never touch each other.*

(c) Infinitely many solutions: Two parallel lines overlap each other.

Figure 3.1.: *Three possible situations for (3.1).*

Clearly, the next task is 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.

Figure 3.2.: *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 pairwise along three non-intersecting parallel lines.*

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 a 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 (because it cannot have no solution). We defer the arguments for showing Theorem 3.1.2, as well as the discussion about non-square systems, until we actually start solving linear systems in the next subsection.

► Short Exercise: By inspection, determine the number of solutions for the

[1]Some readers may think if there can be finitely many solutions only. Unfortunately, it is 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.

following linear systems.[2]

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

3.2. Solving Linear Systems

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

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 Definition 2.2.9 and Theorem 2.2.10, but with $A\vec{x} = \vec{h}$ instead of $AA^{-1} = I$. Let A_{rref} be the reduced row echelon form of A, and $E_1, E_2, \ldots, E_n$ 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

$$\begin{aligned} (E_n \cdots E_2 E_1)A\vec{x} &= (E_n \cdots E_2 E_1)\vec{h} \\ (E_n \cdots E_2 E_1 A)\vec{x} = A_{\text{rref}}\vec{x} &= (E_n \cdots E_2 E_1)\vec{h} \end{aligned} \tag{3.2}$$

hence $\vec{x}$ will be the solution to $A_{\text{rref}}\vec{x} = \tilde{h}$ too at the same time, where $\tilde{h} = E_n \cdots E_2 E_1 \vec{h}$. Therefore, the solutions of $A\vec{x} = \vec{h}$ and $A_{\text{rref}}\vec{x} = \tilde{h}$ coincide,

which can be inferred more directly from the latter reduced 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 using 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 of reducing the coefficient matrix on the left to the identity matrix.

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 1 & 1 & 4 & 10 \\ 2 & 0 & 3 & 8 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 1 & 3 \\ 0 & 1 & 3 & 7 \\ 0 & 0 & 1 & 2 \end{array}\right] \qquad \begin{array}{r} R_2 - R_1 \to R_2 \\ R_3 - 2R_1 \to R_3 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{array}\right] \qquad \begin{array}{r} R_2 - 2R_3 \to R_2 \\ R_1 - R_3 \to R_1 \end{array}$$

which translates to

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

[2]The coefficient matrices of these two homogeneous linear systems have 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.

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

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

$$\left[\begin{array}{ccc|c} 3 & 7 & 2 & 8 \\ 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & 2 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 3 & 7 & 2 & 8 \\ 0 & 2 & 1 & 2 \end{array}\right] \qquad R_1 \leftrightarrow R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 4 & 2 & 2 \\ 0 & 2 & 1 & 2 \end{array}\right] \qquad R_2 - 3R_1 \to R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 2 & 1 & 2 \end{array}\right] \qquad \frac{1}{4}R_2 \to R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & 1 & 0 & 2 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{array}\right] \qquad R_3 - 2R_2 \to R_3$$

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

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

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

Solution. Gaussian Elimination leads to

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

$$\to \left[\begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_3 - R_2 \to R_3 \\ R_1 - 2R_2 \to R_1 \end{array}$$

Now, the last row corresponds to $0 = 0$, which is vacuous and implies that one equation is spurious. This also means that 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 the 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 that appears 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, they form the ***general solution*** $\vec{x}_g = \vec{x}_p + \vec{x}_c$ as the complete set of solutions to the linear system. □

► Short Exercise: Try plugging in any number t to the general solution and verify the consistency.[3]

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 pair of free parameters and column vectors (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 which satisfy $A\vec{x} = \mathbf{0}$ on their own. (We will clarify 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 1. In the 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 the full 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.[4]

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 Footnote 3 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 that 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(s) of full zeros, then 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 the coefficient matrix 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 can still be a unique solution or infinitely many of them.

---

[3]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.

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.

Note that in Earth Science, more often than not, the numbers of equations and unknowns are not exactly equal, hence the concerned system will usually be either overdetermined or underdetermined, e.g. there are more/less data samples (radiosonde soundings) than the variables (temperature, humidity) in a measurement campaign. To this end, let's now look at some worked examples of solving non-square linear systems.

Overdetermined Systems

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

[4]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 the 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 $A\vec{x} = 0$.

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.

$$\left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 2 & 2 & 3 & 8 \\ 1 & 1 & 2 & 3 \\ 0 & 3 & 1 & 5 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & -6 & 3 & 0 \\ 0 & -3 & 2 & -1 \\ 0 & 3 & 1 & 5 \end{array}\right] \qquad \begin{array}{r} R_2 - 2R_1 \to R_2 \\ R_3 - R_1 \to R_3 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & -3 & 2 & -1 \\ 0 & 3 & 1 & 5 \end{array}\right] \qquad -\frac{1}{6}R_2 \to R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & \frac{5}{2} & 5 \end{array}\right] \qquad \begin{array}{r} R_3 + 3R_2 \to R_3 \\ R_4 - 3R_2 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & \frac{5}{2} & 5 \end{array}\right] \qquad 2R_3 \to R_3$$

$$\to \left[\begin{array}{ccc|c} 1 & 4 & 0 & 4 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 10 \end{array}\right] \qquad R_4 - \frac{5}{2}R_3 \to R_4$$

The last row is inconsistent, and hence the overdetermined system has no solution. □

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

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

Solution.

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

$$\to \left[\begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_3 + R_2 \to R_3 \\ R_4 + R_2 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & -1 & 1 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_1 - R_2 \to R_1$$

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.

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

Solution.

$$\left[\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 3 & 2 & -1 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & -1 & 1 & 1 & -1 \end{array}\right] \quad \begin{array}{l} R_2 - R_1 \to R_2 \\ R_3 - R_1 \to R_3 \end{array}$$

$$\to \left[\begin{array}{cccc|c} 1 & 1 & 2 & 1 & 0 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad R_3 + R_2 \to R_3$$

$$\to \left[\begin{array}{cccc|c} 1 & 0 & 3 & 2 & -1 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad R_1 - R_2 \to R_1$$

The last full-zero row is consistent. The first two columns are pivotal, and we can let the remaining two unknowns that are not associated with 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

$$\begin{aligned} A^{-1}A\vec{x} &= (A^{-1}A)\vec{x} = A^{-1}\vec{h} \\ \vec{x} &= I\vec{x} = A^{-1}\vec{h} \quad \text{(Definition 2.2.1 and Properties 2.1.2)} \end{aligned} \tag{3.3}$$

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 (3.3)) 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 computers, the calculation of inverses 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 Statements, 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 the identity 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

For a square linear system of equations, there is another way to find its solution via utilizing determinants, called ***Cramer's Rule***.

Theorem 3.2.2 (Cramer's Rule)**.** Denote the $n \times n$ linear system as $A\vec{x} = \vec{h}$, $\vec{x} = (x_1, x_2, \ldots, x_n)^T$. If the square matrix A is invertible, i.e. $\det(A) \neq 0$, then the solution can be found by

$$x_i = \frac{\det(A_i)}{\det(A)} \tag{3.4}$$

where A_i is the matrix formed after the i-th column of A is replaced by the column vector $\vec{h}$. Usually, we will denote $\det(A)$ by Δ and $\det(A_i)$ by Δ_i in this context.

Example 3.2.8. Use Cramer's Rule to solve

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

Solution. We will first find Δ:

$$\begin{aligned}
\Delta = \det(A) &= \begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & -1 \\ 2 & 0 & 1 \end{vmatrix} \\
&= [(1)(2)(1) + (1)(-1)(2) + (2)(1)(0)] \\
&\quad - [(2)(2)(2) + (0)(-1)(1) + (1)(1)(1)] \\
&= -9
\end{aligned}$$

then each of Δ_i:

$$\begin{aligned}
\Delta_1 = \det(A_1) &= \begin{vmatrix} 1 & 1 & 2 \\ 5 & 2 & -1 \\ 3 & 0 & 1 \end{vmatrix} \\
&= [(1)(2)(1) + (1)(-1)(3) + (2)(5)(0)] \\
&\quad - [(3)(2)(2) + (0)(-1)(1) + (1)(5)(1)] \\
&= -18 \\
\Delta_2 = \det(A_2) &= \begin{vmatrix} 1 & 1 & 2 \\ 1 & 5 & -1 \\ 2 & 3 & 1 \end{vmatrix} \\
&= [(1)(5)(1) + (1)(-1)(2) + (2)(1)(3)] \\
&\quad - [(2)(5)(2) + (3)(-1)(1) + (1)(1)(1)] \\
&= -9
\end{aligned}$$

We leave it to the readers to check that $\Delta_3 = 9$. So the solution of the system is

$$\begin{aligned}
x_1 &= \frac{\Delta_1}{\Delta} = \frac{-18}{-9} = 2 \\
x_2 &= \frac{\Delta_2}{\Delta} = \frac{-9}{-9} = 1 \\
x_3 &= \frac{\Delta_3}{\Delta} = \frac{9}{-9} = -1
\end{aligned}$$

□

Cramer's Rule can also be used for systems in terms of functions.

Example 3.2.9. Use Cramer's Rule to find (u, v) for the system of equations below.

$$\begin{cases} \cos(2x)u + \sin(2x)v = 0 \\ -2\sin(2x)u + 2\cos(2x)v = \tan x \end{cases} \tag{3.5}$$

Solution. Δ and Δ_i will depend on (functions of) x.

$$\Delta = \begin{vmatrix} \cos(2x) & \sin(2x) \\ -2\sin(2x) & 2\cos(2x) \end{vmatrix} = 2\cos^2(2x) + 2\sin^2(2x) = 2$$

$$\begin{aligned} \Delta_1 &= \begin{vmatrix} 0 & \sin(2x) \\ \tan x & 2\cos(2x) \end{vmatrix} \\ &= -\sin(2x)\tan x = -(2\sin x\cos x)(\frac{\sin x}{\cos x}) = -2\sin^2 x \\ \Delta_2 &= \begin{vmatrix} \cos(2x) & 0 \\ -2\sin(2x) & \tan x \end{vmatrix} \\ &= \cos(2x)\tan x = (2\cos^2 x - 1)\tan x \\ &= 2\sin x\cos x - \tan x \end{aligned}$$

Therefore

$$\begin{aligned} u &= \frac{\Delta_1}{\Delta} = -\sin^2 x \\ v &= \frac{\Delta_2}{\Delta} = \sin x\cos x - \frac{1}{2}\tan x \end{aligned}$$

□

3.3. Earth Science Applications

Now we are going to revisit and find the solutions to the four linear systems in Section 1.4.

Example 3.3.1. Solve for the horizontal displacement x and depth of the 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 2×2 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 the top layer are about 94.6 m and 25.4 m respectively. □

Example 3.3.2. Find the radiative loss E_j and hence the 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 E_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.

$$\left[\begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array}\right] \qquad R_N + R_{N+1} \rightarrow R_N$$

$$\rightarrow \left[\begin{array}{ccccccc|c} -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 0 & 0 & -E_{in} \\ 0 & 0 & 0 & & 1 & -1 & 0 & -E_{in} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & -E_{in} \end{array}\right] \qquad R_{N-1} + R_N \rightarrow R_{N-1}$$

$$\rightarrow \quad \vdots \qquad \text{(Keep going up)}$$

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

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

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

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

$$E_j = E_{j-1} + E_{in} \tag{3.6}$$

and inductively we have

$$E_j = jE_{in} \tag{3.7}$$

$E_1 = E_{in}$ is the emission of radiation from the Earth as a whole as viewed from space, and the corresponding *emission temperature* is $T_e = T_1 = \sqrt[4]{E_1/\sigma} = \sqrt[4]{E_{in}/\sigma}$ by the 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 $N = (288/255)^4 - 1 = 0.627$. □

Example 3.3.3. Find the form of the general solution to the underdetermined system about the concentrations of ions in seawater in Example 1.4.3.

Solution. The underdetermined system is

$$\begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \begin{bmatrix} n \\ m \\ c \\ s \end{bmatrix} = \begin{bmatrix} 0 \\ 34.0 \end{bmatrix}$$

or in augmented form,

$$\left[\begin{array}{cccc|c} 1 & 2 & -1 & -2 & 0 \\ 23.0 & 24.3 & 35.5 & 96.1 & 34.0 \end{array}\right]$$

Just a few steps of Gaussian Elimination yields

$$\begin{aligned}
&\left[\begin{array}{cccc|c} 1 & 2 & -1 & -2 & 0 \\ 23.0 & 24.3 & 35.5 & 96.1 & 34.0 \end{array}\right] \\
\to &\left[\begin{array}{cccc|c} 1 & 2 & -1 & -2 & 0 \\ 0 & -21.7 & 58.5 & 142.1 & 34.0 \end{array}\right] && R_2 - 23.0R_1 \to R_2 \\
\to &\left[\begin{array}{cccc|c} 1 & 2 & -1 & -2 & 0 \\ 0 & 1 & -2.70 & -6.55 & -1.57 \end{array}\right] && -\frac{R_2}{21.7} \to R_2 \\
\to &\left[\begin{array}{cccc|c} 1 & 0 & 4.40 & 11.1 & 3.14 \\ 0 & 1 & -2.70 & -6.55 & -1.57 \end{array}\right] && R_1 - 2R_2 \to R_1
\end{aligned}$$

We can assign the non-pivotal unknowns c and s as two free variables, which means that $m = -1.57 + 2.70c + 6.55s$ and $n = 3.14 - 4.40c - 11.1s$, and thus the general solution will be in the form of

$$\begin{bmatrix} n \\ m \\ c \\ s \end{bmatrix} = \begin{bmatrix} 3.14 - 4.40c - 11.1s \\ -1.57 + 2.70c + 6.55s \\ c \\ s \end{bmatrix} = \begin{bmatrix} 3.14 \\ -1.57 \\ 0 \\ 0 \end{bmatrix} + c \begin{bmatrix} -4.40 \\ 2.70 \\ 1 \\ 0 \end{bmatrix} + s \begin{bmatrix} -11.1 \\ 6.55 \\ 0 \\ 1 \end{bmatrix}$$

with the extra physical constraint that all of the ion concentrations should be non-negative, i.e. $n, m, c, s \geq 0$. □

Example 3.3.4. Show that the overdetermined system about air temperature measurements in Example 1.4.4 has no solution.

Solution. The overdetermined system is

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

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

$$\begin{aligned}
\left[\begin{array}{cc|c} 10 & 20 & 0.2 \\ 25 & 15 & 0.3 \\ -10 & 5 & -0.2 \end{array}\right] &\to \left[\begin{array}{cc|c} 1 & 2 & 0.02 \\ 25 & 15 & 0.3 \\ -10 & 5 & -0.2 \end{array}\right] && \frac{1}{10}R_1 \to R_1 \\
&\to \left[\begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & -35 & -0.2 \\ 0 & 25 & 0 \end{array}\right] && \begin{array}{r} R_2 - 25R_1 \to R_2 \\ R_3 + 10R_1 \to R_23 \end{array} \\
&\to \left[\begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 25 & 0 \\ 0 & -35 & -0.2 \end{array}\right] && R_2 \leftrightarrow R_3 \\
&\to \left[\begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 1 & 0 \\ 0 & -35 & -0.2 \end{array}\right] && \frac{1}{25}R_2 \to R_2 \\
&\to \left[\begin{array}{cc|c} 1 & 2 & 0.02 \\ 0 & 1 & 0 \\ 0 & 0 & -0.2 \end{array}\right] && R_3 + 35R_2 \to R_3
\end{aligned}$$

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 (close enough to) constant, there may be measurement errors or limitations in precision for the thermometers. Therefore, it is almost impossible to obtain an exact solution to this type of problem, but often a "good enough" approximation is still useful. We will learn more about how to compute such an optimized solution in Chapter 13. □

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

A1 = np.array([[1., 0., 1.],
               [2., 2., 3.],
               [1., 2., 0.]])
h = np.array([0., -1., 1.])
x = linalg.solve(A1,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, then

```
A2 = 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(A2,h)
print(x)
```

may raise a warning and an unreasonable output of

```
LinAlgWarning: Ill-conditioned matrix
(rcond=3.42661e-18): result may not be accurate.
  x = linalg.solve(A2,h)
[ 4.803839e+15  2.401919e+16 -9.607679e+15 -4.803839e+15]
```

Again, it has been fixed in newer versions. We can also use the `sympy` package for the rescue, as follows.

```
import sympy

A2_sympy = sympy.Matrix(A2)
h_sympy = sympy.Matrix(h)
A2_sympy.solve(h_sympy)
```

which raises the same "not invertible" error as in Section 2.4. For non-square systems, we can use the more general `solveset.linsolve`. Let's use Example 3.2.6 to demonstrate its usage:

```
A3_sympy = sympy.Matrix([[1., 1., 2., 1.],
                         [1., 2., 1., 0.],
                         [1., 0., 3., 2.]])
h_sympy = sympy.Matrix(([0., 1., -1.]))

print(sympy.linsolve([A3_sympy,h_sympy]))
```

which gives

```
{(-3.0*tau0 - 2.0*tau1 - 1.0, 1.0*tau0 + 1.0*tau1 + 1.0, tau0,
    tau1)}
```

where `tau0` and `tau1` are equivalent to s and t in the original example.

Moreover, there are two questions that may come to curious readers when reading the programming sections of these two chapters. First, which of `scipy` and `sympy` should we choose over the other? 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`. For 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.

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

Exercise 3.2 Solve $A\vec{x} = \vec{h}^{(k)}$, $k = 1, 2$, where

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

$$\vec{h}^{(1)} = \begin{bmatrix} -1 \\ 5 \\ 1 \end{bmatrix} \qquad \vec{h}^{(2)} = \begin{bmatrix} \frac{19}{4} \\ 1 \\ \frac{5}{4} \end{bmatrix}$$

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 α so that the system has no solution or infinitely many solutions.

Exercise 3.6 *Ohm's law* relates the 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 (the 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 3.3.

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.

Figure 3.3.: *The circuit for Exercise 3.6.*

Exercise 3.7 The *shallow water equations* (see Figure 3.4) describe the evolution of the gravity wave under some approximations such as *hydrostatic balance* and a sufficiently shallow fluid depth, and have the form of

$$\begin{cases} \dfrac{\partial \eta}{\partial t} + H(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}) = 0 \\ \dfrac{\partial u}{\partial t} = -g\dfrac{\partial \eta}{\partial x} \\ \dfrac{\partial v}{\partial t} = -g\dfrac{\partial \eta}{\partial y} \end{cases} \tag{3.8}$$

when the Coriolis effect is ignored. By assuming a traveling 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 will become

$$
\begin{cases}
\omega\tilde{\eta} - kH\tilde{U} - lH\tilde{V} = 0 \\
\omega\tilde{U} - kg\tilde{\eta} \qquad\qquad = 0 \\
\omega\tilde{V} - lg\tilde{\eta} \qquad\qquad = 0
\end{cases}
\tag{3.9}
$$

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.

Figure 3.4.: *The x-z cross-section of shallow water system in Exercise 3.7. η is the height of free surface, H is the mean depth of the fluid, and u [v] is the fluid velocity along the x-axis [y-axis, into/out of the page].*

Chapter 4

Introduction to Vectors

After three chapters of discussion about matrices, it is time to introduce another closely related type of object in linear algebra: *vectors*. While vectors and *vector spaces* have strictly mathematical definitions that make them abstract, we will take a more physical point of view with the special case of (finite-dimensional) geometric vectors first. Many physical quantities that we commonly come across, like displacements, velocities, and acceleration, in addition to various kinds of interacting forces (gravity, electric, magnetic, etc.), can be represented using vectors. To facilitate, we will learn about how to carry out relevant vector operations for future calculations.

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. $(1, 8, 7, 4)$, $(1 - i, 1 + 3i, 2)$. It has a *magnitude (length)* and *direction*, resembling an arrow. Some real-life examples are:

two-dimensional fluid flow velocity (u, v), the 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*:

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} = (v_1, v_2, v_3, \ldots, v_n)^T \tag{4.1}$$

An 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, 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 in the definition above to reflect that it is in fact a column vector despite being written horizontally.

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-dimensional vectors $\vec{v} = (v_1, v_2, v_3, \ldots, v_n)^T$ as defined in Definition 4.1.1, where v_i can take any *real* only value, for $i = 1, 2, 3, \ldots, n$. Conversely, such objects in $\mathbb{R}^n$ are known as n-dimensional *real* vectors.

Although we have not clearly defined what a vector space is, we note that $\mathbb{R}^n$ satisfies 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.

Figure 4.1.: *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

Table 4.2.: *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).*

An n-dimensional real geometric vector as described in Definitions 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 coordinate axes, where p is from 1 to n. They are denoted as $\hat{e}^{(p)}$. 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 $\mathbb{R}^3$, $\hat{e}^{(1)} = \hat{\imath} = (1, 0, 0)^T$, $\hat{e}^{(2)} = \hat{\jmath} = (0, 1, 0)^T$, and $\hat{e}^{(3)} = \hat{k} = (0, 0, 1)^T$ 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, consisting 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 $q \neq p$.

Figure 4.3 below shows an example of a geometric vector in the three-dimensional xyz space ($\mathbb{R}^3$).

Figure 4.3.: *A 3D vector example in* $\mathbb{R}^3$.

$$\vec{v} = \begin{bmatrix} 1.8 \\ 0.4 \\ 1.1 \end{bmatrix} = 1.8 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 0.4 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1.1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 1.8\hat{\imath} + 0.4\hat{\jmath} + 1.1\hat{k}$$
$$= (1.8, 0.4, 1.1)^T$$

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

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

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

then (see Figure 4.4)

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

Scalar Multiplication

Multiplying a scalar (it can be real or complex, but we will limit ourselves to the case of real only first) to a vector means that all components are multiplied by that scalar. An example is

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

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

(b) *Subtraction: Similar to addition but with the blue vector oriented in the opposite direction.*

Figure 4.4.: *Demonstration of vector addition and subtraction.*

Looking back at vector subtraction, it can be viewed as an 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/l^2-norm***[1], of a vector $\vec{v}$ is based on a generalized version of ***Pythagoras' Theorem***, and is computed 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

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2 + \cdots + v_n^2} \tag{4.2a}$$

$$= \sqrt{\sum_{k=1}^{n} v_k^2} \tag{4.2b}$$

For instance, the length of a two-dimensional vector follows the usual Pythagoras' Theorem as in Figure 4.5 (a) below. While for Figure 4.5 (b), it is

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

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 $\|\vec{v}\|$.

[1] l^2-norm is often seen in Machine Learning as the cost function to be minimized. However, it is not the only norm that can be used: another example will be the l^1-norm (mean absolute error).

(a) *A 2D vector length example.*

(b) *Another example, now for 3D vector length.*

Figure 4.5.: *Demonstration diagrams of deriving vector lengths.*

Definition 4.1.5 (Unit Vector)**.** The unit vector (***direction***) corresponding to a

non-zero vector $\vec{v}$ is denoted as $\hat{v}$ and is given by

$$\hat{v} = \frac{1}{\|\vec{v}\|}\vec{v} \tag{4.3}$$

where the length $\|\vec{v}\|$ is defined as (4.2) in Definition 4.1.4.

Note that even though 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 of Figure 4.5 (b), and verify that it has a length of 1.[2]

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 dimensions. It is the sum of products of the paired components between the two vectors. In other words, it can be regarded to be the matrix product between a row vector ($1 \times n$ matrix) and a column vector ($n \times 1$ matrix).

Definition 4.2.1 (Dot Product (Real))**.** The *real* 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 matrix notation $\mathbf{u}^T\mathbf{v}$. They are defined as

$$\vec{u} \cdot \vec{v} = \mathbf{u}^T\mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3 + \cdots + u_nv_n \tag{4.4a}$$

$$= \sum_{k=1}^{n} u_k v_k \tag{4.4b}$$

which is a real scalar quantity.

Conversely, it can be said that the 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 $\mathbb{R}^n$,

$$A = \begin{bmatrix} \text{—}\vec{u}^{(1)T}\text{—} \\ \hline \text{—}\vec{u}^{(2)T}\text{—} \\ \hline \vdots \\ \hline \text{—}\vec{u}^{(p)T}\text{—} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(p)} \\ | & | & & | \end{bmatrix}^T \quad B = \begin{bmatrix} | & | & & | \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \cdots & \vec{v}^{(q)} \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} \vec{u}_1^{(1)} & \vec{u}_2^{(1)} & \cdots & \vec{u}_n^{(1)} \\ \vec{u}_1^{(2)} & \vec{u}_2^{(2)} & \cdots & \vec{u}_n^{(2)} \\ \vdots & \vdots & & \vdots \\ \vec{u}_1^{(p)} & \vec{u}_2^{(p)} & \cdots & \vec{u}_n^{(p)} \end{bmatrix} \qquad = \begin{bmatrix} \vec{v}_1^{(1)} & \vec{v}_1^{(2)} & \cdots & \vec{v}_1^{(q)} \\ \vec{v}_2^{(1)} & \vec{v}_2^{(2)} & \cdots & \vec{v}_2^{(q)} \\ \vdots & \vdots & & \vdots \\ \vec{v}_n^{(1)} & \vec{v}_n^{(2)} & \cdots & \vec{v}_n^{(q)} \end{bmatrix}$$

(notice the 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 the dot product to express the length of a real vector.

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

Properties 4.2.2. The length of a real vector, as defined by (4.2) in Definition 4.1.4, can be written using its dot product between itself as

$$\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} \qquad \text{or} \qquad \|\vec{v}\|^2 = \vec{v} \cdot \vec{v} \tag{4.5}$$

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, which can never be negative.

Example 4.2.1. If $\vec{u} = (1, 2, 3, 4, 5)^T$ and $\vec{v} = (-1, 0, 2, 1, -2)^T$, find their 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,

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

□

Here are some properties of dot product.

Properties 4.2.3. For three n-dimensional real vectors $\vec{u}, \vec{v}, \vec{w} \in \mathbb{R}^n$, the following properties hold.

$$\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u} \qquad \text{Symmetry Property}$$

$$\vec{u} \cdot (\vec{v} \pm \vec{w}) = \vec{u} \cdot \vec{v} \pm \vec{u} \cdot \vec{w} \qquad \text{Distributive Property}$$
$$(\vec{u} \pm \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} \pm \vec{v} \cdot \vec{w} \qquad \text{Distributive Property}$$
$$(a\vec{u}) \cdot (b\vec{v}) = ab(\vec{u} \cdot \vec{v}) \qquad \text{where } a, b \text{ are some constants}$$

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

$$\vec{u} \cdot (A\vec{v}) = \mathbf{u}^T(A\mathbf{v}) = (A^T\mathbf{u})^T\mathbf{v} = (A^T\vec{u}) \cdot \vec{v}$$
$$(A\vec{u}) \cdot \vec{v} = (A\mathbf{u})^T\mathbf{v} = \mathbf{u}^T(A^T\mathbf{v}) = \vec{u} \cdot (A^T\vec{v})$$

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,

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

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) □

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.

$$\begin{aligned} A\vec{v} &= \begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} (1)(2) + (2)(-1) + (1)(1) \\ (2)(2) + (0)(-1) + (3)(1) \\ (1)(2) + (1)(-1) + (-1)(1) \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 7 \\ 0 \end{bmatrix} \\ \vec{u} \cdot (A\vec{v}) &= (1, 3, 1)^T \cdot (1, 7, 0)^T \\ &= (1)(1) + (3)(7) + (1)(0) \\ &= 22 \end{aligned}$$

On the other hand,

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

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

□

Geometric Meaning of Dot Product

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

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

$$\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta \tag{4.6}$$

where θ 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 $\|\hat{u}\| = \|\hat{v}\| = 1$, it reduces to

$$\hat{u} \cdot \hat{v} = \cos\theta \tag{4.7}$$

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 Figure 4.6 below. While an angle has a clear physical meaning only in a two/three-dimensional space, such a 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.

Figure 4.6.: *Schematic diagram for the relation between dot product and projection.*

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

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

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 Formula (4.6), 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 $\cos\theta = \pm 1$, θ is either 0 or π, 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 (4.6) in Properties 4.2.4, $\cos\theta = 0$ and the angle θ 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 property still holds.

Some may notice that as $-1 \leq \cos\theta \leq 1$, if $|\vec{u} \cdot \vec{v}| > \|\vec{u}\|\|\vec{v}\|$, then θ in (4.6) 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:

$$|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\|\|\vec{v}\| \tag{4.8a}$$

$$|u_1v_1 + u_2v_2 + \cdots + u_nv_n| \leq \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \tag{4.8b}$$

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 positivity (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,

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

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

□

► Short Exercise: Think about under what circumstances the Cauchy-Schwarz Inequality will turn into an equality (i.e. $|\vec{u}\cdot\vec{v}| = \|\vec{u}\|\|\vec{v}\|$).[3]

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

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

Solution. Let's denote the lengths $\|\vec{u}\|$, $\|\vec{v}\|$, $\|(\vec{u}-\vec{v})\|$ by 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 $\mathbb{R}^3$. (See Figure 4.7)

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

1. $\hat{\imath} \times \hat{\jmath} = \hat{k}$, $\hat{\jmath} \times \hat{\imath} = -\hat{k}$,
2. $\hat{\jmath} \times \hat{k} = \hat{\imath}$, $\hat{k} \times \hat{\jmath} = -\hat{\imath}$,
3. $\hat{k} \times \hat{\imath} = \hat{\jmath}$, $\hat{\imath} \times \hat{k} = -\hat{\jmath}$, and
4. $\hat{\imath} \times \hat{\imath} = \hat{\jmath} \times \hat{\jmath} = \hat{k} \times \hat{k} = \mathbf{0}$.

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.

[3]When $\vec{u}$ and $\vec{v}$ are parallel, i.e. $\vec{u} = k\vec{v}$ for some scalar k, or $\vec{v} = \mathbf{0}$.

(a) *A helper cyclic diagram for memorizing Definition 4.2.7. A clockwise/anticlockwise permutation produces a positive/negative standard unit vector of the third.*

(b) *Demonstration of the right-hand rule.*

Figure 4.7.: *Schematic diagrams for the cross product.*

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

$$\vec{u} \times \vec{v} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \tag{4.9}$$

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

$$\begin{aligned}
\vec{u} \times \vec{v} &= (u_1\hat{\imath} + u_2\hat{\jmath} + u_3\hat{k}) \times (v_1\hat{\imath} + v_2\hat{\jmath} + v_3\hat{k}) \\
&= u_1v_1(\hat{\imath} \times \hat{\imath}) + u_1v_2(\hat{\imath} \times \hat{\jmath}) + u_1v_3(\hat{\imath} \times \hat{k}) \\
&\quad + u_2v_1(\hat{\jmath} \times \hat{\imath}) + u_2v_2(\hat{\jmath} \times \hat{\jmath}) + u_2v_3(\hat{\jmath} \times \hat{k}) \\
&\quad + u_3v_1(\hat{k} \times \hat{\imath}) + u_3v_2(\hat{k} \times \hat{\jmath}) + u_3v_3(\hat{k} \times \hat{k}) && \text{(Properties 4.2.8)} \\
&= u_1v_1(\mathbf{0}) + u_1v_2(\hat{k}) - u_1v_3(\hat{\jmath})
\end{aligned}$$

$$
\begin{aligned}
&- u_2v_1(\hat{k}) + u_2v_2(\mathbf{0}) + u_2v_3(\hat{\imath}) \\
&+ u_3v_1(\hat{\jmath}) - u_3v_2(\hat{\imath}) + u_3v_3(\mathbf{0}) && \text{(Definition 4.2.7)} \\
&= (u_2v_3 - u_3v_2)\hat{\imath} + (u_3v_1 - u_1v_3)\hat{\jmath} + (u_1v_2 - u_2v_1)\hat{k}
\end{aligned}
$$

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

$$
\begin{aligned}
\begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} &= \hat{\imath}\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} - \hat{\jmath}\begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} + \hat{k}\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \\
&= (u_2v_3 - u_3v_2)\hat{\imath} + (u_3v_1 - u_1v_3)\hat{\jmath} + (u_1v_2 - u_2v_1)\hat{k}
\end{aligned}
$$

yields an identical result. □

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

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

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

Solution.

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

□

► Short Exercise: Check if $\vec{u} \times \vec{v}$ is orthogonal to $\vec{u}$ and $\vec{v}$ in the above example by finding the corresponding dot products.[4]

► Short Exercise: Following the short exercise above, show that in general, $\vec{u} \cdot (\vec{u} \times \vec{v}) = \vec{v} \cdot (\vec{u} \times \vec{v}) = 0$, i.e. the cross product is orthogonal to the two input vectors.[5]

Geometric Meaning of Cross Product

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

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

$$\|\vec{u} \times \vec{v}\| = \|\vec{u}\|\|\vec{v}\| \sin\theta \tag{4.10}$$

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 π) and $\sin\theta = 0$. This is equivalent to the statement of $\vec{u} \times \vec{u} = \mathbf{0}$.[6] (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.)[7]

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

$$\begin{aligned}
-3\vec{u} \times \vec{v} &= -3 \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 1 & 2 & 3 \\ -1 & 1 & 2 \end{vmatrix} \\
&= -3 \left(\hat{\imath} \begin{vmatrix} 2 & 3 \\ 1 & 2 \end{vmatrix} - \hat{\jmath} \begin{vmatrix} 1 & 3 \\ -1 & 2 \end{vmatrix} + \hat{k} \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} \right) && \text{(Cofactor expansion along the first row)} \\
&= -3(\hat{\imath} - 5\hat{\jmath} + 3\hat{k}) \\
&= -3\hat{\imath} + 15\hat{\jmath} - 9\hat{k} = (-3, 15, -9)^T
\end{aligned}$$

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

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.

---

[4] $\vec{u} \cdot (\vec{u} \times \vec{v}) = (1, 0, 2)^T \cdot (2, 5, -1)^T = (1)(2) + (0)(5) + (2)(-1) = 0$, $\vec{v} \cdot (\vec{u} \times \vec{v}) = (3, -1, 1)^T \cdot (2, 5, -1)^T = (3)(2) + (-1)(5) + (1)(-1) = 0$. The zero dot products in both cases show that they are orthogonal via Properties 4.2.5.

[5] From the derivation of Properties 4.2.9, $\vec{u} \times \vec{v} = (u_2v_3 - u_3v_2)\hat{\imath} + (u_3v_1 - u_1v_3)\hat{\jmath} + (u_1v_2 - u_2v_1)\hat{k}$, and $\vec{u} \cdot (\vec{u} \times \vec{v}) = u_1(u_2v_3 - u_3v_2) + u_2(u_3v_1 - u_1v_3) + u_3(u_1v_2 - u_2v_1) = 0$ where all terms cancel out, and it is similar for $\vec{v}$.

[6] By Properties 4.2.9,

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

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

[7] The anti-symmetric property requires $\vec{u} \times \vec{u} = -\vec{u} \times \vec{u}$ and hence $2(\vec{u} \times \vec{u}) = \mathbf{0}$.

	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

Table 4.8.: *A table comparing the differences between dot/cross product.*

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

$$\overrightarrow{F_{\text{cor}}} = -2\vec{\Omega} \times \vec{v} \tag{4.11}$$

where $\Omega = \|\vec{\Omega}\| = 7.292 \times 10^{-5}\,\text{rad}\,\text{s}^{-1}$ 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 4.9) 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 $\vec{v} = (u, v, w) = u\hat{\imath}+v\hat{\jmath}+w\hat{k}$ as the flow velocity in this local Cartesian coordinate system with unit vectors $\hat{\imath}, \hat{\jmath}, \hat{k}$ along the x, y, z axes. It can be seen that $\vec{\Omega} = (\Omega\cos\varphi)\hat{\jmath} + (\Omega\sin\varphi)\hat{k}$ where φ 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

$$F_{\text{cor},x} = 2\Omega(v\sin\varphi - w\cos\varphi) \tag{4.12a}$$

$$F_{\text{cor},y} = -2\Omega u\sin\varphi \tag{4.12b}$$

$$F_{\text{cor},z} = 2\Omega u\cos\varphi \tag{4.12c}$$

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

Figure 4.9.: *An illustration of the local coordinate frame to derive Coriolis Force in Example 4.3.1.*

Solution. Using (4.9) in Properties 4.2.9 to expand $\overrightarrow{F_{\text{cor}}}$ gives

$$
\begin{aligned}
-2\overrightarrow{\Omega} \times \vec{v} &= -2\left((\Omega\cos\varphi)\hat{\jmath} + (\Omega\sin\varphi)\hat{k}\right) \times (u\hat{\imath} + v\hat{\jmath} + w\hat{k}) \\
&= -2\begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 0 & \Omega\cos\varphi & \Omega\sin\varphi \\ u & v & w \end{vmatrix} \\
&= -2[(w\Omega\cos\varphi - v\Omega\sin\varphi)\hat{\imath} + (u\Omega\sin\varphi)\hat{\jmath} - (u\Omega\cos\varphi)\hat{k}] \\
&= [2\Omega(v\sin\varphi - w\cos\varphi)]\hat{\imath} + (-2\Omega u\sin\varphi)\hat{\jmath} + (2\Omega u\cos\varphi)\hat{k}
\end{aligned}
$$

The $\hat{\imath}$, $\hat{\jmath}$, $\hat{k}$ components then 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 $F_{\text{cor},y} = -fu$. □

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 work 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)) # or np.dot(myVec1, myVec1)**0.5
```

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 Formula (4.7) 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

```
print(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) $\vec{u} \cdot (A\vec{v}) = (A^T\vec{u}) \cdot \vec{v}$,

(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 snapshots. 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	$13\,\mathrm{km\,h^{-1}}$	315°	$(-9.192, 9.192)$
Vicente	2012/07/24, 02:00	$18\,\mathrm{km\,h^{-1}}$	299°	
Linfa	2015/07/09, 23:00			$(-13.595, -6.339)$
Mangkhut	2018/09/16, 22:00		288°	$(\quad, 7.725)$

Figure 4.10.: *The parallelogram constructed by vectors for Exercise 4.6.*

Exercise 4.5 Prove the *Triangle Inequality*.

$$\|\vec{u}+\vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\| \tag{4.13}$$

Exercise 4.6 Prove the *Parallelogram Law*. (See Figure 4.10)

$$2\|\vec{u}\|^2 + 2\|\vec{v}\|^2 = \|\vec{u}+\vec{v}\|^2 + \|\vec{u}-\vec{v}\|^2 \tag{4.14}$$

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 provide 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 relationships between vectors, equations, shapes, *dot/cross products*, and geometry. Some physical applications are to compute areas, volumes, and (great-arc) distances.

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, or (b) vectors. We will start with 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 that controls its extension or contraction, so that it traces out the line when the parameter changed continuously.

Figure 5.1.: *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 the 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 for the example in Figure 5.1 and substitute that value into the expression of $\overrightarrow{OP}$ 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.[1]

5.1.1. Translating Equation Form to Vector Form

The general equation form of a line on an xy 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 by 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

$$\overrightarrow{OP} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{h}{a} - \frac{b}{a}t \\ t \end{bmatrix} = \begin{bmatrix} \frac{h}{a} \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{b}{a} \\ 1 \end{bmatrix}$$

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 can 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 the complementary solution) can be scaled freely. In addition, any initial position vector (particular solution) can be chosen as long as it connects to a point on the line and satisfies the equation. (You may refer to the discussion about particular/complementary solutions 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.[2]

[1] Let's say $t = -0.25$, $\overrightarrow{OP} = (-1, 0.5)^T + (-0.25)(2, 1)^T = (-1.5, 0.25)^T$, $x - 2y = (-1.5) - 2(0.25) = -2$.

► Short Exercise: What is the vector form of the equation $ax + by = h$ for the degenerate case $a = 0$?[3]

5.1.2. Recovering Equation Form from Vector Form

On the other hand, inferring a 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 a dot product with the ***normal vector*** of the line, which is orthogonal to the direction vector, on both sides of the vector form, we will eliminate the parameter and recover the line equation. For example, given that 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

$$\begin{aligned}
\begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} &= \left(\begin{bmatrix} 1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ 4 \end{bmatrix} \right) \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} + t \left(\begin{bmatrix} 1 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -1 \end{bmatrix} \right) \\
4x - y &= ((1)(4) + (3)(-1)) + ((1)(4) + (4)(-1))t = 1 + (0)t = 1 \\
\Rightarrow 4x - y &= 1
\end{aligned}$$

[2]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 other 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

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix} + t_1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \left(\begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right) + 2t_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} + (2t_1 + 2) \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Notice that the coefficients of the equation are the same as the components of the normal vector.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

► Short Exercise: Verify that $(a, b)^T$ is always normal/orthogonal to $(b, -a)^T$, and vice versa.[4]

5.1.3. Generalizing to Higher Dimensions

Similar concepts can be applied to the equation and vector forms for planes. The general form of the 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 that 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 that 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 a line in the last subsection. Again, the coefficients of the plane equation match the components of the normal vector, differing at most by a multiplicative factor. (See Figure 5.2)

[3]The equation is reduced to $y = \frac{h}{b}$ and we select $x = t$ as the free variable instead.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t \\ \frac{h}{b} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{h}{b} \end{bmatrix} + t \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Figure 5.2.: *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 verify 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 a cross product on

[4] $(a, b)^T \cdot (b, -a)^T = (a)(b) + (b)(-a) = 0$

the two direction vectors obtained above. By Properties 4.2.9, we can acquire a normal vector of

$$\begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ -\frac{3}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{vmatrix} = \hat{\imath} + \frac{3}{2}\hat{\jmath} + \frac{1}{2}\hat{k}$$

The next step is to take the dot product on both sides of the vector equation with the normal vector just retrieved.

$$\begin{aligned}
\begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} \\
\begin{bmatrix} x \\ y \\ z \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} &= \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} + s\left(\begin{bmatrix} -\frac{3}{2} \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} \right) + t\left(\begin{bmatrix} -\frac{1}{2} \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} \right) \\
x + \frac{3}{2}y + \frac{1}{2}z &= 2 + s(0) + t(0) = 2 \\
\Rightarrow 2x + 3y + z &= 4
\end{aligned}$$

□

The correspondence between the coefficients of a linear equation and the 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 $(a_1, a_2, a_3, \ldots, a_n)^T$.

The procedures carried out in the last example can be similarly applied to higher-dimensional situations where the equation now represents a ***hyperplane***[5].

[5] A hyperplane in the n-dimensional real space $\mathbb{R}^n$ can be thought of as an "$(n-1)$-dimensional flat surface".

5.2. Further Geometric Applications of Dot Product

5.2.1. Projection

We have mentioned in Properties 4.2.4 that the dot product between two vectors is related to the projection of one vector onto another. By rearranging the formulae 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 number of dimensions, the ***(signed) scalar projection*** of $\vec{v}$ onto $\vec{u}$ is computed according to

$$\text{proj}_u v = \|\vec{v}\| \cos\theta = \frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|} \tag{5.1}$$

If we want to give directionality to the projection, then we can supply its unit vector $\hat{u}$ to make it a ***vector projection***:

$$\overrightarrow{\text{proj}_u v} = (\text{proj}_u v)\hat{u} = \frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|}\hat{u} = \frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|^2}\vec{u} \tag{5.2a}$$

$$= (\text{proj}_u v)\frac{\vec{u}}{\|\vec{u}\|} \tag{5.2b}$$

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 $\vec{v} = -2\hat{\imath} + 3\hat{\jmath} - \hat{k}$ onto $\vec{u} = 4\hat{\imath} + \hat{\jmath} - 3\hat{k}$.

Solution. According to Properties 5.2.1, The signed scalar projection of $\vec{v}$ into $\vec{u}$ is

$$\text{proj}_u v = \frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|}$$

$$= \frac{(-2)(4) + (3)(1) + (-1)(-3)}{\sqrt{(4)^2 + (1)^2 + (-3)^2}}$$
$$= -\frac{2}{\sqrt{26}} = -\frac{\sqrt{26}}{13}$$

and the vector projection is

$$\begin{aligned}\overrightarrow{\text{proj}}_u v &= (\text{proj}_u v)\frac{\vec{u}}{\|\vec{u}\|} \\ &= (-\frac{\sqrt{26}}{13})\left(\frac{4\hat{\imath} + \hat{\jmath} - 3\hat{k}}{\sqrt{26}}\right) \\ &= -\frac{1}{13}(4\hat{\imath} + \hat{\jmath} - 3\hat{k}) = \left(-\frac{4}{13}, -\frac{1}{13}, \frac{3}{13}\right)^T\end{aligned}$$

□

5.2.2. Distance

The 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 Figure 5.3 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 $\hat{\imath} - 2\hat{\jmath} + 3\hat{k}$. 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 $(3, 3, 6)^T - (4, 2, 2)^T = -\hat{\imath} + \hat{\jmath} + 4\hat{k}$. Subsequently, the distance is found from the length of the projection

Figure 5.3.: *Schematic diagram for finding the distance of a point to a line/plane.*

of this vector $-\hat{\imath} + \hat{\jmath} + 4\hat{k}$ onto the normal vector of the plane $\hat{\imath} - 2\hat{\jmath} + 3\hat{k}$. By (5.1) of Properties 5.2.1, it is

$$\frac{(-\hat{\imath} + \hat{\jmath} + 4\hat{k}) \cdot (\hat{\imath} - 2\hat{\jmath} + 3\hat{k})}{\left\|\hat{\imath} - 2\hat{\jmath} + 3\hat{k}\right\|} = \frac{(-1)(1) + (1)(-2) + (4)(3)}{\sqrt{(1)^2 + (-2)^2 + (3)^2}} = \frac{9}{\sqrt{14}}$$

□

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 and the next section are assumed to be of $\mathbb{R}^3$.

5.3.1. Area

The area of the parallelogram formed by two vectors $\vec{u}$, $\vec{v}$ is simply the absolute value of their cross product, as seen in Figure 5.4 below.

Figure 5.4.: *Geometric meaning of the cross product.*

Properties 5.3.1. Directly from (4.10) of Properties 4.2.10, the area of the parallelogram formed by two vectors $\vec{u}$ and $\vec{v}$ is

$$\|\vec{u} \times \vec{v}\| = \|\vec{u}\|\|\vec{v}\| \sin\theta \tag{5.3}$$

Similarly, the area of triangle made by $\vec{u}$ and $\vec{v}$ is half of the above:

$$\frac{1}{2}\|\vec{u} \times \vec{v}\| = \frac{1}{2}\|\vec{u}\|\|\vec{v}\| \sin\theta \tag{5.4}$$

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 Formula (4.9), the cross product between the two given vectors is

$$\begin{aligned}\vec{u} \times \vec{v} &= \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ -1 & -2 & 4 \\ 3 & 0 & 1 \end{vmatrix} \\ &= -2\hat{\imath} + 13\hat{\jmath} + 6\hat{k}\end{aligned}$$

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, the volume of parallelepiped (see Figure 5.5 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 $\mathbb{R}^3$ vectors $\vec{u}$, $\vec{v}$, and $\vec{w}$, is calculated as

$$\|\vec{u} \times \vec{v}\| \|\vec{w}\| \cos\theta = |(\vec{u} \times \vec{v}) \cdot \vec{w}| = \text{abs}\left(\begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}\right) \tag{5.5}$$

Figure 5.5.: *The relationship between the volume of parallelepiped and its scalar triple product.*

where

$$(\vec{u} \times \vec{v}) \cdot \vec{w} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \tag{5.6}$$

is the *scalar triple product* of $\vec{u}$, $\vec{v}$, and $\vec{w}$. Also, by applying Properties 2.3.5, the determinant form of the scalar triple product indicates that

$$\begin{aligned} (\vec{u} \times \vec{v}) \cdot \vec{w} &= (\vec{v} \times \vec{w}) \cdot \vec{u} = (\vec{w} \times \vec{u}) \cdot \vec{v} \\ &= -(\vec{v} \times \vec{u}) \cdot \vec{w} = -(\vec{w} \times \vec{v}) \cdot \vec{u} = -(\vec{u} \times \vec{w}) \cdot \vec{v} \end{aligned} \tag{5.7}$$

Proof. We will prove the determinant formula shown above for $(\vec{u} \times \vec{v}) \cdot \vec{w}$ briefly. By Formula (4.9) from Properties 4.2.9, we have

$$\vec{u} \times \vec{v} = (u_2 v_3 - u_3 v_2)\hat{\imath} + (u_3 v_1 - u_1 v_3)\hat{\jmath} + (u_1 v_2 - u_2 v_1)\hat{k}$$

and then according to Definition 4.2.1

$$\begin{aligned} (\vec{u} \times \vec{v}) \cdot \vec{w} &= (u_2 v_3 - u_3 v_2, u_3 v_1 - u_1 v_3, u_1 v_2 - u_2 v_1)^T \cdot (w_1, w_2, w_3)^T \\ &= (u_2 v_3 - u_3 v_2) w_1 + (u_3 v_1 - u_1 v_3) w_2 + (u_1 v_2 - u_2 v_1) w_3 \end{aligned}$$

which is just 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 given determinant form. □

If the volume of the 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 $\mathbb{R}^3$ vectors $\vec{u}$, $\vec{v}$, and $\vec{w}$, if their scalar triple product $(\vec{u} \times \vec{v}) \cdot \vec{w} = 0$ equals 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 α and β are some scalars, then $\vec{u}$, $\vec{v}$, $\vec{w}$ are co-planar: $(\vec{u} \times \vec{v}) \cdot \vec{w} = (\vec{u} \times \vec{v}) \cdot (\alpha\vec{u} + \beta\vec{v}) = \alpha((\vec{u} \times \vec{v}) \cdot \vec{u}) + \beta((\vec{u} \times \vec{v}) \cdot \vec{v}) = \alpha(0) + \beta(0) = 0$ as both $\vec{u} \cdot (\vec{u} \times \vec{v})$ and $\vec{v} \cdot (\vec{u} \times \vec{v})$ are equal to zero (why?).

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 scalar triple 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 required volume is $|-3| = 3$. □

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 dimensions. 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 the 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 that it is consistent with Properties 5.3.1 about the area of two vectors on the xy 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 will be

$$\text{abs}\left(\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}\right) = |u_1 v_2 - v_1 u_2|$$

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 xy plane. Then according to the previous Properties 5.3.1, the area is computed as $\|\vec{u} \times \vec{v}\|$, and by Properties 4.2.9 it is

$$\begin{aligned}
\vec{u} \times \vec{v} &= \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ u_1 & u_2 & 0 \\ v_1 & v_2 & 0 \end{vmatrix} \\
&= \hat{k} \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \quad \text{(Cofactor Expansion along the third row)} \\
&= (u_1 v_2 - v_1 u_2)\hat{k} = (0, 0, u_1 v_2 - v_1 u_2)^T
\end{aligned}$$

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 utility to some readers.

Properties 5.4.1 (Vector Triple Product)**.** The ***vector triple product*** of three vectors $\vec{u}$, $\vec{v}$, $\vec{w}$ is computed as

$$\vec{u} \times (\vec{v} \times \vec{w}) = (\vec{u} \cdot \vec{w})\vec{v} - (\vec{u} \cdot \vec{v})\vec{w} \tag{5.8}$$

Proof. By Properties 4.2.9, the L.H.S. can be expanded into

$$\begin{aligned}
&\vec{u} \times (\vec{v} \times \vec{w}) \\
&= (u_1\hat{\imath} + u_2\hat{\jmath} + u_3\hat{k}) \\
&\quad \times [(v_2w_3 - v_3w_2)\hat{\imath} + (v_3w_1 - v_1w_3)\hat{\jmath} + (v_1w_2 - v_2w_1)\hat{k}] \\
&= \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ u_1 & u_2 & u_3 \\ v_2w_3 - v_3w_2 & v_3w_1 - v_1w_3 & v_1w_2 - v_2w_1 \end{vmatrix}
\end{aligned}$$

The $\hat{\imath}$ component along the x-direction is

$$u_2(v_1w_2 - v_2w_1) - u_3(v_3w_1 - v_1w_3)$$

$$
\begin{aligned}
&= u_2w_2v_1 + u_3w_3v_1 - u_2v_2w_1 - u_3v_3w_1 \\
&= u_1w_1v_1 + u_2w_2v_1 + u_3w_3v_1 - u_1v_1w_1 - u_2v_2w_1 - u_3v_3w_1 \\
&= (u_1w_1 + u_2w_2 + u_3w_3)v_1 - (u_1v_1 + u_2v_2 + u_3v_3)w_1 \\
&= (\vec{u} \cdot \vec{w})v_1 - (\vec{u} \cdot \vec{v})w_1
\end{aligned}
$$

which is equal to the $\hat{\imath}$ component on the R.H.S. and similar results can be shown for the $\hat{\jmath}$, $\hat{k}$ components, and the equality establishes. The geometric meaning of the R.H.S. is a projection onto (the plane constructed by) $\vec{v}$ and $\vec{w}$ but with the $\vec{v}$ and $\vec{w}$ components rotated. □

Properties 5.4.2 (Jacobi Identity)**.**

$$\vec{u} \times (\vec{v} \times \vec{w}) + \vec{v} \times (\vec{w} \times \vec{u}) + \vec{w} \times (\vec{u} \times \vec{v}) = \mathbf{0} \tag{5.9}$$

Proof. By Properties 5.4.1, we have

$$
\begin{aligned}
&\vec{u} \times (\vec{v} \times \vec{w}) + \vec{v} \times (\vec{w} \times \vec{u}) + \vec{w} \times (\vec{u} \times \vec{v}) \\
&= [(\vec{u} \cdot \vec{w})\vec{v} - (\vec{u} \cdot \vec{v})\vec{w}] \\
&\quad + [(\vec{v} \cdot \vec{u})\vec{w} - (\vec{v} \cdot \vec{w})\vec{u}] \\
&\quad + [(\vec{w} \cdot \vec{v})\vec{u} - (\vec{w} \cdot \vec{u})\vec{v}] \\
&= [(\vec{u} \cdot \vec{w})\vec{v} - (\vec{w} \cdot \vec{u})\vec{v}] \\
&\quad + [(\vec{v} \cdot \vec{u})\vec{w} - (\vec{u} \cdot \vec{v})\vec{w}] \\
&\quad + [(\vec{w} \cdot \vec{v})\vec{u} - (\vec{v} \cdot \vec{w})\vec{u}] \\
&= 0\vec{v} + 0\vec{w} + 0\vec{u} = \mathbf{0}
\end{aligned}
$$

The geometric meaning is that, since the vector triple product is basically a signed, rotated projection onto the two other edges of a parallelepiped as seen above, the permutation of the three vector triple products means that we will traverse along the six outer edges of a half-open parallelepiped and return to the initial vertex. □

Properties 5.4.3 (Lagrange's Identity)**.**

$$\|\vec{u} \times \vec{v}\|^2 = \|\vec{u}\|^2\|\vec{v}\|^2 - (\vec{u} \cdot \vec{v})^2 \tag{5.10}$$

Proof. Manipulating the geometric formulae of dot/cross product, we have

$$\begin{aligned}
\|\vec{u} \times \vec{v}\|^2 &= \|\vec{u}\|^2\|\vec{v}\|^2 \sin^2\theta && \text{(Properties 4.2.10)} \\
&= \|\vec{u}\|^2\|\vec{v}\|^2(1 - \cos^2\theta) \\
&= \|\vec{u}\|^2\|\vec{v}\|^2 - \|\vec{u}\|^2\|\vec{v}\|^2 \cos^2\theta \\
&= \|\vec{u}\|^2\|\vec{v}\|^2 - (\vec{u} \cdot \vec{v})^2 && \text{(Properties 4.2.4)}
\end{aligned}$$

By rearranging the identity as $\|\vec{u} \times \vec{v}\|^2 + (\vec{u} \cdot \vec{v})^2 = \|\vec{u}\|^2\|\vec{v}\|^2$, we can identify its geometric meaning with the Pythagoras' theorem: the squared length $(\|\vec{u}\|\|\vec{v}\|)^2$, representing the hypotenuse of a right triangle, is the sum of the squared lengths of the opposite $\|\vec{u} \times \vec{v}\|^2$ and adjacent edge $(\vec{u} \cdot \vec{v})^2$. □

The last identity is the ***Cosine Law for Spherical Trigonometry***.

Properties 5.4.4 (Cosine Law for Spherical Trigonometry)**.**

$$\cos c = \cos a \cos b + \sin a \sin b \cos C \tag{5.11}$$

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 5.6.

Proof. For the given spherical triangle, we can always rotate the coordinate system (see Figure 5.6) 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 $0°$, i.e. $y = 0$). The vector $\hat{v}$ at the end of arc a will then have

Figure 5.6.: *The spherical triangle on a unit sphere as described in Properties 5.4.4.*

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 the xy 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 $\cos c = \cos a \cos b + \sin a \sin b \cos C$. □

It can be seen as the generalization of the usual Cosine Law on a spherical

surface. By first-order Taylor expansion, we have

$$\begin{aligned} 1 - \frac{1}{2}c^2 + \cdots &= (1 - \frac{1}{2}a^2 + \cdots)(1 - \frac{1}{2}b^2 + \cdots) + ab\cos C \\ 1 - \frac{1}{2}c^2 &= 1 - \frac{1}{2}a^2 - \frac{1}{2}b^2 + ab\cos C \\ c^2 &= a^2 + b^2 - 2ab\cos C \end{aligned}$$

so it reduces to the original Cosine Law.

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 °N, 73.94 °W) and Warsaw (52.24 °N, 21.02 °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

$$\begin{aligned} \cos c &= \cos a \cos b + \sin a \sin b \cos C \\ \cos c &= \cos\left(\frac{\pi}{2} - \varphi_1\right)\cos\left(\frac{\pi}{2} - \varphi_2\right) + \sin\left(\frac{\pi}{2} - \varphi_1\right)\sin\left(\frac{\pi}{2} - \varphi_2\right)\cos(\lambda_2 - \lambda_1) \\ \cos c &= \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\lambda_2 - \lambda_1) \end{aligned}$$

The *haversine* of an angle θ is $\operatorname{hav}(\theta) = \sin^2(\frac{\theta}{2}) = \frac{1}{2}(1 - \cos\theta)$ and therefore $\cos\theta = 1 - 2\operatorname{hav}(\theta)$. Subsequently,

$$\cos c = \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 (1 - 2\operatorname{hav}(\lambda_2 - \lambda_1))$$

$$\begin{aligned}
\cos c &= \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 - 2\cos\varphi_1 \cos\varphi_2 \operatorname{hav}(\lambda_2 - \lambda_1) \\
\cos c &= \cos(\varphi_2 - \varphi_1) - 2\cos\varphi_1 \cos\varphi_2 \operatorname{hav}(\lambda_2 - \lambda_1) \\
(1 - 2\operatorname{hav}(c)) &= (1 - 2\operatorname{hav}(\varphi_2 - \varphi_1)) - 2\cos\varphi_1 \cos\varphi_2 \operatorname{hav}(\lambda_2 - \lambda_1) \\
\operatorname{hav}(c) &= \operatorname{hav}(\varphi_2 - \varphi_1) + \cos\varphi_1 \cos\varphi_2 \operatorname{hav}(\lambda_2 - \lambda_1) \qquad (5.12)
\end{aligned}$$

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 °N, 73.94 °W) and Warsaw (52.24 °N, 21.02 °E), $\lambda_1 = -73.94°$, $\lambda_2 = 21.02°$, $\varphi_1 = 40.73°$, $\varphi_2 = 52.24°$, and

$$\begin{aligned}
\operatorname{hav}(c) &= \operatorname{hav}(52.24° - 40.73°) \\
&\quad + \cos(40.73°)\cos(52.24°)\operatorname{hav}(21.02° - (-73.94°)) \\
&= \operatorname{hav}(11.51°) + \cos(40.73°)\cos(52.24°)\operatorname{hav}(94.96°) \\
&= \sin^2\left(\frac{11.51°}{2}\right) + \cos(40.73°)\cos(52.24°)\sin^2\left(\frac{94.96°}{2}\right) \\
\sin^2\left(\frac{c}{2}\right) &\approx 0.26214 \\
c &\approx 61.6° = 1.075\,\text{rad}
\end{aligned}$$

and therefore the required great-arc 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 (minimally) one. □

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 the 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 $90° - 80.7° = 9.3°$. Similarly, the arc from the geographic North Pole to Tokyo is $a = 90° - 35.65° = 54.35°$. We can use the Haversine Formula derived in the last example to obtain the arc from the geomagnetic North Pole to Tokyo, which yields

$$\begin{aligned}
\operatorname{hav}(t) &= \operatorname{hav}(80.7° - 35.65°) \\
&\quad + \cos(35.65°)\cos(80.7°)\operatorname{hav}((-72.7°) - 139.84°) \\
&= \operatorname{hav}(45.05°) + \cos(35.65°)\cos(80.7°)\operatorname{hav}(-212.54°) \\
&\approx 0.26777 \\
t &\approx 62.3°
\end{aligned}$$

Denote the declination angle by D. By Properties 5.4.4, we have

$$\begin{aligned}
\cos d &= \cos a \cos t + \sin a \sin t \cos D \\
\cos(9.3°) &= \cos(54.35°)\cos(62.3°) + \sin(54.35°)\sin(62.3°)\cos D \\
\cos D &\approx 0.9951 \\
D &\approx \pm 5.7°
\end{aligned}$$

To determine the sign, we note that concluding from the longitudes of Tokyo and the geomagnetic North, the geomagnetic North is located to the east of Tokyo, and hence $D = 5.7$ °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. □

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. The area of a parallelogram formed by two vectors is the magnitude of their cross product and the corresponding function is given 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 the 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_parallelepiped(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_parallelepiped(u, v, w))
```

(`np.c_[]` is a shorthand 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

$$\mathrm{Dist}(\vec{u}, \vec{v}) = \frac{(\hat{a} - \hat{b}) \cdot (\hat{l} \times \hat{m})}{\|\hat{l} \times \hat{m}\|}$$

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 the ***Sine Law*** with vector notation by considering the triangle above 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.

Figure 5.7.: *The triangle for Exercise 5.5.*

Exercise 5.6 By extending Properties 5.3.2, derive a vector formula for the volume of a tetrahedron (pyramid).

Figure 5.8.: *The tetrahedron for Exercise 5.6.*

Exercise 5.7 For $\vec{u} = (1, 2, 3)^T$, $\vec{v} = (2, 1, 5)^T$, $\vec{w} = (1, 4, 0)^T$, find

(a) Area of the parallelogram formed by $\vec{u}$ and $\vec{v}$,

(b) Volume of the parallelepiped formed by $\vec{u}$, $\vec{v}$ and $\vec{w}$,

(c) 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*, and *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 to represent vectors in different coordinate systems. Finally, we are going to investigate the so-called *four fundamental subspaces* induced by a matrix and see how they are interconnected via the *Rank-Nullity Theorem*.
Warning: The following three chapters (6-8) will delve into the more theoretical parts of Linear Algebra and involve a lot of heavy, abstract concepts that are difficult for most people. So be prepared for this challenge: it is very normal to feel a bit frustrated. If you are Mathematics-averse, you may want to just skim through them, go directly to Chapter 9, and focus only on manipulating matrices in the remaining parts. However, if you are courageous enough to put the effort into fully understanding these three chapters, you will find the arguments used in the subsequent more application-emphasized chapters much more natural, and obtain a more insightful and valuable understanding of the interwoven principles.

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 under 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 $\mathbf{0} = (0, 0, 0, \ldots, 0)^T$ with n zeros. We will not do it here but interested readers can try to justify all of them.

Building 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 the 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 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);

[1] 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 $\mathbb{R}^n$.

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,

(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.

Figure 6.1.: *Some examples (blue) and non-examples (red) of subspaces in* $\mathbb{R}^2$.

► Short Exercise: Determine if the following subsets of $\mathbb{R}^3$ are subspaces of $\mathbb{R}^3$.[2]

(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,

(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 subsection 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 vectors it contains. To do so, we need to first understand what a ***linear combination*** of vectors is.

Definition 6.1.3 (Linear Combination of Vectors)**.** A linear combination of vectors $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)} \in \mathcal{V}$ where $\mathcal{V}$ is some vector space has the form

of

$$\sum_{j=1}^{q} c_j \vec{v}^{(j)} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \cdots + c_q\vec{v}^{(q)} \tag{6.1}$$

where the coefficients c_j are some scalars (real [complex] numbers for a real [complex] vector space) and the number of vectors q is assumed 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 $\vec{v}$.[3]

Attentive readers may realize that the short exercise above can be thought of 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 $\vec{v}^{(j)} \in \mathbb{R}^n$, $j = 1, 2, \ldots, q$, is equivalent to determining whether the linear system $A\vec{x} = \vec{h}$ has a solution, where A is produced by writing out $\vec{v}^{(j)}$ as a matrix column by column:

$$A = \left[\begin{array}{c|c|c|c} | & | & & | \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \cdots & \vec{v}^{(q)} \\ | & | & & | \end{array}\right]$$

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.

[2]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.

Properties 6.1.4. A linear combination $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \cdots + c_q\vec{v}^{(q)}$ made up of some vectors $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)} \in \mathbb{R}^n$ as in (6.1) of Definition 6.1.3, can be expressed by the matrix product $A\vec{x}$, where

$$A = \left[\begin{array}{c|c|c|c} | & | & & | \\ \vec{v}^{(1)} & \vec{v}^{(2)} & \cdots & \vec{v}^{(q)} \\ | & | & & | \end{array}\right] \qquad \vec{x} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_q \end{bmatrix} \tag{6.2}$$

From now on, we will just simply write $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(q)}]$ 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

$$A\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad A\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad A\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

and it goes similarly for any other column. Below is a small example to demonstrate the equivalence between matrix-vector products and linear combinations.

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

[3] $(1,4)^T = 4(1,2)^T - (3,4)^T$.

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

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

$$= \begin{bmatrix} -6 \\ 4 \\ 1 \end{bmatrix}$$

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

$$\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}$$

has no solution. We can apply the method of Gaussian Elimination as demonstrated in Section 3.2.1, which leads to

$$\left[\begin{array}{cc|c} -1 & 1 & 2 \\ 0 & 1 & 4 \\ 1 & 0 & 3 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & 4 \\ -1 & 1 & 2 \end{array}\right] \qquad R_1 \leftrightarrow R_3$$

$$\to \left[\begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{array}\right] \qquad R_3 + R_1 - R_2 \to R_3$$

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 $\vec{v}^{(2)}$. □

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 $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}\}$ where all of them are from the same vector space $\mathcal{V}$, i.e. $\vec{v}^{(j)} \in \mathcal{V}$ for $j = 1, 2, \ldots, q$, is another set that contains all their possible linear combinations as given by (6.1) in Definition 6.1.3, and is denoted by

$$\mathcal{B} = \text{span}(\beta) = \left\{ \sum_{j=1}^{q} c_j \vec{v}^{(j)} \mid \text{for all possible values of the scalars } c_j \text{ with } \vec{v}^{(j)} \in \beta \right\} \tag{6.3}$$

Again we will limit ourselves to the cases where the scalar 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

$$\mathcal{B} = \text{span}(\beta) = \{A\vec{x} \mid \text{for any } \vec{x} \in \mathbb{R}^q\} \tag{6.4}$$

with $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(q)}]$ is an $n \times q$ matrix and $\vec{x} = (c_1, c_2, \ldots, c_q)^T$ being the coefficient vector.

For example, the span of $\beta_1 = \{(-1, 1)^T\}$ is simply $t(-1, 1)^T$ where $-\infty < t < \infty$, or the line $y = -x$. The span of $\beta_2 = \{(1, 0, 2)^T, (0, 1, -1)^T\}$ (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 to the *spanning set* does not always imply that the corresponding span will be larger. For example, the span of $\beta_3 = \{(1, 0)^T, (0, 1)^T\}$ and $\beta_4 = \{(1, 0)^T, (0, 1)^T, (1, 1)^T, (1, -1)^T\}$ 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 β_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 two 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 combination of the constituent vectors in the spanning set (here β_4). □

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) views 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 $\mathcal{W}$ of $\mathcal{V}$ is always some span (not necessarily unique) of some vectors in $\mathcal{V}$ (assumed a finite dimension).

We leave the proof for showing the span $\rightarrow$ subspace direction in the footnote[4] and that for the subspace $\rightarrow$ span direction in Appendix A.2. Subsequently, we say $\mathcal{W} = \text{span}(\beta)$ is a subspace of $\mathcal{V}$ ($\mathcal{W} \subseteq \mathcal{V}$) *generated* by the set β, and β is known as a ***spanning/generating set*** for $\mathcal{W}$. 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 precisely consists 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 $\mathcal{W} \subseteq \mathcal{V}$ 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 $\vec{v}^{(j)} \in \mathbb{R}^n$ that comprise $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(q)}]$ 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 $\text{span}(\mathcal{B}')$.

[4]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 $\vec{u} = \sum_{j=1}^{q} a_j\vec{v}^{(j)} \in \text{span}(\beta)$ and $\vec{v} = \sum_{j=1}^{q} b_j\vec{v}^{(j)} \in \text{span}(\beta)$ are both in the span for some sets of constants a_j and b_j, then their sum $\vec{u} + \vec{v} = \sum_{j=1}^{q} a_j\vec{v}^{(j)} + \sum_{j=1}^{q} b_j\vec{v}^{(j)} = \sum_{j=1}^{q}(a_j + b_j)\vec{v}^{(j)} = \sum_{j=1}^{q} g_j\vec{v}^{(j)} \in \text{span}(\beta)$ where $g_j = a_j + b_j$ is also in the span, and hence it closed under addition. Similarly, writing $a\vec{v} = a(\sum_{j=1}^{q} b_j\vec{v}^{(j)}) = \sum_{j=1}^{q}(ab_j)\vec{v}^{(j)}$ shows that $a\vec{v} \in \text{span}(\beta)$ 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 *linear independence*, which has profound implications in Linear Algebra. Given a set of vectors, if every one of them cannot 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 the linear independence of q vectors, one may indeed directly show that for every vector $\vec{v}^{(j)}$ in the set, $j = 1, 2, 3, \ldots, q$, 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 to look 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 that significantly simplifies our work.

Theorem 6.1.9. For a set of vectors $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}\}$ where $\vec{v}^{(j)} \in \mathcal{V}$, $j = 1, 2, \ldots, q$ that are from the same vector space, they are linearly independent if and only if, the equation

$$c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \cdots + c_q\vec{v}^{(q)} = \mathbf{0} \tag{6.5}$$

has the trivial solution where all the (real [complex] for a real [complex] vector space) coefficients are zeros (i.e. $c_j = \mathbf{0}$) as its unique solution. Using the language in Properties 6.1.4, if $\vec{v}^{(j)} \in \mathbb{R}^n$ come from the 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 $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ are linearly independent. We can prove the contrapositive where the opposite

of the conclusion, $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ 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 $c_1 = 1, c_2 = -a_2, c_3 = -a_3, \ldots, c_q = -a_q$ 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 $c_1 = 1 \neq 0$ particularly).

The "only if" direction: We want to show the converse that linear independence of $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ 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 $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ 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 β are linearly dependent. □

As a corollary, any set containing the zero vector $\mathbf{0}$ must be linearly dependent. (Why?)[5]

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 deciding 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 checking 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 $\vec{u} = (1, 1, 3)^T$, $\vec{v} = (1, 3, 2)^T$, $\vec{w} = (2, 8, 3)^T$.[6]

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 number of vectors is more than their dimension, then $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\cdots|\vec{v}^{(q)}]$ is a $p \times q$ matrix which has more columns (q) than rows (p). In this case, $A\vec{x} = \mathbf{0}$ must have at least one free variable 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 has been done in Example 6.1.4. Gaussian Elimination still works for any square matrix,

[5] For any such a set $\beta_0 = \{\vec{u}_1, \vec{u}_2, \ldots, \mathbf{0}\}$, 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.

and in the case of linear independence [dependence], A will [not] be reduced to an identity matrix.

In many cases, the number of vectors is indeed not equal to their dimension (e.g different random perturbations (as vectors) of multiple variables (dimension) in an ensemble experiment), 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 Statements, 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 the identity 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 $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(n)}$ of $\mathbb{R}^n$ as in $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\cdots|\vec{v}^{(n)}]$ 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

[6]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.

when determining linear (in)dependence. Let's illustrate this with 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

$$\begin{aligned}
A &= [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\vec{v}^{(4)}|\vec{v}^{(5)}] \\
&= \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 4 & -1 & 1 & -3 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix}
\end{aligned}$$

by concatenating the five vectors column by column. Now we carry out Gaussian Elimination as follows.

$$\begin{aligned}
\begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 4 & -1 & 1 & -3 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix} &\to \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & -3 & -3 & -3 \\ 0 & 0 & -2 & -2 & -2 \end{bmatrix} && \begin{aligned} R_2 - 2R_1 &\to R_2 \\ R_3 - R_1 &\to R_3 \end{aligned} \\
&\to \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 & -2 \end{bmatrix} && -\frac{1}{3}R_2 \to R_2 \\
&\to \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} && R_3 + 2R_2 \to R_3 \\
&\to \begin{bmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} && R_1 - R_2 \to R_1
\end{aligned}$$

The new columns in the above RREF matrix $A_{\text{rref}} = [\vec{v}^{(1)\prime}|\vec{v}^{(2)\prime}|\vec{v}^{(3)\prime}|\vec{v}^{(4)\prime}|\vec{v}^{(5)\prime}]$ follow the exact same dependence relation: $\vec{v}^{(2)\prime} = 2\vec{v}^{(1)\prime}$, $\vec{v}^{(4)\prime} = \vec{v}^{(1)\prime} + \vec{v}^{(3)\prime}$ and $\vec{v}^{(5)\prime} = -\vec{v}^{(1)\prime} + \vec{v}^{(3)\prime}$. $\vec{v}^{(1)\prime}$ and $\vec{v}^{(3)\prime}$ are clearly still linearly independent of each other too. This demonstrates that dependence relations (and by extension linearly independent vectors) are preserved under elementary row operations during Gaussian Elimination. (The detailed argument is put in Appendix A.2) We now introduce a helper theorem so that we may proceed.

Theorem 6.1.11 (Plus/Minus Theorem)**.** Let $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}\}$ 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:

(a) If β is a linearly independent set and $\vec{v}$ is not in span(β), then $\beta \cup \{\vec{v}\}$, 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 β) that can be expressed as a linear combination of other vectors (so β is now a linearly dependent set), then the new set $\beta - \{\vec{w}\}$ remained after removing $\vec{w}$ from β has the same span, i.e.

$$\text{span}(\beta) = \mathcal{B} = \text{span}(\beta - \{\vec{w}\})$$

Proof. We include the proof for (a) as a footnote since (a) is less of a concern.[7] For (b), assign the linearly dependent vector $\vec{v}^{(k)}$ that is being removed where $1 \leq k \leq q$ as $\vec{w}$. We can write $\vec{w} = a_1\vec{v}^{(1)} + a_2\vec{v}^{(2)} + \cdots + a_{k-1}\vec{v}^{(k-1)} + a_{k+1}\vec{v}^{(k+1)} +$

[7] We will prove the contrapositive that, given β is a linearly independent set then, $\beta \cup \{\vec{v}\}$ is linearly dependent only if $\vec{v} := \vec{v}^{(q+1)}$ is in span(β): if $\beta \cup \{\vec{v}\}$ 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 β is required to be linearly independent, $d_{q+1} \neq 0$, for otherwise $d_{q+1} = 0$ and then at least one of the $c_j = d_j$, $j \neq v$, will be non-zero due to the linear dependence of the union set $\beta \cup \{\vec{v}\}$ 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 β alone, so we have $d_1\vec{v}^{(1)} + d_2\vec{v}^{(2)} + \cdots + d_q\vec{v}^{(q)} + d_{q+1}\vec{v}^{(q+1)} = \mathbf{0}$

$\cdots + a_q\vec{v}^{(q)}$ using other vectors in β 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 span(β) with b_j being some coefficients, it can be rewritten as a linear combination of the remaining vectors:

$$\begin{aligned}
\vec{v} &= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_k\vec{v}^{(k)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} \\
&= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} + b_k\vec{v}^{(k)} \\
&= b_1\vec{v}^{(1)} + b_2\vec{v}^{(2)} + \cdots + b_{k-1}\vec{v}^{(k-1)} + b_{k+1}\vec{v}^{(k+1)} + \cdots + b_q\vec{v}^{(q)} \\
&\quad + b_k(a_1\vec{v}^{(1)} + a_2\vec{v}^{(2)} + \cdots + a_{k-1}\vec{v}^{(k-1)} + a_{k+1}\vec{v}^{(k+1)} + \cdots + a_q\vec{v}^{(q)}) \\
&= (b_1 + b_k a_1)\vec{v}^{(1)} + (b_2 + b_k a_2)\vec{v}^{(2)} + (b_{k-1} + b_k a_{k-1})\vec{v}^{(k-1)} \\
&\quad + (b_{k+1} + b_k a_{k+1})\vec{v}^{(k+1)} + \cdots + (b_q + b_k a_q)\vec{v}^{(q)} \\
&\in \text{span}(\beta - \{\vec{v}^{(k)}\}) = \text{span}(\beta - \{\vec{w}\})
\end{aligned}$$

Therefore for all $\vec{v} \in \text{span}(\beta)$, $\vec{v} \in \text{span}(\beta - \{\vec{w}\})$ and hence $\text{span}(\beta) \subseteq \text{span}(\beta - \{\vec{w}\})$. It is trivial to show $\text{span}(\beta - \{\vec{w}\}) \subseteq \text{span}(\beta)$, and thus $\text{span}(\beta) = \text{span}(\beta - \{\vec{w}\}) = \mathcal{B}$. This part of the theorem is very relevant to the span of sets β_3 and β_4 in the previous Example 6.1.3. □

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 manner 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

---

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 $\vec{v}^{(j)} \in \beta$, $1 \leq j \leq q$.

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 $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}\}$, $\vec{v}^{(j)} \in \mathcal{V}$ for $j = 1, 2, \ldots, q$, which are linearly independent, any vector $\vec{v} \in \text{span}(\beta)$ in their span can be written as a unique linear combination of these generating vectors in β. Otherwise, if the vectors in β 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.[8] Return to our example, where

$$\begin{aligned} A &= [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\vec{v}^{(4)}|\vec{v}^{(5)}] \\ &= \begin{bmatrix} 1 & 2 & 1 & 2 & 0 \\ 2 & 4 & -1 & 1 & -3 \\ 1 & 2 & -1 & 0 & -2 \end{bmatrix} \end{aligned}$$

We have found that the corresponding RREF is

$$A_{\text{rref}} = \begin{bmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[8] We will show the first part only. Since $\vec{v}$ already belongs to $\text{span}(\beta)$, it must be possible to express $\vec{v}$ as some linear combination(s) of vectors $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ in β 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 β 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 that are not exactly the same. Subtracting one expression by another leads to

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 linearly 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 exactly the corresponding columns in the RREF which retain the dependence relations, with row(s) of all zeros removed. For instance,

$$\begin{aligned}
\vec{v}^{(5)} &= -\vec{v}^{(1)} + \vec{v}^{(3)} \\
\begin{bmatrix} 0 \\ -3 \\ 2 \end{bmatrix} &= -1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} && \text{(Properties 6.1.4)} \\
&= [\vec{v}^{(1)}|\vec{v}^{(3)}] \begin{bmatrix} -1 \\ 1 \end{bmatrix} = C \begin{bmatrix} -1 \\ 1 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
(d_1\vec{v}^{(1)} + d_2\vec{v}^{(2)} + d_3\vec{v}^{(3)} + \cdots + d_q\vec{v}^{(q)}) \\
- (g_1\vec{v}^{(1)} + g_2\vec{v}^{(2)} + g_3\vec{v}^{(3)} + \cdots + g_q\vec{v}^{(q)})
\end{aligned} = \vec{v} - \vec{v}$$

$$(d_1 - g_1)\vec{v}^{(1)} + (d_2 - g_2)\vec{v}^{(2)} + (d_3 - g_3)\vec{v}^{(3)} + \cdots + (d_q - g_q)\vec{v}^{(q)} = \mathbf{0}$$

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. This contradicts our assumed linear independence of β and hence the linear combination of $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(q)}$ to generate $\vec{v}$ must be unique.

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_{rref}, then similarly

$$\vec{v}^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = CR_1$$

$$\vec{v}^{(2)} = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = CR_2$$

$$\vec{v}^{(3)} = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = CR_3$$

$$\vec{v}^{(4)} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = CR_4$$

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?)[9] 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 an RREF of A_{rref}, is given by $A = CR$, where C contains the r linearly independent columns of A at which the r leading 1s of A_{rref} are located, and R is simply the first r rows of A_{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.

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

[9] 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),

$$\begin{aligned}
A[B_1|B_2|\cdots|B_n] &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & & a_{2r} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mr} \end{bmatrix} \left[\begin{array}{c|c|c|c} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rn} \end{array}\right] \\
&= \left[\begin{array}{c|c|c|c} \sum_{k=1}^r a_{1k}b_{k1} & \sum_{k=1}^r a_{1k}b_{k2} & \cdots & \sum_{k=1}^r a_{1k}b_{kn} \\ \sum_{k=1}^r a_{2k}b_{k1} & \sum_{k=1}^r a_{2k}b_{k2} & & \sum_{k=1}^r a_{2k}b_{kn} \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^r a_{mk}b_{k1} & \sum_{k=1}^r a_{mk}b_{k2} & \cdots & \sum_{k=1}^r a_{mk}b_{kn} \end{array}\right] \\
&= [AB_1|AB_2|\cdots|AB_n]
\end{aligned}$$

where B_j now denotes the j-th column of B:

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.

$$\left[\begin{array}{ccc|c} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 1 & -1 & 2 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 \end{array}\right] \qquad R_1 \leftrightarrow R_4$$

$$\to \left[\begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 3 & -1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 3 & -4 & 0 \end{array}\right] \qquad \begin{array}{r} R_2 - R_1 \to R_2 \\ R_3 + R_1 \to R_3 \\ R_4 - 2R_1 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 3 & -4 & 0 \end{array}\right] \qquad \frac{1}{3}R_2 \to R_2$$

$$B_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{rj} \end{bmatrix} \qquad \text{and hence } AB_j = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & & a_{2r} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mr} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{rj} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{r} a_{1k} b_{kj} \\ \sum_{k=1}^{r} a_{2k} b_{kj} \\ \vdots \\ \sum_{k=1}^{r} a_{mk} b_{kj} \end{bmatrix}$$

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

$$\rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 2 & 0 \\ 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -3 & 0 \end{array}\right] \qquad \frac{1}{3}R_3 \rightarrow R_3$$

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

(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. The backward phase concerning the matrix A itself is instantaneous, yielding its RREF:

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

By Properties 6.1.13, the CR Factorization of A is then trivially

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

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[10], then its

[10] Bear in mind that it is necessary to have $m \geq n$.

RREF will be in the form of

$$A_{\text{rref}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \end{bmatrix}$$

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.

$$\begin{aligned}
\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ -1 & 1 & -2 \\ 1 & -1 & 2 \end{bmatrix} &\to \begin{bmatrix} 1 & -1 & 2 \\ 1 & 2 & -1 \\ -1 & 1 & -2 \\ 2 & 1 & 1 \end{bmatrix} && R_1 \leftrightarrow R_4 \\
&\to \begin{bmatrix} 1 & -1 & 2 \\ 0 & 3 & -3 \\ 0 & 0 & 0 \\ 0 & 3 & -3 \end{bmatrix} && \begin{aligned} R_2 - R_1 &\to R_2 \\ R_3 + R_1 &\to R_3 \\ R_4 - 2R_1 &\to R_4 \end{aligned} \\
&\to \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 3 & -3 \end{bmatrix} && \frac{1}{3}R_2 \to R_2 \\
&\to \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} && R_4 - 3R_2 \to R_4
\end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad R_1 + R_2 \rightarrow R_1$$

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

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

□

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 $\hat{e}^{(1)}, \hat{e}^{(2)}, \ldots, \hat{e}^{(n)}$ 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 $\vec{x} = (x_1, x_2, \ldots, x_n)^T = x_1\hat{e}^{(1)} + x_2\hat{e}^{(2)} + \cdots + x_n\hat{e}^{(n)}$ 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*** $\mathcal{S} = \{\hat{e}^{(1)}, \hat{e}^{(2)}, \ldots, \hat{e}^{(n)}\}$. 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 β for $\mathbb{R}^n$ should consists of n vectors $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}\}$ 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 $q \geq n$.[11] Hence the number of vectors q must be equal to n.

[11] To formally show this, express the span of q $\mathbb{R}^n$ vectors $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + c_3\vec{v}^{(3)} + \cdots + c_q\vec{v}^{(q)}$ by $A\vec{x}$ where $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\vec{v}^{(3)}|\cdots|\vec{v}^{(q)}]$ is an $n \times q$ matrix and $\vec{x} = (c_1, c_2, c_3, \ldots, c_q)^T$ 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 $\mathbb{R}^n$.

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 $\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}$ are linear independent with $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(n)}]$ being an $n \times n$ 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 $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}\}$ (Definition 6.1.5) so it implies that these vectors can generate the entire $\mathbb{R}^n$.

Spanning $\mathbb{R}^n \rightarrow$ Linear Independence: Assume the opposite of the implication that $\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}$ are linearly dependent, then by (c) and (e) of Theorem 6.1.10 the reduced row echelon form of $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(n)}]$ is not the identity matrix and contains at least one row of full zeros. Following a logic similar to Footnote 11, these vectors cannot span $\mathbb{R}^n$ and the contrapositive is proved. □

Example 6.2.1. Show that $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\} = \{(1,2,1)_S^T, (-1,1,0)_S^T, (1,-1,2)_S^T\}$ forms a basis for $\mathbb{R}^3$ and express $\vec{v}$ in the β basis (denoted by $[\vec{v}]_\beta$) where $[\vec{v}]_S = (2,1,2)_S^T$, the subscript S $[\beta]$ emphasizes that the coordinates are relative to the standard basis $\mathcal{S}$ [the new basis β].

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 β are linearly independent. By (b) to (e) of Theorem 6.1.10, we can simply check if $\det(A)$ is non-zero where

$$A = \left[[\vec{v}^{(1)}]_S|[\vec{v}^{(2)}]_S|[\vec{v}^{(3)}]_S\right] = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix}$$

A simple calculation reveals that $\det(A) = 6 \neq 0$, so β is indeed a valid basis for $\mathbb{R}^3$. To express $(2,1,2)_S^T$ in β is to find $[\vec{v}]_\beta = ([v_1]_\beta, [v_2]_\beta, [v_3]_\beta)_\beta^T$ where $[v_j]_\beta$ is the j-th component (coefficient) of $\vec{v}$ in the β 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:

$$[v_1]_\beta(\vec{v}^{(1)}) + [v_2]_\beta(\vec{v}^{(2)}) + [v_3]_\beta(\vec{v}^{(3)}) = \vec{v}$$
$$[v_1]_\beta(1,2,1)_S^T + [v_2]_\beta(-1,1,0)_S^T + [v_3]_\beta(1,-1,2)_S^T = (2,1,2)_S^T$$

or put in matrix form,

$$[v_1]_\beta[\vec{v}^{(1)}]_S + [v_2]_\beta[\vec{v}^{(2)}]_S + [v_3]_\beta[\vec{v}^{(3)}]_S = [\vec{v}]_S$$
$$A[\vec{v}]_\beta = \left[[\vec{v}^{(1)}]_S | [\vec{v}^{(2)}]_S | [\vec{v}^{(3)}]_S\right] \begin{bmatrix} [v_1]_\beta \\ [v_2]_\beta \\ [v_3]_\beta \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} [v_1]_\beta \\ [v_2]_\beta \\ [v_3]_\beta \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$$

We can either use matrix inverse or Gaussian Elimination to solve for the $[v_j]_\beta$, yielding $[v_1]_\beta = 1$, $[v_2]_\beta = -\frac{1}{2}$, $[v_3]_\beta = \frac{1}{2}$, and hence $[\vec{v}]_\beta = (1, -\frac{1}{2}, \frac{1}{2})_\beta^T$. Notice that the matrix equation

$$A[\vec{v}]_\beta = \left[[\vec{v}^{(1)}]_S | [\vec{v}^{(2)}]_S | [\vec{v}^{(3)}]_S\right] \begin{bmatrix} [v_1]_\beta \\ [v_2]_\beta \\ [v_3]_\beta \end{bmatrix} = [\vec{v}]_S \tag{6.6}$$

shows that the matrix A transforms the coordinate system from the β to $\mathcal{S}$ basis in which a given vector $\vec{v}$ is expressed, and we will write $A = P_\beta^S$ (thus $P_\beta^S[\vec{v}]_\beta = [\vec{v}]_S$) for clarity in the future. Also, the vector $\vec{v}$ is still the same one despite having different coordinate representations since only the reference coordinate frame/system is changed. For this, we will henceforth use the equivalence symbol to write along the lines of $\vec{v} \equiv [\vec{v}]_\beta \equiv [\vec{v}]_S$.

Notice that $P_\beta^S[\vec{v}^{(j)}]_\beta = P_\beta^S(e^{(j)})_\beta$ returns the j-th basis vector in β (j-th column of P_β^S) expressed in the original standard basis $\mathcal{S}$, namely $[\vec{v}^{(j)}]_S$, where $(e^{(j)})_\beta = [\vec{v}^{(j)}]_\beta$ is the numeric coordinate representation (emphasized by the

absence of the hat symbol over e) of the j-th basis vector in the β coordinate system with the j-th component being 1 and other being 0. For instance[12],

$$P_\beta^S[\vec{v}^{(2)}]_\beta = P_\beta^S(e^{(2)})_\beta = \begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = [\vec{v}^{(2)}]_S$$

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. □

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 $\mathbb{R}^n$. 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

[12] In contrast, the standard unit vector $\hat{e}^{(2)}$ with a hat is deemed as a "true" vector with $[\hat{e}^{(2)}]_S = (0, 1, 0)_S^T$. Then, $P_\beta^S[\hat{e}^{(2)}]_\beta = [\hat{e}^{(2)}]_S$ and thus

$$[\hat{e}^{(2)}]_\beta = (P_\beta^S)^{-1}[\hat{e}^{(2)}]_S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{5}{6} & \frac{1}{6} & \frac{1}{2} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \left(\frac{1}{3}, \frac{1}{6}, -\frac{1}{6}\right)_\beta^T$$

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 $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(q)}\}$, its basis can be found by applying CR Factorization (Properties 6.1.13) over $A = \left[\vec{v}^{(1)} | \vec{v}^{(2)} | \ldots | \vec{v}^{(q)}\right]$. 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 β for the subspace generated by $\gamma = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\}$. 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.

$$\begin{aligned} \begin{bmatrix} 1 & 1 & 0 & -2 \\ 0 & -1 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & -2 & 0 & 1 \end{bmatrix} &\to \begin{bmatrix} 1 & 1 & 0 & -2 \\ 0 & -1 & 0 & 1 \\ 0 & -1 & 1 & 4 \\ 0 & -3 & 0 & 3 \end{bmatrix} && \begin{aligned} R_3 - 2R_1 &\to R_3 \\ R_4 - R_1 &\to R_4 \end{aligned} \\ &\to \begin{bmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 1 & 4 \\ 0 & -3 & 0 & 3 \end{bmatrix} && -R_2 \to R_2 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 1 & 0 & -2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{matrix} R_3 + R_2 \rightarrow R_3 \\ R_4 + 3R_2 \rightarrow R_4 \end{matrix}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad R_1 - R_2 \rightarrow R_1$$

So a possible basis for the subspace: $\text{span}(\gamma)$, contains the first three generating vectors in γ, so that $\beta = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T, (0, 0, 1, 0)^T\}$. To express $\vec{v} = (3, -1, 4, 0)^T$ in this basis, we need to find $[\vec{v}]_\beta$ just like in Example 6.2.1, which is derived by

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 2 & 1 & 1 \\ 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} [v_1]_\beta \\ [v_2]_\beta \\ [v_3]_\beta \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 4 \\ 0 \end{bmatrix}$$

We repeat the same steps of Gaussian Elimination over the augmented matrix where now the left portion consists of the three linearly independent basis vectors only.

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -1 \\ 2 & 1 & 1 & 4 \\ 1 & -2 & 0 & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -1 \\ 0 & -1 & 1 & -2 \\ 0 & -3 & 0 & -3 \end{array}\right] \qquad \begin{matrix} R_3 - 2R_1 \rightarrow R_3 \\ R_4 - R_1 \rightarrow R_4 \end{matrix}$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 1 & -2 \\ 0 & -3 & 0 & -3 \end{array}\right] \qquad -R_2 \rightarrow R_2$$

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

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_1 - R_2 \rightarrow R_1$$

The last full-zero row is consistent and $[\vec{v}]_\beta = ([v_1]_\beta, [v_2]_\beta, [v_3]_\beta)_B^T = (2, 1, -1)_B^T$. 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 $(-2, 1, 0, 1)^T$. □

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.[13] 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 numbers 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.[14]

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.

[13] 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 as rows and the replacement is effectively multiplications and additions of rows), plus (a) of Properties 6.2.7 for linear independence.

[14] It, along with other results below, actually comes from a more general theorem called the *Steinitz Replacement Theorem* which is proved in Appendix A.2.

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 $\mathcal{V}$ is generated by 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,

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 γ with a finite amount of vectors, then some subset γ' of γ 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 γ_1 of $\mathcal{V}$ with $m \leq n$ vectors can be extended to a basis for $\mathcal{V}$, i.e. there exists another subset γ_2 of $\mathcal{V}$ with $n - m$ (linearly independent) vectors such that $\beta = \gamma_1 \cup \gamma_2$ 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 some 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

$$\mathcal{W}_1 + \mathcal{W}_2 = \{\vec{w}_1 + \vec{w}_2 \mid \text{all } \vec{w}_1 \in \mathcal{W}_1, \vec{w}_2 \in \mathcal{W}_2\} \tag{6.7}$$

consisted of all possible vectors resulting from addition between any pair of vectors $\vec{w}_1 \in \mathcal{W}_1, \vec{w}_2 \in \mathcal{W}_2$ from $\mathcal{W}_1, \mathcal{W}_2 \subseteq \mathcal{V}$ respectively. Note that $(\mathcal{W}_1 + \mathcal{W}_2) \subseteq \mathcal{V}$ is also 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

$$\mathcal{W}_1 + \mathcal{W}_2 + \cdots + \mathcal{W}_n = \{\vec{w}_1 + \vec{w}_2 + \cdots \vec{w}_n \mid \vec{w}_j \in \mathcal{W}_j, 1 \leq j \leq n\} \tag{6.8}$$

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 $\mathcal{W}_1, \mathcal{W}_2 \subseteq \mathcal{V}$ is their subspace sum $\mathcal{W}_1 + \mathcal{W}_2$ as defined by (6.7) in Definition 6.2.8, which additionally satisfies $\mathcal{W}_1 \cap \mathcal{W}_2 = \{\mathbf{0}\}$ (their intersection is empty), 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 $\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(p)}$ and $\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(q)}$ 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***[15] to each other in $\mathcal{W}_1 \oplus \mathcal{W}_2$, as any non-zero vector from the direct sum $\mathcal{W}_1 \oplus \mathcal{W}_2$ that has purely zero components in one of the two smaller subspaces, let's say $\mathcal{W}_1$, will always completely reside in the other subspace, $\mathcal{W}_2$.

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 $\mathcal{V}$.

[15] A subspace complement is not the same as the set-theoretic complement.

As a counter-example, consider Example 6.2.2, suppose that $\mathcal{W}_1 = \text{span}(\beta_1) = \text{span}(\{(1,0,2,1)^T, (1,-1,1,-2)^T\})$ and $\mathcal{W}_2 = \text{span}(\beta_2) = \text{span}(\{(0,0,1,0)^T, (-2,1,0,1)^T\})$ are the two subspaces spanned by the first/last two vectors in γ respectively. It is not hard to see that β_1 and β_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 generate a three-dimensional subspace. Hence $\dim(\mathcal{W}_1 + \mathcal{W}_2) = \dim(\text{span}(\gamma)) = 3 \neq 4 = 2 + 2 = \dim(\mathcal{W}_1) + \dim(\mathcal{W}_2)$, and therefore they cannot form a direct sum. Geometrically, these two subspaces are like two planes intersecting along a straight line.

The direct sum of multiple subspaces is 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 is 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 equal 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 $\mathcal{W}_2$

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

is pivotal after Gaussian Elimination, as follows.

$$\begin{aligned}
\begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 2 & 0 & 1 & -1 \\ 1 & 0 & 0 & -2 \\ 0 & -1 & 0 & 1 \end{bmatrix} &\to \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & -4 & 1 & -1 \\ 0 & -2 & 0 & -2 \\ 0 & -1 & 0 & 1 \end{bmatrix} && \begin{aligned} R_3 - 2R_1 &\to R_3 \\ R_4 - R_1 &\to R_4 \end{aligned} \\
&\to \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 13 & -1 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 3 & 1 \end{bmatrix} && \begin{aligned} R_3 + 4R_2 &\to R_3 \\ R_4 + 2R_2 &\to R_4 \\ R_5 + R_2 &\to R_5 \end{aligned} \\
&\to \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 13 & -1 \end{bmatrix} && R_3 \leftrightarrow R_5 \\
&\to \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 13 & -1 \end{bmatrix} && \frac{1}{3}R_3 \to R_3
\end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & -4 \\ 0 & 0 & 0 & -\frac{16}{3} \end{bmatrix} \qquad \begin{aligned} R_4 - 6R_3 &\rightarrow R_4 \\ R_5 - 13R_3 &\rightarrow R_5 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{16}{3} \end{bmatrix} \qquad -\frac{1}{4}R_4 \rightarrow R_4$$

$$\rightarrow \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad R_5 + \frac{16}{3}R_4 \rightarrow R_5$$

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

$$\left[\begin{array}{cccc|c} 1 & 2 & 0 & 0 & 1 \\ 0 & 1 & 3 & 0 & 1 \\ 2 & 0 & 1 & -1 & -3 \\ 1 & 0 & 0 & -2 & 0 \\ 0 & -1 & 0 & 1 & -1 \end{array}\right]$$

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 exact same reduction steps performed above, which would lead to

$$\left[\begin{array}{cccc|c} 1 & 2 & 0 & 0 & 1 \\ 0 & 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 & -\frac{1}{4} \\ 0 & 0 & 0 & 0 & -\frac{7}{3} \end{array}\right]$$

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, $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3 = \mathbb{R}^5$. □

The importance of direct sum is that the coordinates of two vectors in the 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 $\beta_1 = \{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T\}$ and $\beta_2 = \{(0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\}$ as bases for $\mathcal{W}_1$ and $\mathcal{W}_2$, the coordinates $(1, 2)^T_{\beta_1}$ and $(1, -1)^T_{\beta_2}$ in the β_1 and β_2 system, represent the vectors

$$\begin{aligned} (1, 0, 2, 1, 0)^T + 2(2, 1, 0, 0, -1)^T &= (5, 2, 2, 1, -2)^T \\ \text{and } (0, 3, 1, 0, 0)^T - (0, 0, -1, -2, 1)^T &= (0, 3, 2, 2, -1)^T \end{aligned}$$

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 β_1 and β_2 will be

$$\beta_1 \cup \beta_2 = \begin{array}{l} \{(1, 0, 2, 1, 0)^T, (2, 1, 0, 0, -1)^T, \\ (0, 3, 1, 0, 0)^T, (0, 0, -1, -2, 1)^T\} \end{array}$$

and the merged coordinates $(1, 2, 1, -1)^T_{\beta_1+\beta_2}$ 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 $(1, 2, 1, -1)^T_{\beta_1+\beta_2}$ is unique as $\beta_1 \cup \beta_2$ 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 $(1, 2, 1, -1)^T_{\beta_1+\beta_2} = (1, 2)^T_{\beta_1} + (1, -1)^T_{\beta_2}$.

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 instead as a demonstration, where $\beta_1 = \{(1, 0, 2, 1)^T, (1, -1, 1, -2)^T\}$ and $\beta_2 = \{(0, 0, 1, 0)^T, (-2, 1, 0, 1)^T\}$ and we have already shown that they are not linearly independent when combined. Take

$$\begin{aligned}(1, 2)^T_{\beta_1} &= (1, 0, 2, 1)^T + 2(1, -1, 1, -2)^T = (3, -2, 4, -3)^T \\ \text{and } (-1, 1)^T_{\beta_2} &= -(0, 0, 1, 0)^T + (-2, 1, 0, 1)^T = (-2, 1, -1, 1)^T\end{aligned}$$

Their concatenated sum will be

$$\begin{aligned}&(1, 2, -1, 1)^T_{\beta_1+\beta_2} \\ &= (1, 0, 2, 1)^T + 2(1, -1, 1, -2)^T - (0, 0, 1, 0)^T + (-2, 1, 0, 1)^T \\ &= (3, -2, 4, -3)^T + (-2, 1, -1, 1)^T \\ &= (1, -1, 3, -2)^T\end{aligned}$$

but

$$\begin{aligned}(0, 1)^T_{\beta_1} + (2, 0)^T_{\beta_2} = (0, 1, 2, 0)^T_{\beta_1+\beta_2} &= (1, -1, 1, -2)^T + 2(0, 0, 1, 0)^T \\ &= (1, -1, 3, -2)^T\end{aligned}$$

is aptly an alternative representation.

Another aspect of direct sum is that all finite-dimensional, particularly n-dimensional vector spaces with some basis consisting 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.

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.*

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 an $m \times n$ 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 $\vec{v}^{(j)}$, $j = 1, 2, \ldots, n$. 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, $\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)} \in \mathbb{R}^m$ as in

$$A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(n)}] \tag{6.9}$$

Meanwhile its row space is the subspace spanned by its m row vectors $\vec{w}^{(1)T}, \vec{w}^{(2)T}, \ldots, \vec{w}^{(m)T} \in \mathbb{R}^n$ as in

$$A = \begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \hline \vdots \\ \hline \vec{w}^{(m)T} \end{bmatrix} \tag{6.10}$$

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 the 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 independence between column vectors, hence

Properties 6.3.3. Elementary row operations do 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 do 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 resulting 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

$$\begin{aligned} &= \operatorname{span}(\{\ldots, \vec{w}^{(p)}, \ldots, \vec{w}^{(q)}, \ldots, \vec{w}^{(r)}\}) \\ C(A^T) &= \operatorname{span}(\{\ldots, \vec{w}^{(p)}, \ldots, \vec{w}^{(q)}, \ldots\}) \\ &= \operatorname{span}(\{\ldots, \vec{w}^{(r)}, \ldots, \vec{w}^{(q)}, \ldots\}) = C(A'^T) \end{aligned}$$

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 its reduced row echelon form has the same number of (linearly independent) vectors for 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 0s 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 dimensions of row/column space of the reduced row echelon form are both 3. With these observations, we can derive the desired result, sometimes referred to as *"Column rank equals 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.

$$\dim(C(A)) = \dim(C(A^T)) \tag{6.11}$$

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. □

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

$$\begin{bmatrix} 1 & 1 & -2 & 1 \\ 1 & 2 & 1 & -1 \\ 1 & 0 & -5 & 3 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & -2 & 1 \\ 0 & 1 & 3 & -2 \\ 0 & -1 & -3 & 2 \end{bmatrix} \qquad \begin{aligned} R_2 - R_1 &\to R_2 \\ R_3 - R_1 &\to R_3 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 1 & -2 & 1 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad R_3 - R_2 \rightarrow R_3$$

$$\rightarrow \begin{bmatrix} 1 & 0 & -5 & 3 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad R_1 - R_2 \rightarrow R_1$$

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

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ -2 & 1 & -5 \\ 1 & -1 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \begin{aligned} &R_3 + 5R_1 - 3R_2 \rightarrow R_3, \\ &R_4 - 3R_1 + 2R_2 \rightarrow R_4 \end{aligned}$$

and the next step is straightforward:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad R_2 - R_1 \rightarrow R_2$$

$$\rightarrow \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad R_1 - R_2 \rightarrow R_1$$

which reveals that the first two columns (representing the first two row vectors in A) are linearly independent and the third column (representing 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. □

Finally, in view of Definitions 6.1.5 and 6.3.1, the analysis of solving linear systems in Section 3.2 can be summarized as follows.

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 at the end of the 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 solution space. In Section 3.2.1 we know that it is made up of the particular and complementary solution, where the latter corresponds 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 $\mathbb{R}^n$, and this subspace is then called the ***null space*** of A. To verify, 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 $a\vec{x}_0^{(1)} \in \vec{x}_0$.

Definition 6.3.8 (Null Space)**.** For an $m \times n$ real matrix A, its null space $\mathcal{N}(A)$ is the subspace consisting 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 number 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 applying elementary matrices. Hence the null space and nullity of A will be the same as that of A_{rref}. Previously we have assigned free variables to the non-pivotal columns (let's say there is k of them) of A_{rref} and derive $\vec{x}_0$ where they are generated by k pairs of free variables and column vectors $(\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \ldots, \vec{x}_0^{(k)})$. 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 free variables (see Example 3.2.6 for an instance). We claim that they also span the entire null space of A_{rref}.[16] Hence by the definition given in Section 6.2.2 they form a basis for the null space of A_{rref} 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 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

$$\dim(C(A)) + \dim(N(A)) = \text{rank}(A) + \text{nullity}(A) = n \tag{6.12a}$$

$$= \dim(C(A^T)) + \dim(N(A)) \tag{6.12b}$$

[16] Assume the contrary so that the span of $\{\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \ldots, \vec{x}_0^{(k)}\}$ 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 $\{\vec{x}_0^{(1)}, \vec{x}_0^{(2)}, \ldots, \vec{x}_0^{(k)}, \vec{x}_0^{(k+1)}\}$ is a 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 13), and by doing so non-zero components of $\vec{x}_0^{(k+1)}$ only appear in positions corresponding to leading 1s in A_{rref}. This causes a contradiction since $A_{\text{rref}}\vec{x}_0^{(k+1)} = \mathbf{0}$ then implies that there exists a non-trivial dependence relation between the pivotal column vectors themselves.

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*[17], by Definition 6.3.2 and 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.

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 $\mathcal{N}(A)$ 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

$$A = \left[\begin{array}{c} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \hline \vdots \\ \hline \vec{w}^{(m)T} \end{array}\right]$$

and the corresponding homogeneous system $A\vec{x} = \mathbf{0}$ then can be written as

$$A\vec{x} = \left[\begin{array}{c} \vec{w}^{(1)T} \\ \hline \vec{w}^{(2)T} \\ \hline \vdots \\ \hline \vec{w}^{(m)T} \end{array}\right] \vec{x} = \begin{bmatrix} \vec{w}^{(1)T} \cdot \vec{x} \\ \vec{w}^{(2)T} \cdot \vec{x} \\ \vdots \\ \vec{w}^{(m)T} \cdot \vec{x} \end{bmatrix} = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

[17] An $m \times n$ matrix A is said to be full-rank if $\text{rank}(A) = \min(m, n)$.

where for a solution $\vec{x} = \vec{x}_0$ in the null space of A, each of the dot products $\vec{w}^{(i)T} \cdot \vec{x}_0 = 0$, $i = 1, 2, \ldots, m$, has to be equal to zero. Any vector in the row space of A can be expressed as $\vec{w} = c_1\vec{w}^{(1)} + c_2\vec{w}^{(2)} + \cdots + c_m\vec{w}^{(m)}$ by Definitions 6.3.1 and 6.1.5, and subsequently, its dot product with $\vec{x}_0$

$$\begin{aligned}\vec{w}^T \cdot \vec{x}_0 &= (c_1\vec{w}^{(1)} + c_2\vec{w}^{(2)} + \cdots + c_m\vec{w}^{(m)})^T \cdot \vec{x}_0 \\ &= c_1(\vec{w}^{(1)T} \cdot \vec{x}_0) + c_2(\vec{w}^{(2)T} \cdot \vec{x}_0) + \cdots + c_m(\vec{w}^{(m)T} \cdot \vec{x}_0) \\ &= c_1(0) + c_2(0) + \cdots + c_m(0) = 0\end{aligned}$$

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 $N(A)$. □

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

$$\begin{aligned}\vec{u}^{(1)} \cdot (c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)}) = c_1(\vec{u}^{(1)} \cdot \vec{u}^{(1)}) + c_2(\vec{u}^{(1)} \cdot \vec{u}^{(2)}) &= \vec{u}^{(1)} \cdot \mathbf{0} \\ c_1\left\|\vec{u}^{(1)}\right\|^2 + c_2(0) &= c_1\left\|\vec{u}^{(1)}\right\|^2 = 0\end{aligned}$$

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. □

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 straightforward, 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 $\dim \mathcal{C}(A) = \text{rank}(A) = 2$. So $\text{rank}(A) + \text{nullity}(A) = 2 + 2 = 4$, and Theorem 6.3.9 is true. □

► Short Exercise: Show that[18]

$$\dim(\mathcal{C}(A^T)) + \dim(\mathcal{N}(A^T)) = \dim(\mathcal{C}(A)) + \dim(\mathcal{N}(A^T)) = m$$

$\mathcal{N}(A^T)$ 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 $\mathcal{C}(A^T) \oplus \mathcal{N}(A)$ 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 $\mathcal{C}(A^T) \subseteq \mathbb{R}^n$, $\mathcal{N}(A) \subseteq \mathbb{R}^n$ and hence $\mathcal{C}(A^T) \oplus \mathcal{N}(A) \subseteq \mathbb{R}^n$, from Theorem 6.3.9 and Properties 6.2.5, we conclude that $\mathcal{C}(A^T) \oplus \mathcal{N}(A)$ 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 same can be 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

$$\mathcal{C}(A^T) \oplus \mathcal{N}(A) = \mathbb{R}^n \qquad \mathcal{C}(A) \oplus \mathcal{N}(A^T) = \mathbb{R}^m \tag{6.13}$$

where $\mathcal{C}(A^T)^\perp = \mathcal{N}(A)$, $\mathcal{N}(A)^\perp = \mathcal{C}(A^T)$, and $\mathcal{C}(A)^\perp = \mathcal{N}(A^T)$, $\mathcal{N}(A^T)^\perp = \mathcal{C}(A)$, 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.

Figure 6.3.: *The relationships between the four fundamental subspaces for an $m \times n$ real matrix A of rank r: the row space $\mathcal{C}(A^T)$, null space $\mathcal{N}(A)$, column space $\mathcal{C}(A)$, left null space $\mathcal{N}(A^T)$. Any vector $\vec{x} \in \mathbb{R}^n$ can be partitioned into $\vec{x} = \vec{x}_r + \vec{x}_n$ uniquely, where $\vec{x}_r \in \mathcal{C}(A^T) \subseteq \mathbb{R}^n$ and $\vec{x}_n \in \mathcal{N}(A) \subseteq \mathbb{R}^n$ 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 $\vec{h} \in \mathcal{C}(A) \subseteq \mathbb{R}^m$ in the column space of A. The total effect on $\vec{x}$ multiplied by A, is the sum of the two responses: $A\vec{x} = A(\vec{x}_r + \vec{x}_n) = A\vec{x}_r + A\vec{x}_n = \vec{h} + \mathbf{0} = \vec{h}$.*

Properties 6.3.14. For two (real) $m \times r$ and $r \times n$ matrices A and B, the rank of AB

$$\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B)) \tag{6.14}$$

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

$$\begin{aligned} AB &= \left[A_1|A_2|\cdots|A_r\right] \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rn} \end{bmatrix} \\ &= \left[b_{11}A_1 + b_{21}A_2 + \cdots + b_{r1}A_r | \cdots | b_{1n}A_1 + b_{2n}A_2 + \cdots + b_{rn}A_r\right] \end{aligned}$$

(where A_j is the j-th column of A) which shows that the columns of AB are linear combinations of columns 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 $\operatorname{rank}(AB) = \dim(C(AB)) \leq \dim(C(A)) = \operatorname{rank}(A)$. The same argument on the row space of AB and B similarly shows that $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$ and the desired inequality follows. □

► Short Exercise: Show that $\operatorname{rank}(AB) = \operatorname{rank}(A)$ if B is an invertible square matrix.[19]

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

[18]Replace A by A^T in Theorem 6.3.9 to get $\dim(C(A^T)) + \dim(\mathcal{N}(A^T)) = m$.

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())
```

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([
```

[19] $\text{rank}(A) = \text{rank}(ABB^{-1}) \leq \text{rank}(AB) \leq \text{rank}(A)$ by applying Properties 6.3.14 twice. So $\text{rank}(AB)$ is "sandwiched" by and must be equal to $\text{rank}(A)$.

```
[-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 $\vec{v}^{(1)} = \begin{bmatrix}1\\1\\0\end{bmatrix}$, $\vec{v}^{(2)} = \begin{bmatrix}0\\1\\1\end{bmatrix}$, $\vec{v}^{(3)} = \begin{bmatrix}1\\0\\1\end{bmatrix}$, 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 $\gamma = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \vec{v}^{(4)}\}$ 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 γ.

Exercise 6.4 For the basis $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$, 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 β frame,

(b) Transform $(1, -1, 1)^T_\beta$ from the β system back to the the standard basis $\mathcal{S}$.

Exercise 6.5 Show that $\beta = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \vec{w}^{(3)}\}$, 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 into what a matrix actually represents in the big picture. Matrices, by nature, are the rules of *linear transformations* (or *linear mappings*) between two vector spaces. We are going to study several special types of linear transformations, and this ultimately reveals the relationship between any n-dimensional real vector space and the real n-space $\mathbb{R}^n$, as an *isomorphism*. We will 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. This is needed in Earth Science as we often have to switch between different coordinate systems and particularly want to work with the nicer ones that facilitate computation.

7.1. Ideas of Linear Transformations

7.1.1. Linear Transformations 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 vectors

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 (also called a linear map) from a vector space $\mathcal{U}$ to another vector space $\mathcal{V}$ is a mapping: $T : \mathcal{U} \to \mathcal{V}$, such that for all possible pairs of 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 then $T(a\vec{u}^{(1)} + b\vec{u}^{(2)}) = aT(\vec{u}^{(1)}) + bT(\vec{u}^{(2)})$, 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 $\vec{u} \equiv [\vec{u}]_\beta$[1] in $\mathcal{U}$ by n coordinates using some basis β (similarly for $\vec{v} \equiv [\vec{v}]_\eta \in \mathcal{V}$ that has m coordinates in some basis η), then $[T]^\eta_\beta = A$, where A is any $m \times n$ matrix, will satisfy the requirements of and is a linear transformation from $\mathcal{U}$ to $\mathcal{V}$ according to the rule $T(\vec{u}) \equiv [T]^\eta_\beta [\vec{u}]_\beta = A[\vec{u}]_\beta$. (Short Exercise: show this satisfies the conditions outlined in Definition 7.1.1 above![2]) 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 $\beta = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n)}\}$ for $\mathcal{U}$ (again, similarly we have $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(m)}\}$ for $\mathcal{V}$). For each $\vec{u}^{(j)}$, $j = 1, 2, \ldots, n$, denote $\vec{w}^{(j)} = T(\vec{u}^{(j)})$ as the

[1] We use "≡" to associate any vector to its representation in some fixed coordinate system.

resulting vectors in $\mathcal{V}$ after applying the transformation T over the basis vectors for $\mathcal{U}$. Notice that $[\vec{u}^{(j)}]_\beta = (e^{(j)})_\beta$ where the j-th basis vector of β is explicitly represented in the 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 β 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 $\vec{w}^{(j)} = a_1^{(j)}\vec{v}^{(1)} + a_2^{(j)}\vec{v}^{(2)} + \cdots + a_m^{(j)}\vec{v}^{(m)} = \sum_{i=1}^m a_i^{(j)}\vec{v}^{(i)}$ of the basis vectors $\vec{v}^{(i)}$ from η, i.e.

$$T(\vec{u}^{(j)}) = \sum_{i=1}^{m} a_i^{(j)}\vec{v}^{(i)} \tag{7.1}$$

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 $T(\vec{u}) \equiv A[\vec{u}]_\beta$. Subsequently,

$$\begin{aligned}
A[\vec{u}^{(j)}]_\beta &= a_i^{(j)}(e^{(j)})_\beta \\
&= \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(j)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(j)} & & a_2^{(n)} \\ \vdots & & & \vdots & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(j)} & \cdots & a_m^{(n)} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \text{ (the } j\text{-th entry)} \\ \vdots \\ 0 \text{ (the last index is } n) \end{bmatrix} \\
&= \begin{bmatrix} a_1^{(j)} \\ a_2^{(j)} \\ \vdots \\ a_m^{(j)} \end{bmatrix}
\end{aligned}$$

Due to the structure of $(e^{(j)})_\beta$, 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,

[2] $T(\vec{u}^{(1)} + \vec{u}^{(2)}) \equiv A([\vec{u}^{(1)}]_\beta + [\vec{u}^{(2)}]_\beta) = A[\vec{u}^{(1)}]_\beta + A[\vec{u}^{(2)}]_\beta \equiv T(\vec{u}^{(1)}) + T(\vec{u}^{(2)})$ and $T(a\vec{u}^{(1)}) \equiv A(a[\vec{u}^{(1)}]_\beta) = a(A[\vec{u}^{(1)}]_\beta) \equiv aT(\vec{u}^{(1)})$

the coordinates of $\vec{w}^{(j)} = T(\vec{u}^{(j)})$ in the η system

$$
\begin{aligned}
T(\vec{u}^{(j)}) \equiv [T(\vec{u}^{(j)})]_\eta &= [\vec{w}^{(j)}]_\eta \\
&= [\sum_{i=1}^{m} a_i^{(j)} \vec{v}^{(i)}]_\eta \\
&= \sum_{i=1}^{m} a_i^{(j)} [\vec{v}^{(i)}]_\eta \\
&= \sum_{i=1}^{m} a_i^{(j)} (e^{(i)})_\eta \\
&= a_1^{(j)} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2^{(j)} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + a_m^{(j)} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \text{ (the last index is } m) \end{bmatrix} \\
&= \begin{bmatrix} a_1^{(j)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_2^{(j)} \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ a_m^{(j)} \end{bmatrix} = \begin{bmatrix} a_1^{(j)} \\ a_2^{(j)} \\ \vdots \\ a_m^{(j)} \end{bmatrix}
\end{aligned}
$$

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 $[A]_\beta^\eta = a_i^{(j)}$ to the linear transformation T is consistent, where we have now added the subscript β and superscript η to emphasize the transformation is 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 β basis and find its coordinates in the η frame, namely $[\vec{w}^{(j)}]_\eta$, which readily become the j-th column of the matrix to be found. To be more clear, we have

$$
\begin{aligned}
[T]_\beta^\eta &= \left[[T(\vec{u}^{(1)})]_\eta | [T(\vec{u}^{(2)})]_\eta | \cdots | [T(\vec{u}^{(n)})]_\eta \right] \\
&= \left[[\vec{w}^{(1)}]_\eta | [\vec{w}^{(2)}]_\eta | \cdots | [\vec{w}^{(n)}]_\eta \right]
\end{aligned}
$$

$$= \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(n)} \\ \vdots & & \ddots & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(n)} \end{bmatrix} = a_i^{(j)} = [A]_\beta^\eta \tag{7.2}$$

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 $T : \mathcal{U} \to \mathcal{V}$ as defined in Definition 7.1.1 between two finite-dimensional vector spaces, with respect to the bases $\beta = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n)}\}$ for $\mathcal{U}$ and $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(m)}\}$ for $\mathcal{V}$, has a unique matrix representation of

$$[T]_\beta^\eta = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(n)} \\ \vdots & & \ddots & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(n)} \end{bmatrix}$$

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, $[T]_\beta^\eta [\vec{u}]_\beta = [\vec{v}]_\eta$.

Properties 7.1.3. For two linear transformations S and T, if $S(\vec{v}) = T(\vec{v})$ for all $\vec{v} \in \mathcal{V}$, then $S = T$ must be the same linear transformation.

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 $\beta = \{(1, 2, 1)^T, (0, 1, -1)^T, (2, -1, 1)^T\}$ is a basis for $\mathcal{U}$, and the same goes for $\mathcal{V}$ with a basis $\eta = \{(1, 2)^T, (2, -1)^T\}$. If a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^2$ 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 $[T]_\beta^\eta$ with respect to the β and η 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 be applied on the basis vectors in β. 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 η:

$$(5, 0)^T = 1(1, 2)^T + 2(2, -1)^T$$

Hence $a_1^{(1)} = 1$, $a_2^{(1)} = 2$, and this gives us the first column of $[T]_\beta^\eta$ 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 β, 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

$$[T]_\beta^\eta = \begin{bmatrix} 1 & -\frac{2}{5} & \frac{8}{5} \\ 2 & \frac{6}{5} & -\frac{4}{5} \end{bmatrix}$$

For the second part, we start by expressing $(-1,4,1)^T$ in the basis β. As $(-1,4,-1)^T = 1(1,2,1)^T + 1(0,1,-1)^T - 1(2,-1,1)^T$, we have $(-1,4,1)^T = (1,1,-1)^T_\beta$, and then

$$\begin{aligned}
[T((1,1,-1)^T_\beta)]_\eta &= [T]^\eta_\beta (1,1,-1)^T_\beta \\
&= \begin{bmatrix} 1 & -\frac{2}{5} & \frac{8}{5} \\ 2 & \frac{6}{5} & -\frac{4}{5} \end{bmatrix}^\eta_\beta \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}_\beta \\
&= \begin{bmatrix} -1 \\ 4 \end{bmatrix}_\eta = (-1,4)^T_\eta
\end{aligned}$$

implying that $T((-1,4,-1)^T) = (-1,4)^T_\eta = -1(1,2)^T + 4(2,-1)^T = (7,-6)^T$ 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. □

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.[3] Two prime examples of abstract vector spaces are the set of (real) polynomials $\mathcal{P}^n$ with a degree up to n[4] 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.

[3] Again, this appeals to the isomorphic nature of same-dimensional real vector spaces, which will be discussed soon.

[4] We shall argue for some criteria in Definition 6.1.1 for $\mathcal{P}^n$ here. For instance, 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.

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 $\beta = \{1, x, x^2\}$ and $\eta = \{1, x\}$. (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 $\{1, x, x^2, \cdots, x^{n-1}, x^n\}$ and thus $(n+1)$-dimensional. Readers are advised to justify why they constitute a basis for the polynomial spaces.) Let $T : \mathcal{U} \to \mathcal{V}$ be $T[p(x)] = p'(x)$ the differentiation operator and find its matrix representation with respect to β and η.

Solution. We essentially do the same thing as in Example 7.1.1, but it is 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

$$[T]_\beta^\eta = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Notice that we can express, quite trivially

$$\begin{aligned} 1 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_\beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}_\eta \\ x &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}_\beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}_\eta \\ x^2 &= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}_\beta \end{aligned}$$

using vector notation with respect to the two given standard bases. We can verify the form of $[T]_\beta^\eta$ by a test polynomial $c_0 + c_1x + c_2x^2$, whose vector representation in β is clearly $(c_0, c_1, c_2)_\beta^T$. Then, multiplying $[T]_\beta^\eta$ to its left gives

$$[T((c_0, c_1, c_2)_\beta^T)]_\eta = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}_\beta^\eta \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}_\beta = \begin{bmatrix} c_1 \\ 2c_2 \end{bmatrix}_\eta$$

which corresponds to the polynomial $c_1 + 2c_2x$. This coincides with the usual result of differentiation, that is, $\frac{d}{dx}(c_0 + c_1x + c_2x^2) = c_1 + 2c_2x$. □

In each of the previous examples, we consider a linear transformation between two vector spaces that are of the same type (the usual real n-space vectors/polynomials). Shown below is what happens 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 $\beta = \{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T\}$ and $\eta = \{1, x, x^2\}$ 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 β and η.

Solution. Again, we repeat what we have done in the previous two examples. It is apparent that

$$\begin{aligned} T((1,0,0)_\beta^T) &= (1) + (0)x + (0)x^2 = 1 = (1,0,0)_\eta^T \\ T((0,1,0)_\beta^T) &= (0) + (1)x + (0)x^2 = x = (0,1,0)_\eta^T \\ T((0,0,1)_\beta^T) &= (0) + (0)x + (1)x^2 = x^2 = (0,0,1)_\eta^T \end{aligned}$$

So by Definition 7.1.2, the desired matrix representation is simply

$$[T]_\beta^\eta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which is the 3×3 identity matrix. This is expected as the linear transformation is essentially $T[(c_0, c_1, c_2)_\beta^T] = (c_0, c_1, c_2)_\eta^T$ where $(c_0, c_1, c_2)_\eta^T = c_0 + c_1 x + c_2 x^2$, which means that the numeric coordinate representation of vectors in the two vector spaces is preserved under such a linear transformation between them and the only visible change is the subscript. □

Most of the readers should find it boring in the above example as we are just stating the obvious. It is a straightforward, "one-to-one" association between the standard bases of the real n-space and the 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 with a fixed 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.

Definition 7.1.4 (Injective Transformation)**.** A transformation $T : \mathcal{U} \to \mathcal{V}$ 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?)[5] 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.

Definition 7.1.5 (Surjective Transformation)**.** A transformation $T : \mathcal{U} \to \mathcal{V}$ 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 $\vec{v} = d_0 + d_1 x \in \mathcal{P}^1$, all preimages in the form of $\vec{u} = K + d_0 x + \frac{d_1}{2}x^2 \in \mathcal{P}^2$, 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 $\vec{v} = d_0 + d_1 x$ with specific d_0 and d_1, and note that both $\vec{u}^{(1)} = K_1 + d_0 x + \frac{d_1}{2}x^2$ and $\vec{u}^{(2)} = K_2 + d_0 x + \frac{d_1}{2}x^2$ where K_1, K_2 are distinct satisfy $T(\vec{u}^{(1)}) = T(\vec{u}^{(2)}) = \vec{v}$, but $\vec{u}^{(1)} \neq \vec{u}^{(2)}$.

However, in other cases, it may not be 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

[5] $T(\mathbf{0}) = T(0\vec{u}) = 0T(\vec{u}) = \mathbf{0}$ for arbitrary $\vec{u}$ due to the homogeneity property as required in Definition 7.1.1.

it requires the transformation to be linear (and here is where the linearity comes to play).

Theorem 7.1.6. A linear transformation $T : \mathcal{U} \to \mathcal{V}$ (between two finite-dimensional vector spaces) is one-to-one if and only if given any basis $\beta = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n)}\}$ for $\mathcal{U}$, $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \ldots, T(\vec{u}^{(n)}) \in \mathcal{V}$ are linearly independent.

Theorem 7.1.7. A linear transformation $T : \mathcal{U} \to \mathcal{V}$ (between two finite-dimensional vector spaces) is onto if and only if given any basis $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(m)}\}$ 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 7.1.6: The "if" direction is proved by showing $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \ldots, T(\vec{u}^{(n)})$ are linearly independent implies that, if $T(\vec{u}) = \mathbf{0}$ then $\vec{u} = \mathbf{0}$ as suggested by the alternative condition in Definition 7.1.4. By Theorem 6.1.9, the equation $c_1T(\vec{u}^{(1)}) + c_2T(\vec{u}^{(2)}) + \cdots + c_nT(\vec{u}^{(n)}) = \mathbf{0}$ will only have $c_j = \mathbf{0}$ as the trivial solution. Now by linearity from Definition 7.1.1, we have

$$\begin{aligned} c_1T(\vec{u}^{(1)}) + c_2T(\vec{u}^{(2)}) + \cdots + c_nT(\vec{u}^{(n)}) &= T(c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \cdots + c_n\vec{u}^{(n)}) \\ &= \mathbf{0} \end{aligned}$$

Since $c_j = 0$ is the only possibility, this means that if $T(c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \cdots + c_n\vec{u}^{(n)}) = \mathbf{0}$ then $\vec{u} = c_1\vec{u}^{(1)} + c_2\vec{u}^{(2)} + \cdots + c_n\vec{u}^{(n)}$ must be $\mathbf{0}$, hence $T(\vec{u}) = \mathbf{0}$ implies $\vec{u} = 0$ and we are done. The converse is similarly proved with the argument going in a reverse direction.

Theorem 7.1.7: We compare Theorem 7.1.7 against Definition 7.1.5 to show the part of the "if" direction. Since $\eta = \{\vec{v}^{(i)}\}$, $i = 1, 2, \ldots, m$, is a basis for $\mathcal{V}$, any $\vec{v} \in \mathcal{V}$ can be written as a linear combination of $\vec{v} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + \cdots + c_m\vec{v}^{(m)}$. 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

$$\vec{v} = c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} + \cdots + c_m\vec{v}^{(m)}$$

$$= c_1 T(\vec{w}^{(1)}) + c_2 T(\vec{w}^{(2)}) + \cdots + c_m T(\vec{w}^{(m)})$$
$$= T(c_1 \vec{w}^{(1)} + c_2 \vec{w}^{(2)} + \cdots + c_m \vec{w}^{(m)})$$

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 Definition 7.1.5 for all possible i. □

Example 7.1.4. Given a linear transformation $T : \mathcal{U} \to \mathcal{V}$ where $\mathcal{U}$ and $\mathcal{V}$ have a dimension of 3 and 4 respectively, if its matrix representation corresponding to some bases β and η is

$$[T]_\beta^\eta = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

determine whether it is (a) one-to-one, as well as (b) onto, or not.

Solution. (a) By Theorem 7.1.6, 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 β. Their coordinate representation in the β system is trivially $[\vec{u}^{(1)}]_\beta = (e^{(1)})_\beta = (1, 0, 0)_\beta^T$, $[\vec{u}^{(2)}]_\beta = (e^{(2)})_\beta = (0, 1, 0)_\beta^T$ and $[\vec{u}^{(3)}]_\beta = (e^{(3)})_\beta = (0, 0, 1)_\beta^T$, and hence

$$\begin{aligned} [T(\vec{u}^{(1)})]_\eta &= [T]_\beta^\eta (e^{(1)})_\beta \\ &= \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}_\beta^\eta \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_\beta = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}_\eta \end{aligned}$$

which is just the first column of $[T]^{\eta}_{\beta}$. Similarly, $[T(\vec{u}^{(2)})]_{\eta} = (-1, 1, 0, 1)^T_{\eta}$, $[T(\vec{u}^{(3)})]_{\eta} = (0, 1, -1, 0)^T_{\eta}$ are then the second/third column of $[T]^{\eta}_{\beta}$. From this, we see that in general, the coordinates in η after transformation $[T(\vec{u}^{(j)})]_{\eta}$ is just the j-th column of $[T]^{\eta}_{\beta}$. (Actually, this has been observed when we are deriving the matrix representation of linear transformations at the beginning of this chapter.) So the problem is reduced to deciding whether the column vectors constituting $[T]^{\eta}_{\beta}$ are linearly independent or not. By Theorem 6.1.9, we can accomplish this by showing if the solution $[T]^{\eta}_{\beta}[\vec{u}]_{\beta} = \mathbf{0}$ consists of the trivial solution only, and we have

$$\left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 2 & 0 & 0 \end{array}\right] \qquad \begin{array}{l} R_3 - R_1 \to R_3 \\ R_4 - R_1 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & -2 & 0 \end{array}\right] \qquad \begin{array}{l} R_3 - R_2 \to R_3 \\ R_4 - R_2 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 \end{array}\right] \qquad -\frac{1}{2}R_3 \to R_3$$

$$\to \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_4 + 2R_3 \to R_4$$

As every column in this homogeneous system contains a pivot, it demonstrates that $[T]^{\eta}_{\beta}[\vec{u}]_{\beta} = \mathbf{0}$ indeed only has the trivial solution $[\vec{u}]_{\beta} = \mathbf{0}$, and therefore the linear transformation in question is one-to-one.

(b) By Definition 7.1.5, it is equivalent to showing that if the $\{T(\vec{u}^{(j)})\}$ span $\mathcal{W}$, or expressed in terms of the β/η coordinates, whether the three transformed vectors $\{[T]_\beta^\eta[\vec{u}^{(1)}]_\beta, [T]_\beta^\eta[\vec{u}^{(2)}]_\beta, [T]_\beta^\eta[\vec{u}^{(3)}]_\beta\}$ 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. □

Accompanying injectivity and surjectivity are the ideas of ***kernel*** and ***range***. For a (linear) transformation $T : \mathcal{U} \to \mathcal{V}$, its kernel consists of vectors in $\mathcal{U}$ that are 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 7.1.8. For a (linear) transformation $T : \mathcal{U} \to \mathcal{V}$, its kernel is defined to be

$$\mathrm{Ker}(T) = \{\vec{u} \in \mathcal{U} | T(\vec{u}) = \mathbf{0}_V\} \tag{7.3}$$

whereas its range is

$$\mathcal{R}(T) = \{\vec{v} \in \mathcal{V} | T(\vec{u}) = \vec{v} \text{ for some } \vec{u} \in \mathcal{U}\} \tag{7.4}$$

Also, notice that the kernel and range are a subspace of $\mathcal{U}$ and $\mathcal{V}$ respectively.[6] 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

[6]For $\vec{u}^{(1)}, \vec{u}^{(2)} \in \mathrm{Ker}(T) \subseteq \mathcal{U}$, $T(a\vec{u}^{(1)} + b\vec{u}^{(2)}) = aT(\vec{u}^{(1)}) + bT(\vec{u}^{(2)}) = a\mathbf{0}_V + b\mathbf{0}_V = \mathbf{0}_V$ for any scalar a and b so $a\vec{u}^{(1)} + b\vec{u}^{(2)} \in \mathrm{Ker}(T)$ and by Theorem 6.1.2 it is a subspace of $\mathcal{U}$. We leave showing that the range is a subspace of $\mathcal{V}$ as an exercise to the readers.

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 has 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 $T : \mathcal{U} \to \mathcal{V}$ where $\mathcal{U}$ and $\mathcal{V}$ are now both having a dimension of 3, if its matrix representation corresponding to some bases β and η is

$$[T]_\beta^\eta = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

find its kernel and range.

Solution. According to Definition 7.1.8, $\text{Ker}(T)$ is the set of $\vec{u}$ that satisfies $T(\vec{u}) = \mathbf{0}$, or using basis representation (Definition 7.1.2), $[T]_\beta^\eta [\vec{u}]_\beta = \mathbf{0}$. Therefore, it is equivalent to finding the null space (Definition 6.3.8) of $[T]_\beta^\eta$:

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right] \qquad \begin{array}{c} R_2 - R_1 \to R_2 \\ R_3 - R_1 \to R_3 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{array}\right] \qquad R_2 \leftrightarrow R_3$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_3 + R_2 \to R_3$$

The nullity is 1 and we can let $[u_3]_\beta = t$ be the free variable, and we have $[u_1]_\beta = -t$ and $[u_2]_\beta = 0$ from the first two rows. So the kernel takes the form

of

$$\mathrm{Ker}(T) = \begin{bmatrix} -t \\ 0 \\ t \end{bmatrix}_\beta = t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}_\beta$$

where $-\infty < t < \infty$, or in other words, $\mathrm{Ker}(T) = \mathrm{span}(\{(-1, 0, 1)_\beta^T\})$ with a dimension of 1. Similarly, the range of T will be the column space of $[T]_\beta^\eta$. 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 $\mathcal{R}(T) = \mathrm{span}(\{(1, 1, 1)_\beta^T, (0, -1, 1)_\beta^T\})$ and has a dimension of 2 (the rank of the $[T]_\beta^\eta$ matrix). □

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. Hence, the Rank-nullity Theorem 6.3.9 can be extended to any linear transformation in general as

Theorem 7.1.9 (Rank-nullity Theorem for Linear Transformations)**.** For a linear transformation $T : \mathcal{U} \to \mathcal{V}$, we have

$$\dim(\mathrm{Ker}(T)) + \dim(\mathcal{R}(T)) = \dim(\mathcal{U}) \tag{7.5}$$

Finally, we can rewrite Definitions 7.1.4 and 7.1.5 using the notion of kernel and range (alongside Properties 6.2.5).

Properties 7.1.10. A linear transformation $T : \mathcal{U} \to \mathcal{V}$ is one-to-one if and only if the dimension of its kernel $\mathrm{Ker}(T)$ is zero, i.e. $\mathrm{Dim}(\mathrm{Ker}(T)) = 0$. Meanwhile, it is onto if and only if the dimension of range $\mathcal{R}(T)$ (rank) is the 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 7.1.11 (Composition of Linear Transformations)**.** For two (linear) transformations $T : \mathcal{U} \to \mathcal{V}$ and $S : \mathcal{V} \to \mathcal{W}$, their composition is defined as $S \circ T : \mathcal{U} \to \mathcal{W}$ 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

$$
\begin{aligned}
(S \circ T)(\vec{u}^{(j)}) = S(T(\vec{u}^{(j)})) &= S\left(\sum_{k=1}^{r} a_k^{(j)} \vec{v}^{(k)}\right) \\
&= \sum_{k=1}^{r} a_k^{(j)} S(\vec{v}^{(k)}) \qquad \text{(Linearity)}
\end{aligned}
$$

$$= \sum_{k=1}^{r} a_k^{(j)} \left(\sum_{i=1}^{m} b_i^{(k)} \vec{w}^{(i)} \right)$$
$$= \sum_{i=1}^{m} \left(\left(\sum_{k=1}^{r} b_i^{(k)} a_k^{(j)} \right) \vec{w}^{(i)} \right)$$

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 $T : \mathcal{U} \to \mathcal{V}$ and $S : \mathcal{V} \to \mathcal{U}$ 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. $(S \circ T)(\vec{u}) = \mathrm{id}_U(\vec{u}) = \vec{u} \in \mathcal{U}$ and $(T \circ S)(\vec{v}) = \mathrm{id}_V(\vec{v}) = \vec{v} \in \mathcal{V}$, where $\mathrm{id} : \mathcal{U} \to \mathcal{U}$ (or $\mathcal{V} \to \mathcal{V}$) denotes the ***identity transformation/identity mapping*** that simply returns the input vector as the output.

Definition 7.1.12. For a (linear) transformation $T : \mathcal{U} \to \mathcal{V}$, if there is another linear transformation $S : \mathcal{V} \to \mathcal{U}$ such that $S \circ T = \mathrm{id}_U$ and $T \circ S = \mathrm{id}_V$, 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 β and η. Then both the linear mappings $T \equiv [T]_\beta^\eta$ and $S \equiv [S]_\eta^\beta$[7] have a corresponding matrix representation with a shape of $m \times n$ and $n \times m$. It is also not hard to accept that $\mathrm{id}_U \equiv [\mathrm{id}_U]_\beta^\beta = I_n$ and $\mathrm{id}_V \equiv [\mathrm{id}_V]_\eta^\eta = I_m$. Subsequently, for S to be the inverse of T, $S \circ T = \mathrm{id}_U \equiv I_n = [S]_\eta^\beta [T]_\beta^\eta$, and similarly we need $[T]_\beta^\eta [S]_\eta^\beta = I_m$. We claim that $m = n$[8], and thus invertible linear transformations

[7] "$\equiv$" will also be used to associate any linear transformation to its matrix representation.

[8] Assume without loss of generality, $m < n$, then by Properties 6.3.14, $\mathrm{rank}([S]_\eta^\beta [T]_\beta^\eta) \leq \min(\mathrm{rank}([S]_\eta^\beta), \mathrm{rank}([T]_\beta^\eta)) \leq m < n$, but $\mathrm{rank}(I_n) = n$ so it is impossible that $[S]_\eta^\beta [T]_\beta^\eta = I_n$.

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 7.1.13. For an invertible linear transformation $T : \mathcal{U} \to \mathcal{V}$ 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 $T^{-1} = S : \mathcal{V} \to \mathcal{U}$, then T^{-1} will be unique. The matrix representation of this inverse would be $[T^{-1}]_\eta^\beta = ([T]_\beta^\eta)^{-1}$. Moreover, $(T^{-1})^{-1} = T$ and $(X \circ T)^{-1} = T^{-1} \circ X^{-1}$ if $X : \mathcal{V} \to \mathcal{W}$ 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 Definitions 7.1.4 and 7.1.5, for a linear transformation $T : \mathcal{U} \to \mathcal{V}$ 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 say a transformation is bijective *between* two same-dimensional vector spaces. There are two major results regarding invertibility here. The first one is

Theorem 7.1.14. There always exists a bijective linear mapping between $\mathcal{V}$ itself, i.e. $T : \mathcal{V} \to \mathcal{V}$, that transforms the coordinates of any fixed vector in $\mathcal{V}$ between two different bases (denote them by β and β') of its. Such a change of

coordinates in $\mathcal{V}$ has a matrix representation $[T]_\beta^{\beta'} = P_\beta^{\beta'}$ that is invertible.

Proof. Since it is the same vector space $\mathcal{V}$ but just represented in different bases, the number of dimensions will stay the same, let's say n, and the bases β and β' both are made up of n basis vectors (Properties 6.2.4). Denote them by $\beta = \{\vec{v}_\beta^{(1)}, \vec{v}_\beta^{(2)}, \ldots, \vec{v}_\beta^{(n)}\}$ and $\beta' = \{\vec{v}_{\beta'}^{(1)}, \vec{v}_{\beta'}^{(2)}, \ldots, \vec{v}_{\beta'}^{(n)}\}$. The desired mapping $T : \mathcal{V} \to \mathcal{V}$ is in fact

$$T(\vec{v}) = \mathrm{id}(\vec{v}) = \vec{v} \tag{7.6}$$

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 onto itself, and is described by

$$[\vec{v}]_{\beta'} = [T]_\beta^{\beta'} [\vec{v}]_\beta \tag{7.7}$$

following Definition 7.1.2 with $\mathcal{U} = \mathcal{V}$ and $\vec{u} = \vec{v}$. Now note that $[\vec{v}]_{\beta'} = [T]_\beta^{\beta'} [\vec{v}]_\beta$ has a unique solution $[\vec{v}]_\beta$ for any $[\vec{v}]_{\beta'}$ 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 $[\vec{v}]_{\beta'}$ is associated to one and only one $[\vec{v}]_\beta$. Part (d) to (a) of Theorem 3.2.1 then shows that $[T]_\beta^{\beta'}$ is an invertible matrix. According to the discussion prior to Definition 7.1.2, $[T]_\beta^{\beta'}$ takes the form of

$$\begin{aligned} P_\beta^{\beta'} = [T]_\beta^{\beta'} &= \left[[\mathrm{id}(\vec{v}_\beta^{(1)})]_{\beta'} | [\mathrm{id}(\vec{v}_\beta^{(2)})]_{\beta'} | \cdots | [\mathrm{id}(\vec{v}_\beta^{(n)})]_{\beta'} \right] \\ &= \left[[\vec{v}_\beta^{(1)}]_{\beta'} | [\vec{v}_\beta^{(2)}]_{\beta'} | \cdots | [\vec{v}_\beta^{(n)}]_{\beta'} \right] \end{aligned} \tag{7.8}$$

So we have to find how each of the basis vectors in β is expressed in the β' system. Conversely,

$$P_{\beta'}^{\beta} = ([T]_\beta^{\beta'})^{-1} = [T]_{\beta'}^{\beta} = \left[[\vec{v}_{\beta'}^{(1)}]_\beta | [\vec{v}_{\beta'}^{(2)}]_\beta | \cdots | [\vec{v}_{\beta'}^{(n)}]_\beta \right] \tag{7.9}$$

□

Be aware that despite "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 $T : \mathcal{U} \to \mathcal{V}$ 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 justifies our fixing of bases during several arguments in the previous subsections.

Example 7.1.6. Show that $\beta = \{(1,0,1)^T, (0,2,1)^T, (-1,1,2)^T\}$ and $\beta' = \{(0,0,1)^T, (2,0,1)^T, (1,-1,0)^T\}$ 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

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

are non-zero or not. A simple computation shows that $\det(B) = 5$ and $\det(B') = -2$ and thus both β and β' are bases for $\mathbb{R}^3$. By Theorem 7.1.14, the matrix representation for the change of basis abides

$$[\mathrm{id}]_\beta^{\beta'} = [T]_\beta^{\beta'} = \left[[\vec{v}_\beta^{(1)}]_{\beta'} | [\vec{v}_\beta^{(2)}]_{\beta'} | \cdots | [\vec{v}_\beta^{(n)}]_{\beta'} \right]$$

where each of $[\vec{v}_\beta^{(j)}]_{\beta'}$ is found via the equation

$$[(\vec{v}_\beta^{(j)})_1]_{\beta'}(\vec{v}_{\beta'}^{(1)}) + [(\vec{v}_\beta^{(j)})_2]_{\beta'}(\vec{v}_{\beta'}^{(2)}) + [(\vec{v}_\beta^{(j)})_3]_{\beta'}(\vec{v}_{\beta'}^{(3)}) = \vec{v}_\beta^{(j)}$$

just as in Example 6.2.1 with $[(\vec{v}_\beta^{(j)})_i]_{\beta'}$ being the i-th component of $\vec{v}_\beta^{(j)}$ in the β' frame, or equivalently,

$$\left[\vec{v}_{\beta'}^{(1)}|\vec{v}_{\beta'}^{(2)}|\vec{v}_{\beta'}^{(3)}\right]\begin{bmatrix}[(\vec{v}_\beta^{(j)})_1]_{\beta'}\\ [(\vec{v}_\beta^{(j)})_2]_{\beta'}\\ [(\vec{v}_\beta^{(j)})_3]_{\beta'}\end{bmatrix} = \vec{v}_\beta^{(j)}$$

$$\begin{aligned}[\vec{v}_\beta^{(j)}]_{\beta'} = \begin{bmatrix}[(\vec{v}_\beta^{(j)})_1]_{\beta'}\\ [(\vec{v}_\beta^{(j)})_2]_{\beta'}\\ [(\vec{v}_\beta^{(j)})_3]_{\beta'}\end{bmatrix} &= \left[\vec{v}_{\beta'}^{(1)}|\vec{v}_{\beta'}^{(2)}|\vec{v}_{\beta'}^{(3)}\right]^{-1}\vec{v}_\beta^{(j)}\\ &= B'^{-1}\vec{v}_\beta^{(j)}\end{aligned}$$

Subsequently,

$$\begin{aligned}[T]_\beta^{\beta'} &= \left[[\vec{v}_\beta^{(1)}]_{\beta'}|[\vec{v}_\beta^{(2)}]_{\beta'}|\cdots|[\vec{v}_\beta^{(n)}]_{\beta'}\right]\\ &= \left[B'^{-1}\vec{v}_\beta^{(1)}|B'^{-1}\vec{v}_\beta^{(2)}|B'^{-1}\vec{v}_\beta^{(3)}\right]\\ &= B'^{-1}\left[\vec{v}_\beta^{(1)}|\vec{v}_\beta^{(2)}|\vec{v}_\beta^{(3)}\right]\\ &= B'^{-1}B\end{aligned}$$

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 9 of Chapter 6), and the required matrix representation for the coordinate change is

$$P_\beta^{\beta'} = [T]_\beta^{\beta'} = B'^{-1}B = \begin{bmatrix}0 & 2 & 1\\ 0 & 0 & -1\\ 1 & 1 & 0\end{bmatrix}^{-1}\begin{bmatrix}1 & 0 & -1\\ 0 & 2 & 1\\ 1 & 1 & 2\end{bmatrix}$$

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

Let's take $(2,2,3)^T = 2(1,0,1)^T + 1(0,2,1)^T + 0(-1,1,2)^T = (2,1,0)^T_\beta$ for double-checking:

$$P^{\beta'}_{\beta}(2,1,0)^T_\beta = \begin{bmatrix} \frac{1}{2} & 0 & 2 \\ \frac{1}{2} & 1 & 0 \\ 0 & -2 & -1 \end{bmatrix}^{\beta'}_{\beta} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}_\beta = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}_{\beta'}$$

and indeed $(2,2,3)^T = 1(0,0,1)^T + 2(2,0,1)^T + (-2)(1,-1,0)^T = (1,2,-2)^T_{\beta'}$. □

More generally, the relation

$$P^{\beta'}_{\beta} = [\mathrm{id}]^{\beta'}_{\beta} = B'^{-1}B \tag{7.10}$$

still remains valid where B and B' are matrices composed by the basis vectors from the β and β' systems, relative to a third basis (without loss of generality we have assumed it is the standard basis $\mathcal{S}$[9], but the readers are advised to extend this for any other arbitrary basis), where the basis vectors are arranged in columns. To see this from another perspective, take any vector $\vec{v}$ that is expressed in the β coordinates, $[\vec{v}]_\beta$. We can view the change in coordinates from β to β' in two steps: first from β to $\mathcal{S}$, and then from $\mathcal{S}$ to β'. From (6.6) in Section 6.2.1, we already know that the former constitutes $[\vec{v}]_S = B[\vec{v}]_\beta$, and the latter is done by $[\vec{v}]_{\beta'} = B'^{-1}[\vec{v}]_S$. Combining these two operations together we have $[\vec{v}]_{\beta'} = B'^{-1}[\vec{v}]_S = B'^{-1}B[\vec{v}]_\beta$ and hence $[\mathrm{id}]^{\beta'}_{\beta} = B'^{-1}B$.

The second major result of this subsection is

[9]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.

Theorem 7.1.15. 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. Denote the basis chosen for $\mathcal{V}$ by $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}\}$ and we use the standard basis $\mathcal{S} = \{\hat{e}^{(1)}, \hat{e}^{(2)}, \ldots, \hat{e}^{(n)}\}$ for $\mathbb{R}^n$. Then the linear mapping $T : \mathcal{V} \to \mathbb{R}^n$ that abides

$$T(\vec{v}^{(j)}) = \hat{e}^{(j)} \tag{7.11}$$

where $j = 1, 2, \ldots, n$, is bijective as desired. To see this, by Theorem 7.1.6, as for every $\vec{v}^{(j)}$, $T(\vec{v}^{(j)}) = \hat{e}^{(j)}$ leads to the standard unit vectors that are linearly independent, thus T is one-to-one. Meanwhile, a direct use of Theorem 7.1.7 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 $\vec{x} \in \mathbb{R}^n$, $[\vec{x}]_S = [T]_\beta^S [\vec{v}]_\beta$ has a unique solution $[\vec{v}]_\beta$, and part (d) to (a) of Theorem 6.1.10 then shows that the matrix representation $[T]_\beta^S$ is invertible where $T^{-1} \equiv ([T]_\beta^S)^{-1}$. The converse follows the same argument running in an opposite direction. □

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 7.1.16. 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 dimensions. 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 7.1.15 and bijectivity is transitive.[10] For the "only if" direction, let 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 $T\colon \mathcal{U} \to \mathcal{V}$ the m transformed basis vectors $T(\vec{u}^{(1)}), T(\vec{u}^{(2)}), \ldots, T(\vec{u}^{(m)})$ which will be unable to span the n-dimensional $\mathcal{V}$ and by Definition 7.1.5 all of them are not surjective.

Example 7.1.7. Explicitly show that $\mathcal{U} = \mathcal{P}^3$ and $\mathcal{V} = \text{span}(\eta)$, where $\eta = \{e^x, xe^x, x^2e^x, x^3e^x\}$, are isomorphic by considering $T : \mathcal{U} \to \mathcal{V}$, $T[p(x)] = \int_{-\infty}^{x} e^x p(x)dx$.

Solution. It is clear that both $\mathcal{U}$ and $\mathcal{V}$ are four-dimensional and by the above corollary, they are isomorphic. Take $\beta = \{1, x, x^2, x^3\}$ the standard polynomial basis for $\mathcal{U} = \mathcal{P}^3$ and the linearly independent η is automatically the basis for $\mathcal{V}$. Now we compute the matrix representation $[T]_\beta^\eta$ as follows. By elementary Calculus,

$$T(1) = \int_{-\infty}^{x} e^x dx = e^x$$
$$T(x) = \int_{-\infty}^{x} xe^x dx = xe^x - e^x$$

[10] Let's say $S\colon \mathcal{U} \to \mathbb{R}^n$ and $T\colon \mathcal{V} \to \mathbb{R}^n$ are the two respective isomorphisms. Then $T^{-1} \circ S\colon \mathcal{U} \to \mathcal{V}$ will be the required bijective linear transformation.

$$T(x^2) = \int_{-\infty}^{x} x^2 e^x dx = x^2 e^x - 2xe^x + 2e^x$$
$$T(x^3) = \int_{-\infty}^{x} x^3 e^x dx = x^3 e^x - 3x^2 e^x + 6xe^x - 6e^x$$

and thus

$$[T]_\beta^\eta = \begin{bmatrix} 1 & -1 & 2 & -6 \\ 0 & 1 & -2 & 6 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is an upper-triangular matrix and its determinant is simply the product of diagonal entries $(1)^4 = 1 \neq 0$. Therefore, by Theorem 6.1.10, $[T]_\beta^\eta$ is invertible, and the given transformation, as well as $\mathcal{U}$ and $\mathcal{V}$ themselves, is/are isomorphic according to Theorem 7.1.15. □

► Short Exercise: Redo the above example by considering $T[p(x)] = e^x p(x)$ this time.[11]

7.2. Additional Discussions about Coordinate Bases

7.2.1. Linear Change of Coordinates

In previous parts, we have already mentioned about the change of coordinates between bases several times, where such a mapping is 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 for change of coordinates for vectors has been discussed substantially in Examples 6.2.1, 7.1.6 and explained through Theorem 7.1.14. 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 xy plane. Given a basis β 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 β.

Solution. As before, $P^S_\beta = [\vec{v}^{(1)}|\vec{v}^{(2)}]$, and it can be seen that

$$P^S_\beta = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \qquad P^\beta_S = (P^S_\beta)^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}$$

Hence the coordinates of $\vec{v}$ in the β system according to (6.6), is

$$[\vec{v}]_\beta = P^\beta_S [\vec{v}]_S = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}^\beta_S \begin{bmatrix} 2 \\ 1 \end{bmatrix}_S = \begin{bmatrix} 1 \\ 1 \end{bmatrix}_\beta$$

The geometry of this problem is shown in Figure 7.1 on the next page where each grid line separation represents one unit length of the axis vectors. □

In this example, we can see that in the two bases $\mathcal{S}$ and β, their axis vectors (reversed in the figure) can be converted via $T : \mathbb{R}^2 \to \mathbb{R}^2$ with $T(\hat{e}^{(j)}) = \vec{v}^{(j)}$. The corresponding matrix representation is $\vec{v}^{(j)} = [\vec{v}^{(1)}|\vec{v}^{(2)}]\hat{e}^{(j)} = P^S_\beta \hat{e}^{(j)}$. Meanwhile, the coordinate transformation of a fixed vector follows $[\vec{v}]_\beta = (P^S_\beta)^{-1}[\vec{v}]_S = P^\beta_S [\vec{v}]_S$, 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

[11] It becomes trivial and the matrix representation is simply the identity matrix.

Figure 7.1.: *The geometry of Example 7.2.1.*

the coordinate frame but keeps the vector unchanged physically 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 the square formed by the new two basis vectors is enlarged by a factor of $|\det(P^S_\beta)| = 3$. 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 the area of an identical square is shrunk to $|\det(P^\beta_S)| = |\det((P^S_\beta)^{-1})| = |\det(P^S_\beta)|^{-1} = \frac{1}{3}$ 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 $T : \mathcal{U} \to \mathcal{V}$ that has a matrix representation of $[\vec{v}]_\eta = [T]^\eta_\beta [\vec{u}]_\beta$ where β and η are bases for $\mathcal{U}$ and $\mathcal{V}$ respectively. If we want to change the basis for $\mathcal{U}$ from β to some other basis β' (and similarly η' for $\mathcal{V}$), then the new matrix representation of the linear transformation would be $[\vec{v}]_{\eta'} = [T]^{\eta'}_{\beta'} [\vec{u}]_{\beta'}$. 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 7.1.14 with $[\vec{u}]_\beta = [\mathrm{id}]^\beta_{\beta'} [\vec{u}]_{\beta'} = P^\beta_{\beta'} [\vec{u}]_{\beta'}$ and $[\vec{v}]_\eta = [\mathrm{id}]^\eta_{\eta'} [\vec{v}]_{\eta'} = Q^\eta_{\eta'} [\vec{v}]_{\eta'}$, where we denote the coordinate transition matrix from β' to β by $P^\beta_{\beta'}$ (and similarly η' to η by $Q^\eta_{\eta'}$). Subsequently,

$$\begin{aligned}
[\vec{v}]_\eta &= [T]^\eta_\beta [\vec{u}]_\beta \\
Q^\eta_{\eta'} [\vec{v}]_{\eta'} &= [T]^\eta_\beta P^\beta_{\beta'} [\vec{u}]_{\beta'} \\
[\vec{v}]_{\eta'} &= \left((Q^\eta_{\eta'})^{-1} [T]^\eta_\beta P^\beta_{\beta'} \right) [\vec{u}]_{\beta'}
\end{aligned}$$

Comparing with the assumed form of $[\vec{v}]_{\eta'} = [T]^{\eta'}_{\beta'} [\vec{u}]_{\beta'}$, we can identify $[T]^{\eta'}_{\beta'}$ with $(Q^\eta_{\eta'})^{-1} [T]^\eta_\beta P^\beta_{\beta'}$, 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 $T : \mathcal{U} \to \mathcal{V}$ from bases β and η to β' and η' for $\mathcal{U}$ and $\mathcal{V}$ respectively, follows the relation

$$[T]^{\eta'}_{\beta'} = (Q^\eta_{\eta'})^{-1} [T]^\eta_\beta P^\beta_{\beta'} \tag{7.12}$$

where $P^\beta_{\beta'}$ and $Q^\eta_{\eta'}$ are matrices for change of coordinates on vectors from β' to β and η' to η individually.

Another way to derive the above formula is to consider the linear transformation with respect to the basis β' to η' as three smaller steps: firstly, convert the input vector from the basis β' back to β ($P^{\beta}_{\beta'}$); subsequently, carry out the transformation in terms of β and η ($[T]^{\eta}_{\beta}$); finally, map the vector from the basis η to η' ($Q^{\eta'}_{\eta} = (Q^{\eta}_{\eta'})^{-1}$). This flow is illustrated in the schematic of Figure 7.2.

Figure 7.2.: *A schematic showing how the change of coordinate bases works for linear transformation.*

Example 7.2.2. Use Properties 7.2.1 to redo Example 7.1.2 with respect to new bases $\beta' = \{1, x-1, (x-1)^2\}$ and $\eta' = \{1, x+1\}$.

Solution. First it is instructive to find $P^{\beta}_{\beta'}$ and $Q^{\eta}_{\eta'}$. We leave it to the readers to verify that

$$P^{\beta}_{\beta'} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \qquad Q^{\eta}_{\eta'} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and hence by Properties 7.2.1,

$$\begin{aligned}
[T]^{\eta'}_{\beta'} &= (Q^{\eta}_{\eta'})^{-1}[T]^{\eta}_{\beta}P^{\beta}_{\beta'} \\
&= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -4 \\ 0 & 0 & 2 \end{bmatrix}
\end{aligned}$$

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 β' basis is $(-1,-1,1)^T_{\beta'}$, and the transformation can be described by

$$[T]^{\eta'}_{\beta'}(-1,-1,1)^T_{\beta'} = \begin{bmatrix} 0 & 1 & -4 \\ 0 & 0 & 2 \end{bmatrix}^{\eta'}_{\beta'} \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}_{\beta'} = \begin{bmatrix} -5 \\ 2 \end{bmatrix}_{\eta'}$$

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. □

Most of the time, 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. $T : \mathcal{V} \to \mathcal{V}$[12]. Often we also use the same basis β for the input and output vector. Subsequently, to change the basis for both of them at the same time, let's say β', if the matrix for change of coordinates on vectors from β' to β is denoted as $P = P^{\beta}_{\beta'}$, then (7.12) of Properties 7.2.1 is reduced to $[T]^{\beta'}_{\beta'} = (P^{\beta}_{\beta'})^{-1}[T]^{\beta}_{\beta}P^{\beta}_{\beta'} = P^{-1}AP$ where $A = [T]^{\beta}_{\beta}$ is the original matrix representation of the endomorphism. When

[12] 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.

it is clear from the context, we will simply write $[T]_{\beta}^{\beta}$ and $[T]_{\beta'}^{\beta'}$ as $[T]_{\beta}$ and $[T]_{\beta'}$.

Properties 7.2.2. For a linear endomorphism/operator $T : \mathcal{V} \to \mathcal{V}$, the change of coordinates for its matrix representation from the old basis β to the new one β' is described by the formula

$$[T]_{\beta'} = (P_{\beta'}^{\beta})^{-1}[T]_{\beta}P_{\beta'}^{\beta} \tag{7.13}$$

Speaking loosely, the change of coordinates for a matrix in general takes the form of

$$A' = P^{-1}AP \tag{7.14}$$

In this case, A' and A (or $[T]_{\beta'}$ and $[T]_{\beta}$) are said to be ***similar***.

Example 7.2.3. For a two-dimensional vector space $\mathcal{V}$ with a basis $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}\}$, if a linear endomorphism $T : \mathcal{V} \to \mathcal{V}$ 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 β. Subsequently, if a new basis β' 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

$$\begin{aligned}[T]_{\beta} = \left[[T(\vec{v}^{(1)})]_{\beta} | [T(\vec{v}^{(2)})]_{\beta}\right] &= \left[[\vec{v}^{(1)}]_{\beta} | [\vec{v}^{(1)} + \vec{v}^{(2)}]_{\beta}\right] \\ &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\end{aligned}$$

with respect to the old basis β. The appropriate $P_{\beta'}^{\beta}$ matrix that will be used for Properties 7.2.2, by (7.8), is

$$P_{\beta'}^{\beta} = \left[[\vec{v}^{(1)'}]_{\beta} | [\vec{v}^{(2)'}]_{\beta}\right] = \left[[2\vec{v}^{(1)} - \vec{v}^{(2)}]_{\beta} | [-\vec{v}^{(1)} + \vec{v}^{(2)}]_{\beta}\right]$$

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

and thus by (7.13) the desired new matrix representation of the endomorphism with respect to β' is

$$\begin{aligned} [T]_{\beta'} &= (P^{\beta}_{\beta'})^{-1}[T]_{\beta}P^{\beta}_{\beta'} \\ &= \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \end{aligned}$$

□

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 is orthogonal to all the previously processed vectors by removing the parallel components projected onto them (blue in Figure 7.3) while retaining the orthogonal part (red).

Definition 7.2.3 (Algorithm for Gram-Schmidt Orthogonalization)**.** Given a coordinate basis consisted of $\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \ldots, \vec{u}^{(n)} \in \mathbb{R}^m$, Gram-Schmidt Orthogonalization transforms them into an orthogonal one with $\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots, \vec{v}^{(n)} \in \mathbb{R}^m$ (m and n are not necessarily equal, specifically $n \leq m$) according to the following formulae:

$$\vec{v}^{(1)} = \vec{u}^{(1)}$$

Figure 7.3.: *Similar to Figure 4.6 but with the parallel and orthogonal components of the projection highlighted.*

$$
\begin{aligned}
\vec{v}^{(2)} &= \vec{u}^{(2)} - \text{proj}_{v^{(1)}}\vec{u}^{(2)} = \vec{u}^{(2)} - \frac{\vec{u}^{(2)} \cdot \vec{v}^{(1)}}{\left\|\vec{v}^{(1)}\right\|^2}\vec{v}^{(1)} \\
\vec{v}^{(3)} &= \vec{u}^{(3)} - \text{proj}_{v^{(1)}}\vec{u}^{(3)} - \text{proj}_{v^{(2)}}\vec{u}^{(3)} \\
&= \vec{u}^{(3)} - \frac{\vec{u}^{(3)} \cdot \vec{v}^{(1)}}{\left\|\vec{v}^{(1)}\right\|^2}\vec{v}^{(1)} - \frac{\vec{u}^{(3)} \cdot \vec{v}^{(2)}}{\left\|\vec{v}^{(2)}\right\|^2}\vec{v}^{(2)} \\
&\vdots \\
\vec{v}^{(n)} &= \vec{u}^{(n)} - \text{proj}_{v^{(1)}}\vec{u}^{(n)} - \text{proj}_{v^{(2)}}\vec{u}^{(n)} - \cdots - \text{proj}_{v^{(n-1)}}\vec{u}^{(n)} \\
&= \vec{u}^{(n)} - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(1)}}{\left\|\vec{v}^{(1)}\right\|^2}\vec{v}^{(1)} - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(2)}}{\left\|\vec{v}^{(2)}\right\|^2}\vec{v}^{(2)} - \cdots - \frac{\vec{u}^{(n)} \cdot \vec{v}^{(n-1)}}{\left\|\vec{v}^{(n-1)}\right\|^2}\vec{v}^{(n-1)}
\end{aligned}
\tag{7.15}
$$

In general, for $j \geq 2$, the j-th new vector is computed by

$$
\vec{v}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \text{proj}_{v^{(k)}}\vec{u}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \frac{\vec{u}^{(j)} \cdot \vec{v}^{(k)}}{\left\|\vec{v}^{(k)}\right\|^2}\vec{v}^{(k)} \tag{7.16}
$$

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 $\left\|\hat{v}^{(j)}\right\| = 1$ for all j, and the resulting 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)**.**

$$\begin{aligned}
\hat{v}^{(1)} &= \frac{\vec{u}^{(1)}}{\left\|\vec{u}^{(1)}\right\|} \\
\hat{v}^{(2)} &= \frac{\vec{u}^{(2)} - (\vec{u}^{(2)} \cdot \hat{v}^{(1)})\hat{v}^{(1)}}{\left\|\vec{u}^{(2)} - (\vec{u}^{(2)} \cdot \hat{v}^{(1)})\hat{v}^{(1)}\right\|} \\
\hat{v}^{(3)} &= \frac{\vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)})\hat{v}^{(2)}}{\left\|\vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)})\hat{v}^{(2)}\right\|} \\
&\vdots \\
\hat{v}^{(n)} &= \frac{\vec{u}^{(n)} - (\vec{u}^{(n)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(n)} \cdot \hat{v}^{(2)})\hat{v}^{(2)} - \cdots - (\vec{u}^{(n)} \cdot \hat{v}^{(n-1)})\hat{v}^{(n-1)}}{\left\|\vec{u}^{(n)} - (\vec{u}^{(n)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(n)} \cdot \hat{v}^{(2)})\hat{v}^{(2)} - \cdots - (\vec{u}^{(n)} \cdot \hat{v}^{(n-1)})\hat{v}^{(n-1)}\right\|}
\end{aligned} \tag{7.17}$$

For $j \geq 2$, the general formula is

$$\hat{v}^{(j)} = \frac{\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}}{\left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right\|} \tag{7.18}$$

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 formulae 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

$$\vec{u}^{(2)} - (\vec{u}^{(2)} \cdot \hat{v}^{(1)})\hat{v}^{(1)}$$

$$= \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} - \left[(1)(\frac{1}{3}) + (-1)(\frac{2}{3}) + (0)(\frac{2}{3}) \right] \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} - (-\frac{1}{3}) \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} = \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix}$$

$$\hat{v}^{(2)} = \frac{1}{\sqrt{(\frac{10}{9})^2 + (-\frac{7}{9})^2 + (\frac{2}{9})^2}} \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix} = \frac{3}{\sqrt{17}} \begin{bmatrix} \frac{10}{9} \\ -\frac{7}{9} \\ \frac{2}{9} \end{bmatrix} = \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix}$$

By the same essence, we have the third vector as

$$\vec{u}^{(3)} - (\vec{u}^{(3)} \cdot \hat{v}^{(1)})\hat{v}^{(1)} - (\vec{u}^{(3)} \cdot \hat{v}^{(2)})\hat{v}^{(2)}$$

$$= \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} - \left[(3)(\frac{1}{3}) + (-1)(\frac{2}{3}) + (1)(\frac{2}{3}) \right] \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$$

$$- \left[(3)(\frac{10}{3\sqrt{17}}) + (-1)(-\frac{7}{3\sqrt{17}}) + (1)(\frac{2}{3\sqrt{17}}) \right] \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} - \frac{13}{\sqrt{17}} \begin{bmatrix} \frac{10}{3\sqrt{17}} \\ -\frac{7}{3\sqrt{17}} \\ \frac{2}{3\sqrt{17}} \end{bmatrix} = \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix}$$

$$\hat{v}^{(3)} = \frac{1}{\sqrt{(\frac{2}{17})^2 + (\frac{2}{17})^2 + (-(\frac{3}{17}))^2}} \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix} = \sqrt{17} \begin{bmatrix} \frac{2}{17} \\ \frac{2}{17} \\ -\frac{3}{17} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{17}} \\ \frac{2}{\sqrt{17}} \\ -\frac{3}{\sqrt{17}} \end{bmatrix}$$

□

► Short Exercise: Verify that $\hat{v}^{(1)}, \hat{v}^{(2)}, \hat{v}^{(3)}$ are pairwise orthogonal.[13]

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 an upper-triangular matrix (non-zero entries only found along or above the main diagonal) where the entries 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 $A = [\vec{u}^{(1)}|\vec{u}^{(2)}|\vec{u}^{(3)}|\cdots|\vec{u}^{(n)}]$, and the matrix $Q = [\hat{v}^{(1)}|\hat{v}^{(2)}|\hat{v}^{(3)}|\cdots|\hat{v}^{(n)}]$, where the $\hat{v}^{(j)} \in \mathbb{R}^m$ are orthonormal vectors that come from carrying out Gram-Schmidt orthogonalization on the basis vectors $\vec{u}^{(j)} \in \mathbb{R}^m$, $m \geq n$, and $j = 1, 2, \ldots, n$, according to the Definition 7.2.4, we have $A = QR$, where

$$R = \begin{bmatrix} \vec{u}^{(1)} \cdot \hat{v}^{(1)} & \vec{u}^{(2)} \cdot \hat{v}^{(1)} & \vec{u}^{(3)} \cdot \hat{v}^{(1)} & \cdots & \vec{u}^{(n)} \cdot \hat{v}^{(1)} \\ 0 & \vec{u}^{(2)} \cdot \hat{v}^{(2)} & \vec{u}^{(3)} \cdot \hat{v}^{(2)} & & \vec{u}^{(n)} \cdot \hat{v}^{(2)} \\ 0 & 0 & \vec{u}^{(3)} \cdot \hat{v}^{(3)} & & \vec{u}^{(n)} \cdot \hat{v}^{(3)} \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \vec{u}^{(n)} \cdot \hat{v}^{(n)} \end{bmatrix} \tag{7.19}$$

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

[13] 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

$$
\begin{aligned}
QR^{(j)} &= \left[\hat{v}^{(1)}|\hat{v}^{(2)}|\cdots|\hat{v}^{(j)}|\cdots|\hat{v}^{(n)}\right]\begin{bmatrix}\vec{u}^{(j)}\cdot\hat{v}^{(1)}\\ \vec{u}^{(j)}\cdot\hat{v}^{(2)}\\ \vdots\\ \vec{u}^{(j)}\cdot\hat{v}^{(j)}\\ \vdots\\ 0\end{bmatrix}\\
&= (\vec{u}^{(j)}\cdot\hat{v}^{(1)})\hat{v}^{(1)} + (\vec{u}^{(j)}\cdot\hat{v}^{(2)})\hat{v}^{(2)} + \cdots + (\vec{u}^{(j)}\cdot\hat{v}^{(j)})\hat{v}^{(j)} + 0\\
&= \sum_{k=1}^{j}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)} = \sum_{k=1}^{j-1}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)} + (\vec{u}^{(j)}\cdot\hat{v}^{(j)})\hat{v}^{(j)}
\end{aligned}
$$

By (7.18) from Definition 7.2.4, we have

$$
\hat{v}^{(j)} = \frac{\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)}}{\left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)}\right\|}
$$

[14] 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

$$
\vec{u}^{(j)}\cdot\hat{v}^{(j)} = \left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)}\right\|
$$

[14] Some may ask if $\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)}\cdot\hat{v}^{(k)})\hat{v}^{(k)}\|$ 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 $\hat{v}^{(k)}$ is a linear combination of $\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(k)}$, 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 $c_1\vec{u}^{(1)} + \cdots + c_{j-1}\vec{u}^{(j-1)} + c_j\vec{u}^{(j)} = \mathbf{0}$ where particularly $c_j = 1 \neq 0$.

Consider

$$\left[\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right] \cdot \hat{v}^{(j)} = \vec{u}^{(j)} \cdot \hat{v}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})(\hat{v}^{(k)} \cdot \hat{v}^{(j)})$$
$$= \vec{u}^{(j)} \cdot \hat{v}^{(j)}$$

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,

$$\left[\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right] \cdot \hat{v}^{(j)} = \left(\left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right\|\hat{v}^{(j)}\right) \cdot \hat{v}^{(j)}$$
$$= \left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right\|(\hat{v}^{(j)} \cdot \hat{v}^{(j)})$$
$$= \left\|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\right\|$$

as $\hat{v}^{(j)} \cdot \hat{v}^{(j)} = \left\|\hat{v}^{(j)}\right\|^2 = 1$. The required equality is then established and the result follows. The invertibility of R can be shown by noting that all diagonal components $\vec{u}^{(j)} \cdot \hat{v}^{(j)} = \|\vec{u}^{(j)} - \sum_{k=1}^{j-1}(\vec{u}^{(j)} \cdot \hat{v}^{(k)})\hat{v}^{(k)}\|$ of the upper-triangular R matrix are non-zero (see Footnote 14). □

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 (7.19) in 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 two small remarks, one related to computing the coordinates relative to an orthonormal basis, and another related to the concept of orthogonal complement discussed at the end of Section 6.3.2.

Properties 7.2.6. The coordinates of a vector $\vec{v}$ in an orthonormal coordinate system β are simply its dot products (projections onto) with the orthonormal basis vectors in β.

Proof. As β is an orthonormal coordinate system, according to the discussion in Section 6.2.2, $\vec{v}$ can be written as a unique linear combination of the orthonormal basis vectors $\vec{v} = [v_1]_\beta \hat{e}_\beta^{(1)} + [v_2]_\beta \hat{e}_\beta^{(2)} + \cdots + [v_n]_\beta \hat{e}_\beta^{(n)}$. By taking the dot product with $\hat{e}_\beta^{(1)}$, we have

$$\begin{aligned} \vec{v} \cdot \hat{e}_\beta^{(1)} &= [v_1]_\beta (\hat{e}_\beta^{(1)} \cdot \hat{e}_\beta^{(1)}) + [v_2]_\beta (\hat{e}_\beta^{(2)} \cdot \hat{e}_\beta^{(1)}) + \cdots + [v_n]_\beta (\hat{e}_\beta^{(n)} \cdot \hat{e}_\beta^{(1)}) \\ &= [v_1]_\beta (1) + [v_2]_\beta (0) + \cdots + [v_n]_\beta (0) \\ &\quad \text{(Orthonormality of the basis vectors)} \\ &= [v_1]_\beta \end{aligned}$$

which is also the projection of $\vec{v}$ onto $\hat{e}_\beta^{(1)}$ by Properties 5.2.1. The same applies to the other coordinate components. □

Properties 7.2.7. For an orthogonal(-normal) basis $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \cdots, \vec{v}^{(n)}\}$ for a finite, n-dimensional vector space $\mathcal{V}$, the subspaces $\mathcal{V}_\gamma$ and $\mathcal{V}_\eta$ formed by $\gamma = \{\vec{v}^{(I)}\}$ and $\eta = \{\vec{v}^{(J)}\}$ 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 $\mathcal{V}_\gamma^\perp = \mathcal{V}_\eta$ ($\mathcal{V}_\eta^\perp = \mathcal{V}_\gamma$) and $\mathcal{V}_\gamma \oplus \mathcal{V}_\eta = \mathcal{V}$.

7.3. Python Programming

We can define a function for 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 7.1.14, which is equal to $B'^{-1}B$ as discussed at 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 β, we have

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

and $B' = I$ as $\beta' = \mathcal{S}$ 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.

```
B1 = np.array([[1.,  1.],
               [2., -1.]])

P1 = P_matrix(B1, np.identity(2))
old_v = np.array([2., 1.])
new_v = coord_trans_vector(old_v, P1)
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

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

Subsequently,

```
B2 = np.array([[2., -1.],
              [-1.,  1.]])
old_A = np.array([[1.,  1.],
               [0.,  1.]])

P2 = P_matrix(B2, np.identity(2))
new_A = coord_trans_matrix(old_A, P2)
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 $\beta = \{(1,1,1)^T, (1,1,0)^T, (1,0,0)^T\}$ and $\eta = \{(1,2,3)^T, (1,-1,0)^T, (2,-1,-1)^T\}$ as bases for $\mathcal{V}$ and $\mathcal{W}$. If a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ is defined by $T((x,y,z)^T) = (x+y+z, 2x+y, x-$

$y - 3z)^T$, find its matrix representation $[T]_\beta^\eta$ 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 $\beta = \{\cos x, \sin x, x\cos x, x\sin x\}$ and $T : \mathcal{V} \to \mathcal{V}, T[f(x)] = f'(x)$ be the differentiation operator over $\mathcal{V}$. Find the matrix representation of T with respect to β, 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 $T : \mathcal{V} \to \mathcal{W}$ 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 $T : \mathcal{V} \to \mathcal{V}, T(\vec{v}) = \mathrm{id}(\vec{v}) = \vec{v}$ for a finite-dimensional vector space $\mathcal{V}$ with respect to a fixed basis β throughout always has a matrix representation of an identity matrix such that $[T]_\beta = I$.

Exercise 7.6 For a linear operator $T : \mathcal{P}^2 \to \mathcal{P}^2$, $T(p(x)) = p(x - 1)$, 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 $T : \mathcal{U} \to \mathcal{V}$ and $S : \mathcal{V} \to \mathcal{W}$, where $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are finite-dimensional with bases β, η, κ, find the formula for change of coordinates to new bases β', η', κ' 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) $\vec{u}^{(1)} = (1, -1, 2, 1)^T, \vec{u}^{(2)} = (1, 2, 0, 2)^T, \vec{u}^{(3)} = (2, 1, -1, 0)^T$.

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 *complex vector spaces*, and compare them to real vector spaces that we just learnt in the previous chapters. Working with complex numbers and *Euler's Formula* will make many calculations faster, such as solving systems of Ordinary Differential Equations in Chapters 9, and the Discrete Fourier Transform in Chapter 14. The second one is about the *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 led to the square root of a negative quantity, which was undefined back in the day. 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 $-\frac{2}{3} \pm \frac{1}{3}i$.

Definition 8.1.1 (Complex Number)**.** Complex numbers are scalars in the form of $z = a + bi$, where a and b are some real numbers. Their real and imaginary parts are denoted by $\mathrm{Re}\{z\} = a$ and $\mathrm{Im}\{z\} = b$.

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 $\mathrm{Re}\{z_1\} = a = c = \mathrm{Re}\{z_2\}$ and $\mathrm{Im}\{z_1\} = b = d = \mathrm{Im}\{z_2\}$.

For every complex number, there exists another corresponding complex number known as the ***(complex) conjugate*** associated with it, formed by flipping the sign of its imaginary part.

Definition 8.1.3 (Complex Conjugate)**.** For a complex number $z = a + bi$, its complex conjugate is defined as $\overline{z} = a - bi$.

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 are 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

$$\begin{aligned} z_1 \pm z_2 &= (a + bi) \pm (c + di) \\ &= (a \pm c) + (b \pm d)i \\ &= (\mathrm{Re}\{z_1\} \pm \mathrm{Re}\{z_2\}) + (\mathrm{Im}\{z_1\} \pm \mathrm{Im}\{z_2\})i \end{aligned} \tag{8.1}$$

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

$$\begin{aligned}(a+bi)(c+di) &= a(c+di)+bi(c+di)\\ &= ac+adi+bci+bdi^2\\ &= (ac-bd)+(ad+bc)i \qquad (i^2=-1)\end{aligned} \tag{8.2}$$

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

$$\begin{aligned}\frac{1}{a+bi} &= \frac{1}{a+bi}\frac{a-bi}{a-bi}\\ &= \frac{a-bi}{a^2-(-b^2)-abi+bai}\\ &= \frac{a-bi}{a^2+b^2}\end{aligned} \tag{8.3}$$

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 denominator 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

Figure 8.1.: *A complex number $z = 3 + 4i$ represented in the complex plane.*

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 a vector is

$$|z| = \sqrt{\mathrm{Re}\{z\}^2 + \mathrm{Im}\{z\}^2} \tag{8.4}$$

which is called the ***modulus*** of the corresponding complex number. In the diagram (Figure 8.1) 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

$$\theta = \arctan(\mathrm{Im}\{z\}/\mathrm{Re}\{z\}) \tag{8.5}$$

in the same figure. Since its complex conjugate $\overline{z}$ 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 $\mathrm{Re}\{z\} = |z|\cos\theta$ and $\mathrm{Im}\{z\} = |z|\sin\theta$. Hence z can be represented as $z = \mathrm{Re}\{z\} + i\,\mathrm{Im}\{z\} = |z|(\cos\theta + i\sin\theta)$. 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

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{8.6}$$

where θ is taken to be real.

Hence z can be further written as $z = |z|e^{i\theta}$, and $\overline{z} = \mathrm{Re}\{z\} - i\,\mathrm{Im}\{z\} = |z|(\cos\theta - i\sin\theta) = |z|(\cos(-\theta) + i\sin(-\theta)) = |z|e^{-i\theta}$. Conversely, the quantity $e^{i\theta}$ can be regarded as a complex number that has a modulus of 1 and an argument of θ. Additionally, this provides formulae to express sines and cosines with complex exponentials.

Properties 8.1.7. For any θ which is confined to be real,

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \tag{8.7a}$$

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} \tag{8.7b}$$

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. □

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 $z_1 z_2 = (|z_1|e^{i\theta_1})(|z_2|e^{i\theta_2}) = |z_1||z_2|e^{i(\theta_1+\theta_2)}$.[1] 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 θ_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

$$\begin{aligned}\frac{1}{z} = \frac{1}{|z|e^{i\theta}} = \frac{1}{|z|}\frac{e^{-i\theta}}{e^{i\theta}e^{-i\theta}} = \frac{1}{|z|}\frac{e^{-i\theta}}{e^{i(\theta-\theta)}} = \frac{1}{|z|}\frac{e^{-i\theta}}{e^{0}} &= \frac{1}{|z|}e^{-i\theta} \\ &= \frac{1}{|z|^2}(|z|e^{-i\theta})\end{aligned}$$

[1] We take it for granted that $e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$.

$$= \frac{1}{|z|^2}\overline{z}$$

which is consistent with the discussion about complex division in the last subsection, where $|z|^2 = a^2 + b^2$ arises in the denominator. In addition, we can observe that $|z|^2 = z\overline{z}$. This is not a coincidence, as

$$\begin{aligned} z\overline{z} &= |z|e^{i\theta}|z|e^{-i\theta} \\ &= |z|^2 e^{i(\theta-\theta)} = |z|^2 e^0 = |z|^2 \end{aligned} \tag{8.8}$$

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 θ, 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 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

$$(e^{i\theta})^n = e^{i(n\theta)} \tag{8.9a}$$

$$(\cos\theta + i\sin\theta)^n = \cos(n\theta) + i\sin(n\theta) \tag{8.9b}$$

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 entries. A complex vector is simply a vector that has complex numbers 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 at the end of this section. A complex matrix is similarly a matrix with complex entries, 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 component-wise. Multiplication by a scalar is also similar, applied to all components. 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 components, but additionally with the conjugate operation applied on the second

complex vector beforehand.

$$\begin{aligned}\vec{u}\cdot\vec{v} &= \mathbf{u}^T\overline{\mathbf{v}}\\ &= u_1\overline{v_1}+u_2\overline{v_2}+\cdots+u_n\overline{v_n}=\sum_{k=1}^{n}u_k\overline{v_k}\end{aligned}\tag{8.10}$$

The bar on $\overline{\mathbf{v}}$ means carrying out conjugate on every entry of $\mathbf{v}$. If $\mathbf{v} = \mathrm{Re}\{\mathbf{v}\} + i\,\mathrm{Im}\{\mathbf{v}\}$, where $\mathrm{Re}\{\mathbf{v}\}$ and $\mathrm{Im}\{\mathbf{v}\}$ are the vectors consisting of the real/imaginary parts of every entry in $\mathbf{v}$, then $\overline{\mathbf{v}} = \mathrm{Re}\{\mathbf{v}\} - i\,\mathrm{Im}\{\mathbf{v}\}$.

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

$$\begin{aligned}\|\vec{v}\| &= \sqrt{\vec{v}\cdot\vec{v}}=\sqrt{\mathbf{v}^T\overline{\mathbf{v}}}\\ &= \sqrt{v_1\overline{v_1}+v_2\overline{v_2}+\cdots+v_n\overline{v_n}}\\ &= \sqrt{|v_1|^2+|v_2|^2+\cdots+|v_n|^2} && \text{(by (8.8))}\\ &= \sqrt{\sum_{k=1}^{n}|v_k|^2}\end{aligned}\tag{8.11}$$

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

$\vec{u}\cdot\vec{v} = \overline{\vec{v}\cdot\vec{u}}$	Conjugate-symmetric 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
$(a\vec{u})\cdot(b\vec{v}) = a\bar{b}(\vec{u}\cdot\vec{v})$	where a, b are some complex constants

There is no complex analog for cross product.

Example 8.2.1. Show that the conjugate-symmetric property in Properties 8.2.3 holds for $\vec{u} = (1+2i, 3+i)^T$, $\vec{v} = (2-5i, 1+4i)^T$.

Solution.

$$\begin{aligned}\vec{u}\cdot\vec{v} &= (1+2i)(\overline{2-5i}) + (3+i)(\overline{1+4i})\\ &= (1+2i)(2+5i) + (3+i)(1-4i)\\ &= (-8+9i) + (7-11i)\\ &= -1-2i\end{aligned}$$

$$\begin{aligned}\vec{v}\cdot\vec{u} &= (2-5i)(\overline{1+2i}) + (1+4i)(\overline{3+i})\\ &= (2-5i)(1-2i) + (1+4i)(3-i)\\ &= (-8-9i) + (7+11i)\\ &= -1+2i\end{aligned}$$

Hence $\vec{u}\cdot\vec{v} = \overline{\vec{v}\cdot\vec{u}}$. □

► Short Exercise: Find the norm $\|\vec{u}\|$ and $\|\vec{v}\|$ respectively.[2]

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 the 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 entries resulting 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.

Definition 8.2.4 (Conjugate Transpose)**.** The ***conjugate transpose*** of a matrix A, denoted as $A^* = \overline{A^T}$, has its entries as $A^*_{pq} = \overline{A}_{qp}$, where $\overline{A}$ is the conjugate of the matrix A produced by changing every entry 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 entries 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 $A^* = A$.

Note that the diagonal entries of a Hermitian matrix are always real. Properties of conjugate transpose are alike to those for real transpose stated in Properties 2.1.4. Furthermore, for complex dot product, there is 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. $(A^*)^* = A$;

[2] $\|\vec{u}\| = \sqrt{(1+2i)(1-2i)+(3+i)(3-i)} = \sqrt{(1^2+2^2)+(3^2+1^2)} = \sqrt{15}$. Similarly, $\|\vec{v}\| = \sqrt{46}$.

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

$$\vec{u} \cdot (A\vec{v}) = \mathbf{u}^T \overline{(A\mathbf{v})} = (\overline{A^T}\mathbf{u})^T \overline{\mathbf{v}} = (A^*\vec{u}) \cdot \vec{v}$$
$$(A\vec{u}) \cdot \vec{v} = (A\mathbf{u})^T \overline{\mathbf{v}} = \mathbf{u}^T (A^T \overline{\mathbf{v}}) = \vec{u} \cdot (A^*\vec{v})$$

where Properties 2.1.4 and Definitions 8.2.1, 8.2.4 are used.

With the 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 entries of the complex matrix product $A\overline{B}$ where a conjugate is applied on the second matrix.

Example 8.2.2. For two complex matrices

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

Verify that $(AB)^* = B^* A^*$.

Solution.

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

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

$$= \begin{bmatrix} -i & i \\ 3-4i & 3+2i \end{bmatrix}$$

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

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

$$= \begin{bmatrix} i & 3+4i \\ -i & 3-2i \end{bmatrix}$$

$$(AB)^* = \begin{bmatrix} -i & i \\ 3-4i & 3+2i \end{bmatrix}$$

□

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, and the result is

$$\det(A) = -(1+i)\begin{vmatrix} 3 & 2 \\ -2i & 1 \end{vmatrix} + (0)\begin{vmatrix} 1-i & 2 \\ 2 & 1 \end{vmatrix} - (i)\begin{vmatrix} 1-i & 3 \\ 2 & -2i \end{vmatrix}$$

$$
\begin{aligned}
&= -(1+i)(3+4i) - (i)(-8-2i) \\
&= -1+i
\end{aligned}
$$

□

Example 8.2.4. Find the inverse of the matrix A in the last example.

Solution. The computation of the 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

$$
\begin{aligned}
C &= \begin{bmatrix} \begin{vmatrix} 0 & i \\ -2i & 1 \end{vmatrix} & -\begin{vmatrix} 1+i & i \\ 2 & 1 \end{vmatrix} & \begin{vmatrix} 1+i & 0 \\ 2 & -2i \end{vmatrix} \\ -\begin{vmatrix} 3 & 2 \\ -2i & 1 \end{vmatrix} & \begin{vmatrix} 1-i & 2 \\ 2 & 1 \end{vmatrix} & -\begin{vmatrix} 1-i & 3 \\ 2 & -2i \end{vmatrix} \\ \begin{vmatrix} 3 & 2 \\ 0 & i \end{vmatrix} & -\begin{vmatrix} 1-i & 2 \\ 1+i & i \end{vmatrix} & \begin{vmatrix} 1-i & 3 \\ 1+i & 0 \end{vmatrix} \end{bmatrix} \\
&= \begin{bmatrix} -2 & -1+i & 2-2i \\ -3-4i & -3-i & 8+2i \\ 3i & 1+i & -3-3i \end{bmatrix}
\end{aligned}
$$

Thus, by Properties 2.3.11, the inverse of A is

$$
A^{-1} = \frac{1}{\det(A)}\operatorname{adj}(A) = \frac{1}{\det(A)}C^T
$$

$$= -\frac{1+i}{2}\begin{bmatrix} -2 & -3-4i & 3i \\ -1+i & -3-i & 1+i \\ 2-2i & 8+2i & -3-3i \end{bmatrix}$$

$$= \begin{bmatrix} 1+i & -\frac{1}{2}+\frac{7}{2}i & \frac{3}{2}-\frac{3}{2}i \\ 1 & 1+2i & -i \\ -2 & -3-5i & 3i \end{bmatrix}$$

□

► Short Exercise: Find A^{-1} via Gaussian Elimination.[3]

Below are some useful properties of determinants and inverses for complex matrices that can be compared to Properties 2.3.9 and 2.2.3.

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 $\vec{v} = (v_1, v_2, v_3, \ldots, v_n)^T$, where v_i can be any *complex* numbers, for $i = 1, 2, 3, \ldots, n$. 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

[3]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 $\frac{1+i}{2}$ to create a leading 1 and subtracting $1+i$ and 2 times the new first row from the second and third row respectively.

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 7.1.15, 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) to be taken over $\mathbb{C}$ (the former case) onwards. All major results from the last two chapters are still valid if we replace $\mathbb{R}$ and $\mathbb{R}^n$ by $\mathbb{C}$ and $\mathbb{C}^n$ appropriately.[4] In particular, we note that

Properties 8.2.10. The angle between two complex vectors $\vec{u}$ and $\vec{v}$ is found according to

$$\cos\theta = \frac{\mathrm{Re}\{\vec{u} \cdot \vec{v}\}}{\|u\|\|v\|} \tag{8.12}$$

which should be compared to (4.6).

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 2×2 block matrix

[4]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.

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 3×5 matrix at the same time a 2×2 block matrix where

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

are of the shapes 2×3, 2×2, 1×3 and 1×2. We can extend this for block matrices of any partition. For instance, a 3×4 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} 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 2×2 block matrices:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \qquad N = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$$

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 ordinary matrices, but how about the submatrices? In fact, we just carry out the multiplication as if each of them is a single entry, such that

$$MN = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{bmatrix} \tag{8.13}$$

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 $r_1 \times n_1$, $r_1 \times n_2$, where $m_1 + m_2 = m$, $n_1 + n_2 = n$. 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

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \qquad N = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$$

as a 3×3 and 3×2 matrix respectively, with

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

$$C = \begin{bmatrix} 0 & -1 \end{bmatrix} \qquad D = \begin{bmatrix} 1 \end{bmatrix}$$

$$X = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$Z = \begin{bmatrix} -1 \end{bmatrix} \qquad W = \begin{bmatrix} 1 \end{bmatrix}$$

such that the partitions look like

$$M = \left[\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 1 & -2 \\ \hline 0 & -1 & 1 \end{array}\right] \qquad N = \left[\begin{array}{c|c} 0 & 1 \\ 2 & -1 \\ \hline -1 & 1 \end{array}\right]$$

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 Formula (8.13) 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

$$AX = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \qquad BZ = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$AY = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad BW = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$CX = \begin{bmatrix} 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \end{bmatrix} \qquad DZ = \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} = \begin{bmatrix} -1 \end{bmatrix}$$

$$CY = \begin{bmatrix} 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} \qquad DW = \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}$$

Hence

$$MN = \begin{bmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{bmatrix}$$

$$= \left[\begin{array}{c|c} \begin{bmatrix}4\\2\end{bmatrix} + \begin{bmatrix}-1\\2\end{bmatrix} & \begin{bmatrix}-1\\-1\end{bmatrix} + \begin{bmatrix}1\\-2\end{bmatrix} \\ \hline \begin{bmatrix}-2\end{bmatrix} + \begin{bmatrix}-1\end{bmatrix} & \begin{bmatrix}1\end{bmatrix} + \begin{bmatrix}1\end{bmatrix} \end{array}\right] = \left[\begin{array}{c|c} 3 & 0 \\ 4 & -3 \\ \hline -3 & 2 \end{array}\right]$$

The readers can check the answer by computing the matrix product in the usual way. □

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

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & \cdots & M_{1r} \\ M_{21} & M_{22} & M_{23} & & M_{2r} \\ M_{31} & M_{32} & M_{33} & & M_{3r} \\ \vdots & & & \ddots & \vdots \\ M_{m1} & M_{m2} & M_{m3} & & M_{mr} \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} N_{11} & N_{12} & N_{13} & \cdots & N_{1n} \\ N_{21} & N_{22} & N_{23} & & N_{2n} \\ N_{31} & N_{32} & N_{33} & & N_{3n} \\ \vdots & & & \ddots & \vdots \\ N_{r1} & N_{r2} & N_{r3} & & N_{rn} \end{bmatrix}$$

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 2×2 block matrices only, we will handle them exclusively. Specifically, we consider 2×2 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 $\det\left(A^{-1}\right) = (\det(A))^{-1}$.

Proof. For the first matrix, simply note that it is a lower-triangular matrix with all diagonal entries being 1, and hence 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 $\det\left(A^{-1}\right) = (\det(A))^{-1}$ (Properties 2.3.9). If A is invertible, then $\det\left(A^{-1}\right) = (\det(A))^{-1}$ 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

$$\begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -C & I_q \end{bmatrix} \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} I_p A^{-1} + [\mathbf{0}]_{p\times q}[\mathbf{0}]_{q\times p} & I_p[\mathbf{0}]_{p\times q} + [\mathbf{0}]_{p\times q} I_q \\ (-C)A^{-1} + I_q[\mathbf{0}]_{q\times p} & -C[\mathbf{0}]_{p\times q} + I_q I_q \end{bmatrix} \\
&= \begin{bmatrix} A^{-1} & [\mathbf{0}]_{p\times q} \\ -CA^{-1} & I_q \end{bmatrix}
\end{aligned}
$$

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

$$
\begin{aligned}
\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p\times q} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} A^{-1}A + [\mathbf{0}]_{p\times q}C & A^{-1}B + [\mathbf{0}]_{p\times q}D \\ -CA^{-1}A + I_q C & -CA^{-1}B + I_q D \end{bmatrix} \\
&= \begin{bmatrix} I_p & A^{-1}B \\ -C + C & -CA^{-1}B + D \end{bmatrix} \\
&= \begin{bmatrix} I_p & A^{-1}B \\ [\mathbf{0}]_{q\times p} & D - CA^{-1}B \end{bmatrix}
\end{aligned}
$$

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 then 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,

$$
\begin{aligned}
&\begin{bmatrix} A^{-1} & [\mathbf{0}]_{p\times q} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \\
&= \begin{bmatrix} I_p & A^{-1}B \\ [\mathbf{0}]_{q\times p} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} = \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & D - CA^{-1}B \end{bmatrix}
\end{aligned}
$$

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.[5] By Properties 2.2.3, we 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 2×2 block matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A and D are square submatrices, if A and its Schur complement $M/A =$

[5]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

$$\begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & D - CA^{-1}B \end{bmatrix}^{-1} = \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

$D - CA^{-1}B$ of block A are both invertible, then M is invertible with

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{bmatrix} \tag{8.14}$$

Properties 8.3.3. The determinant of the 2×2 block matrix in Properties 8.3.2 is

$$\det(M) = \det(A)\det\left(D - CA^{-1}B\right) = \det(A)\det(M/A) \tag{8.15}$$

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 (Properties 2.3.9) 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

$$\begin{aligned} (\det(A))^{-1}\det(M)(1) &= \det\left(D - CA^{-1}B\right) \\ \det(M) &= \det(A)\det\left(D - CA^{-1}B\right) = \det(A)\det(M/A) \end{aligned}$$

□

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 it to the readers to verify that

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

$$M/A = D - CA^{-1}B = \begin{bmatrix} 2 \end{bmatrix} - \begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \end{bmatrix}$$

Then by Properties 8.3.2, we have

$$\begin{aligned} M^{-1} &= \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{bmatrix} \\ &= \left[\begin{array}{c|c} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} & -\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} \\ \hline -\begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 \end{bmatrix} \end{array}\right] \\ &= \left[\begin{array}{cc|c} -3 & 4 & -2 \\ 2 & -2 & 1 \\ \hline -2 & 3 & -1 \end{array}\right] \end{aligned}$$

Meanwhile, by Properties 8.3.3,

$$\det(M) = \det(A)\det(M/A)$$

$$= \det\left(\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}\right) \det([-1]) = (1)(-1) = -1$$

□

Similar results are also available in terms of the Schur complement using block D instead of A.

Properties 8.3.4. For the 2×2 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

$$M^{-1} = \begin{bmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{bmatrix} \tag{8.16}$$

and its determinant can be computed by

$$\det(M) = \det(D)\det\left(A - BD^{-1}C\right) = \det(D)\det(M/D) \tag{8.17}$$

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 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 (Restriction of a Linear Transformation)**.** Given a linear transformation $T : \mathcal{U} \to \mathcal{V}$ and a proper subspace $\mathcal{W} \subset \mathcal{U}$, the restriction of T to $\mathcal{W}$ is defined as

$$T|_W : \mathcal{W} \to \mathcal{V}, T|_W(\vec{w}) = T(\vec{w}) \text{ for any } \vec{w} \in \mathcal{W}.$$

In simpler terms, $T|_W$ works exactly as T but is 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 β_W with r generating vectors (Properties 6.2.4), which by part (c) of Properties 6.2.7 can be extended to a new basis $\beta' = \beta_W \cup \gamma$ for $\mathcal{U}$, where γ contains $n - r$ vectors and $\beta' = \beta_W \cup \gamma$ 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 standpoint of a block matrix.

To see this, let $\beta_W = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(r)}\}$ and $\gamma = \{\vec{u}^{(r+1)}, \ldots, \vec{u}^{(n)}\}$, and thus $\beta' = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(r)}, \vec{u}^{(r+1)}, \ldots, \vec{u}^{(n)}\}$. By Definition 6.2.9, since the vectors in $\beta' = \beta_W \cup \gamma$ are designed to be linearly independent, the subspace $\mathcal{W}^C$ generated by γ will be the complement of $\mathcal{W}$ as a result of $\mathcal{W} \oplus \mathcal{W}^C$ being a direct sum that produces the n-dimensional $\mathcal{U}$ (Properties 6.2.10). Any $\vec{w} \in \mathcal{W} \subset \mathcal{U}$ will then have a coordinate representation of

$$(w_1, w_2, \ldots, w_r, 0, \ldots, 0)^T$$

in the β' basis where components beyond the r-th index are all zeros. From the perspective of direct sum, it is the same as $\vec{w} \oplus \mathbf{0}_{n-r} = (w_1, w_2, \ldots, w_r)^T_{\beta_W} \oplus (0, \ldots, 0)^T_\gamma$, i.e. the components of $\vec{w}$ for $\mathcal{W}^C$ are all zeros. By Definition 7.1.2, writing out the matrix representation of $[T]^\eta_{\beta'}$ where η is an arbitrary basis for the m-dimensional $\mathcal{V}$ results in

$$[T]^\eta_{\beta'} = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} & a_1^{(r+1)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} & a_2^{(r+1)} & & a_2^{(n)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} & a_m^{(r+1)} & \cdots & a_m^{(n)} \end{bmatrix}$$

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

$$
\begin{aligned}
[T]^{\eta}_{\beta'}[\vec{w}]_{\beta'} &= \left[\begin{array}{cccc|ccc} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} & a_1^{(r+1)} & \cdots & a_1^{(n)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} & a_2^{(r+1)} & & a_2^{(n)} \\ \vdots & & & \vdots & \vdots & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} & a_m^{(r+1)} & \cdots & a_m^{(n)} \end{array}\right]^{\eta}_{\beta_W+\gamma} \left[\begin{array}{c} w_1 \\ w_2 \\ \vdots \\ w_r \\ \hline 0 \\ \vdots \\ 0 \end{array}\right]_{\beta_W+\gamma} \\
&= [(T|_W) \mid (T|_{W^C})]^{\eta}_{\beta_W+\gamma}[\vec{w} \oplus \mathbf{0}]^{T}_{\beta_W+\gamma} \qquad (8.18)
\end{aligned}
$$

We can simply ignore $[T|_{W^C}]^{\eta}_{\gamma}$, the block at the right of the $[T]^{\eta}_{\beta'}$ partition as well as discard the all-zero components of $[\vec{w}]_{\beta'}$ starting from the $(r+1)$-th index, and keep only the other block $[T|_W]^{\eta}_{\beta_W}$ at the left and the $[\vec{w}]_{\beta_W}$ part. The output of the truncated multiplication

$$
[T|_W]^{\eta}_{\beta_W}[\vec{w}]_{\beta_W} = \begin{bmatrix} a_1^{(1)} & a_1^{(2)} & \cdots & a_1^{(r)} \\ a_2^{(1)} & a_2^{(2)} & & a_2^{(r)} \\ \vdots & & & \vdots \\ a_m^{(1)} & a_m^{(2)} & \cdots & a_m^{(r)} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_r \end{bmatrix} \qquad (8.19)
$$

will represent the same vector in $\mathcal{W} \subset \mathcal{U}$ as that mapped from the full form $[T]^{\eta}_{\beta'}[\vec{w}]_{\beta'}$.

Properties 8.3.6. For a linear transformation $T : \mathcal{U} \to \mathcal{V}$ between two finite-dimensional spaces, if a proper subspace $\mathcal{W}$ of $\mathcal{U}$ is generated by a basis $\beta_W = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \ldots, \vec{w}^{(r)}\}$, then the matrix representation of the restriction of T to $\mathcal{W}$ with respect to β_W and η will be given by

$$
[T|_W]^{\eta}_{\beta_W} = \left[[T(\vec{w}^{(1)})]_{\eta} | [T(\vec{w}^{(2)})]_{\eta} | \cdots | [T(\vec{w}^{(r)})]_{\eta}\right] \qquad (8.20)
$$

where η is any basis for $\mathcal{V}$. (This can be compared to Definition 7.1.2.)

In general, the effect of a linear transformation $T : \mathcal{U} \to \mathcal{V}$ applied to $\vec{u} \in \mathcal{U}$ is equivalent to the sum of responses from the restrictions of T to a set of subspaces $\mathcal{W}_1, \mathcal{W}_2, \cdots, \mathcal{W}_s$ 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 $\vec{w}_1 \in \mathcal{W}_1, \vec{w}_2 \in \mathcal{W}_2, \cdots, \vec{w}_s \in \mathcal{W}_s$ of $\vec{u} = \vec{w}_1 + \vec{w}_2 + \cdots + \vec{w}_s$ in these smaller subspaces: $T(\vec{u}) = T(\vec{w}_1) + T(\vec{w}_2) + \cdots + T(\vec{w}_s)$.

Example 8.3.3. Given a linear transformation $T : \mathcal{U} \to \mathcal{V}$ that has a matrix representation of

$$[T]_\beta^\eta = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}$$

with respect to some bases $\beta = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}\}$ and $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$ for $\mathcal{U}$ and $\mathcal{V}$, find the restriction of T to $\mathcal{W}$, where $\mathcal{W} \subset \mathcal{U}$ has a basis of $\beta_W = \{\vec{w}^{(1)}, \vec{w}^{(2)}\}$, 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 $P_{\beta_W}^\beta$ from β_W to β as devised in Theorem 7.1.14 appropriate in this situation is a 3×2 matrix instead since there are only two basis vectors in β_W, and it can be easily seen to be

$$\begin{aligned} P_{\beta_W}^\beta &= \left[[\vec{w}^{(1)}]_\beta | [\vec{w}^{(2)}]_\beta \right] \\ &= \left[[\vec{u}^{(1)} + \vec{u}^{(2)}]_\beta | [\vec{u}^{(1)} + \vec{u}^{(2)} + \vec{u}^{(3)}]_\beta \right] \\ &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

and we are to find $[\vec{w}^{(3)}]_\beta$ to complete a basis $\beta' = \beta_W \cup \{\vec{w}^{(3)}\}$ and hence $P_{\beta'}^\beta$. An algorithm to do so, motivated by Footnote 11 in Chapter 6, is to apply

Gaussian Elimination to $P^{\beta}_{\beta_W}$ 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.

$$\begin{aligned}
\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} &\to \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} && R_2 - R_1 \to R_2 \\
&\to \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} && R_2 \leftrightarrow R_3 \\
&\to \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} && R_1 - R_2 \to R_1
\end{aligned}$$

$$\begin{aligned}
\left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] &\to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] && R_1 + R_2 \to R_1 \\
&\to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right] && R_2 \leftrightarrow R_3 \\
&\to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{array}\right] && R_2 + R_1 \to R_2
\end{aligned}$$

So $[\vec{w}^{(3)}]_\beta = (0,1,0)^T_\beta$ is a possible choice. This simple algorithm remains straightforward when the number of dimensions and vectors to be appended becomes much larger.[6] Now

$$P^{\beta}_{\beta'} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

[6] In fact, we only need to keep track of the row-swapping operations with full-zero rows.

and by Properties 7.2.1

$$
\begin{aligned}
[T]^{\eta}_{\beta'} &= [T]^{\eta}_{\beta} P^{\beta}_{\beta'} \\
&= \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 2 & 4 & 1 \\ 2 & 3 & 0 \\ 0 & 1 & -1 \end{bmatrix}
\end{aligned}
$$

The matrix representation of the restriction of T to $\mathcal{W}$ with respect to β_W then agrees with the first two columns of $[T]^{\eta}_{\beta'}$. The third column of $[T]^{\eta}_{\beta'}$ that characterizes the action of $T|_{W_C}$, is removed. These lead to

$$
[T|_W]^{\eta}_{\beta_W} = \begin{bmatrix} 2 & 4 \\ 2 & 3 \\ 0 & 1 \end{bmatrix}
$$

□

► Short Exercise: Directly apply Properties 8.3.6 to redo the example above.[7]

With the concept of restriction, we can now introduce the matrix analog of a direct sum. For a linear transformation $T : \mathcal{U} \to \mathcal{V}$, if the vector spaces (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

$$
\mathcal{R}(T|_W) \subseteq \mathcal{Y} \qquad\qquad \mathcal{R}(T|_{W^C}) \subseteq \mathcal{Y}^C
$$

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 $\vec{u} = \vec{w} + \vec{w}^C$ 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 $\vec{v} = \vec{y} + \vec{y}^C$ 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 $T_1 : \mathcal{U}_1 \to \mathcal{V}_1$ and $T_2 : \mathcal{U}_2 \to \mathcal{V}_2$ is $T = T_1 \oplus T_2$ such that for any vector direct sum $\vec{u} = \vec{u}_1 + \vec{u}_2$ in $\mathcal{U} = \mathcal{U}_1 \oplus \mathcal{U}_2$, applying T on $\vec{u}$ will yield an output of a vector direct sum $T(\vec{u}) = \vec{v} = \vec{v}_1 + \vec{v}_2$ in $\mathcal{V}_1 \oplus \mathcal{V}_2$ as well, where $\vec{v}_1 = T_1(\vec{u}_1) = T|_{U_1}(\vec{u}_1) \in \mathcal{V}_1$ and $\vec{v}_2 = T_2(\vec{u}_2) = T|_{U_2}(\vec{u}_2) \in \mathcal{V}_2$. The matrix direct sum is then the matrix representation of $T = T_1 \oplus T_2$ with respect to the direct sum basis 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 $\beta_1 = \{\vec{w}^{(1)}, \vec{w}^{(2)}, \ldots, \vec{w}^{(r)}\}$ and $\beta_2 = \{\vec{w}^{(r+1)}, \vec{w}^{(r+2)}, \ldots, \vec{w}^{(n)}\}$, and $\mathcal{V}_1$ and $\mathcal{V}_2$ has a basis $\eta_1 = \{\vec{y}^{(1)}, \vec{y}^{(2)}, \ldots, \vec{y}^{(s)}\}$ and $\eta_2 = \{\vec{y}^{(s+1)}, \vec{y}^{(s+2)}, \ldots, \vec{y}^{(m)}\}$, where r, n, s, m are some integers, then $T = T_1 \oplus T_2$ will have a *block diagonal* matrix representation of

$$[T]^{\eta_1+\eta_2}_{\beta_1+\beta_2} \equiv \begin{bmatrix} ([T_1]^{\eta_1}_{\beta_1})_{s\times r} & [\mathbf{0}]_{s\times(n-r)} \\ [\mathbf{0}]_{(m-s)\times r} & ([T_2]^{\eta_2}_{\beta_2})_{(m-s)\times(n-r)} \end{bmatrix}^{\eta_1+\eta_2}_{\beta_1+\beta_2} \tag{8.21}$$

with respect to the $\beta_1 \cup \beta_2$ and $\eta_1 \cup \eta_2$ bases. 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

$$\begin{aligned} T(\vec{u}) &= [T]^{\eta_1+\eta_2}_{\beta_1+\beta_2}[\vec{u}]_{\beta_1+\beta_2} \\ &= \begin{bmatrix} ([T_1]^{\eta_1}_{\beta_1})_{s\times r} & [\mathbf{0}]_{s\times(n-r)} \\ [\mathbf{0}]_{(m-s)\times r} & ([T_2]^{\eta_2}_{\beta_2})_{(m-s)\times(n-r)} \end{bmatrix}^{\eta_1+\eta_2}_{\beta_1+\beta_2} \begin{bmatrix} ([\vec{u}_1]_{\beta_1})_r \\ ([\vec{u}_2]_{\beta_2})_{n-r} \end{bmatrix}_{\beta_1+\beta_2} \end{aligned}$$

---

[7] $[T(\vec{w}_1)]_\eta = [T(\vec{u}_1 + \vec{u}_2)]_\eta = [T]^\eta_\beta(1,1,0)^T_\beta = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}^\eta_\beta \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}_\beta = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}_\eta$ and this will be the first column of $[T|_W]^\eta_{\beta_W}$. The second column is derived similarly by evaluating $[T(\vec{w}_2)]_\eta$.

$$
\begin{aligned}
&= \begin{bmatrix} ([T_1]_{\beta_1}^{\eta_1})_{s\times r}([\vec{u}_1]_{\beta_1})_r + [\mathbf{0}]_{s\times(n-r)}([\vec{u}_2]_{\beta_2})_{n-r} \\ [\mathbf{0}]_{(m-s)\times r}([\vec{u}_1]_{\beta_1})_r + ([T_2]_{\beta_2}^{\eta_2})_{(m-s)\times(n-r)}([\vec{u}_2]_{\beta_2})_{n-r} \end{bmatrix}_{\eta_1+\eta_2} \\
&= \begin{bmatrix} ([T_1]_{\beta_1}^{\eta_1}[\vec{u}_1]_{\beta_1})_s \\ ([T_2]_{\beta_2}^{\eta_2}[\vec{u}_2]_{\beta_2})_{m-s} \end{bmatrix}_{\eta_1+\eta_2} \\
&\equiv T_1(\vec{u}_1) \oplus T_2(\vec{u}_2)
\end{aligned}
$$

where the image is a direct sum composed of $T_1(\vec{u}_1) \equiv [T_1]_{\beta_1}^{\eta_1}[\vec{u}_1]_{\beta_1} \in \mathcal{V}_1$ and $T_2(\vec{u}_2) \equiv [T_2]_{\beta_2}^{\eta_2}[\vec{u}_2]_{\beta_2} \in \mathcal{V}_2$ 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

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

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 $T_1 : \mathcal{U}_1 \to \mathcal{V}_1$ and $T_2 : \mathcal{U}_2 \to \mathcal{V}_2$, 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 $T_1 \oplus T_2$ 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 $A_1, A_2, A_3, \ldots, A_{n-1}, A_n$ are defined recursively just like a vector direct sum as

$$
\begin{aligned}
&A_1 \oplus A_2 \oplus A_3 \oplus \cdots \oplus A_{n-1} \oplus A_n \\
&= (\cdots((A_1 \oplus A_2) \oplus A_3) \oplus \cdots \oplus A_{n-1}) \oplus A_n
\end{aligned}
$$

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 $T : \mathcal{U} \to \mathcal{V}$ that has a matrix representation of

$$[T]_\beta^\eta = \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix}$$

with respect to some bases $\beta = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \vec{u}^{(4)}\}$, $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}\}$, show that it can turn into an apparent matrix direct sum if the coordinate systems are changed according to $\beta' = \{\vec{u}^{(1)\prime}, \vec{u}^{(2)\prime}, \vec{u}^{(3)\prime}, \vec{u}^{(4)\prime}\}$, $\eta' = \{\vec{v}^{(1)\prime}, \vec{v}^{(2)\prime}, \vec{v}^{(3)\prime}\}$, where

$$\begin{aligned} \vec{u}^{(1)\prime} &= \vec{u}^{(1)} & \qquad \vec{v}^{(1)\prime} &= \vec{v}^{(1)} + \vec{v}^{(2)} \\ \vec{u}^{(2)\prime} &= \vec{u}^{(3)} & \qquad \vec{v}^{(2)\prime} &= -\vec{v}^{(2)} + \vec{v}^{(3)} \\ \vec{u}^{(3)\prime} &= \vec{u}^{(1)} + \vec{u}^{(2)} & \qquad \vec{v}^{(3)\prime} &= \vec{v}^{(1)} - \vec{v}^{(3)} \\ \vec{u}^{(4)\prime} &= \vec{u}^{(1)} + \vec{u}^{(4)} \end{aligned}$$

Solution. The change of coordinate matrices suggested by Properties 7.2.1 are

$$P_{\beta'}^{\beta} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad Q_{\eta'}^{\eta} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

and the new matrix representation of T is

$$[T]_{\beta'}^{\eta'} = (Q_{\eta'}^{\eta})^{-1} [T]_\beta^\eta P_{\beta'}^{\beta}$$

$$
\begin{aligned}
&= \left(\begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 & -2 \\ 0 & 0 & 1 & 0 \\ 1 & -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}
\end{aligned}
$$

where it can be seen that

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

□

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 θ, ϕ

$$\begin{aligned} e^{i\theta} &= \cos\theta + i\sin\theta \\ e^{i\phi} &= \cos\phi + i\sin\phi \\ e^{i(\theta+\phi)} &= \cos(\theta+\phi) + i\sin(\theta+\phi) \end{aligned}$$

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 De Moivre's Formula, prove the double angle formula shown below.

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta$$

$$\sin(2\theta) = 2\sin\theta\cos\theta$$

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,

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

compute AB^*, and verify $(AB^*)^* = BA^*$.

Exercise 8.5 For the matrix

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

find its determinant and inverse.

Exercise 8.6 Prove the formulae in Properties 8.3.4, by noting that

$$\begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} = \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -C & I_q \end{bmatrix}$$

and

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix}$$

Exercise 8.7 Write down the direct sum of the following three matrices.

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

$$B = \begin{bmatrix} 1 \end{bmatrix}$$

Exercise 8.8 Show that given two bases $\beta = \{\cos x, \sin x, 1, x, x^2\}$ and $\eta = \{\cos x, \sin x, 1, x\}$ which generate vector spaces $\mathcal{U}$ and $\mathcal{V}$ respectively, the differentiation operator $T(f(x)) = f'(x) : \mathcal{U} \to \mathcal{V}$ has a 2×2 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 that 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, e.g. chains of chemical reactions in the atmosphere. Moreover, many Earth Science applications suggested in the later chapters also depend heavily on the concepts of eigenvalues and eigenvectors.

9.1. Eigenvalues and Eigenvectors of a Square Matrix

9.1.1. Definition of Eigenvalues and Eigenvectors

Consider a linear operator/endomorphism $T : \mathcal{V} \to \mathcal{V}$, an interesting question is if a vector $\vec{v} \in \mathcal{V}$ under this mapping will remain stationary in direction such

that the image $T(\vec{v}) = \lambda\vec{v}$ 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 λ is the corresponding ***eigenvalue***. Since a linear operator is a mapping between the same 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 $T : \mathcal{V} \to \mathcal{V}$, we call λ and $\vec{v}_\lambda$ its eigenvalue and eigenvector if

$$T(\vec{v}_\lambda) = \lambda\vec{v}_\lambda \tag{9.1}$$

Similarly, given an $n \times n$ square matrix A, λ and $\vec{v}_\lambda$ will be an eigenvalue and eigenvector for it when

$$A\vec{v}_\lambda = \lambda\vec{v}_\lambda \tag{9.2}$$

This is a special case in which a vector space $\mathcal{V}$ is finite-dimensional, $\dim(\mathcal{V}) = n$, and $A = [T]_\beta$ is just the matrix representation of T with respect to some basis β.

Notice that there can be more than one eigenvalue and eigenvector. 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.

► 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 $\lambda = 2$.[1]

Figure 9.1.: *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 that 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 (non-zero) eigenvectors. 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 eigenvector for the same eigenvalue, and the dimension of the 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 λ. 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 for eigenvector rescaling.

Figure 9.2.: *Schematics of eigenvalues that represent: 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 Equation (9.2) in Definition 9.1.1 relating eigenvalues and eigenvectors to obtain

$$\begin{aligned} A\vec{v}_\lambda &= \lambda\vec{v}_\lambda \\ A\vec{v}_\lambda &= \lambda I\vec{v}_\lambda && (I\vec{v} = \vec{v} \text{ for any } \vec{v}) \\ (A - \lambda I)\vec{v}_\lambda &= \mathbf{0} \end{aligned} \tag{9.3}$$

[1] $\begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix}\left(s\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = s\begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = s\begin{bmatrix} 2 \\ 4 \end{bmatrix} = 2\left(s\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$. 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 $s\vec{v}_\lambda$ are also an eigenvector.

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 $\det(B) = \det(A - \lambda I) = 0$, as implied by Theorem 3.1.2. The relationship

$$p(\lambda) = \det(A - \lambda I) = 0 \tag{9.4}$$

is called the ***characteristic equation***. The roots for λ of the *characteristic polynomial* $p(\lambda)$ are then the desired eigenvalues.

Definition 9.1.2 (Characteristic Equation)**.** The characteristic equation for a linear operator $T : \mathcal{V} \to \mathcal{V}$ over a vector space $\mathcal{V}$ with $\dim(\mathcal{V}) = n$ (or an $n \times n$ square matrix A) is

$$\det\big([T]_\beta - \lambda I\big) = 0 \qquad \text{or} \qquad \det(A - \lambda I) = 0$$

where β 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 λ 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.[2]

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

For each eigenvalue, there corresponds at least one linearly independent eigenvector by definition. The required eigenvector(s) that generates the eigenspace for a particular eigenvalue λ_j can be found as the general solution to the matrix equation $(A - \lambda_j I)\vec{v} = \mathbf{0}$ (see (9.3)), 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 λ_j is hence the dimension of the null space (Definition 6.3.8) $\mathcal{N}(A - \lambda_j I)$, known as the ***geometric multiplicity***. A closely related quantity is the ***algebraic multiplicity***, which is the number of times where

λ_j appears as the root to the characteristic polynomial. It can be shown that the geometric multiplicity of λ_j is capped by its algebraic multiplicity.

Theorem 9.1.3. For an eigenvalue λ_j to a square matrix A, its geometric multiplicity is equal to $\dim(\mathcal{N}(A - \lambda_j I))$, the dimension of the null space of $A - \lambda_j I$. Meanwhile, its algebraic multiplicity is the power k_j of the factor $(\lambda_j - \lambda)^{k_j}$ present 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 λ_j.

We defer the proof showing that geometric multiplicity is always smaller than or equal to algebraic multiplicity until we introduce diagonalization later in this chapter.

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 $A - \lambda I = \mathbf{0}$:

$$\left[\begin{array}{cc|c} 1-1 & -1 & 0 \\ 0 & 1-1 & 0 \end{array}\right] = \left[\begin{array}{cc|c} 0 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

[2]They are $\lambda = 1, 2, 3$.

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. □

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)

$$\begin{aligned} \det(A - \lambda I) &= \begin{vmatrix} 1-\lambda & 3 & 1 \\ 0 & 1-\lambda & 0 \\ 1 & 0 & 2-\lambda \end{vmatrix} \\ &= (1-\lambda)(1-\lambda)(2-\lambda) - (1)(1-\lambda)(1) \\ &= (1-\lambda)((2-3\lambda+\lambda^2) - 1) \\ &= (1-\lambda)(1-3\lambda+\lambda^2) \end{aligned}$$

The roots and thus eigenvalues are $\lambda = 1$ and

$$\begin{aligned} \lambda &= \frac{-(-3) \pm \sqrt{(-3)^2 - 4(1)(1)}}{2} \\ &= \frac{3}{2} \pm \frac{\sqrt{5}}{2} \end{aligned}$$

Particularly, for the eigenvalue $\lambda = \frac{3}{2} + \frac{\sqrt{5}}{2}$, the eigenvector is inferred from the homogeneous system $A - \lambda I = \mathbf{0}$

$$\left[\begin{array}{ccc|c} -\frac{1}{2} - \frac{\sqrt{5}}{2} & 3 & 1 & 0 \\ 0 & -\frac{1}{2} - \frac{\sqrt{5}}{2} & 0 & 0 \\ 1 & 0 & \frac{1}{2} - \frac{\sqrt{5}}{2} & 0 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & \frac{1}{2}-\frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{2}-\frac{\sqrt{5}}{2} & 3 & 1 & 0 \end{array}\right] \qquad R_1 \leftrightarrow R_3$$

$$\left(\frac{1}{-\frac{1}{2}-\frac{\sqrt{5}}{2}}\right) R_2 \rightarrow R_2$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & \frac{1}{2}-\frac{\sqrt{5}}{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{array}\right] \qquad R_3 - \left(-\frac{1}{2}-\frac{\sqrt{5}}{2}\right) R_1 \rightarrow R_3$$

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

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. □

► Short Exercise: Find the eigenvectors for the other remaining eigenvalues.[3]

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.

$$\begin{aligned} p(\lambda) &= \det(A - \lambda I) \\ &= \begin{vmatrix} 1-\lambda & 1 & -1 \\ -1 & 3-\lambda & -1 \\ 1 & -1 & 3-\lambda \end{vmatrix} \\ &= [(1-\lambda)(3-\lambda)(3-\lambda) + (1)(-1)(1) + (-1)(-1)(-1)] \\ &\quad - [(-1)(3-\lambda)(1) + (1-\lambda)(-1)(-1) + (1)(-1)(3-\lambda)] \end{aligned}$$

$$
\begin{aligned}
&= -\lambda^3 + 7\lambda^2 - 16\lambda + 12 \\
&= (2-\lambda)^2(3-\lambda)
\end{aligned}
$$

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

$$
\begin{aligned}
\left[\begin{array}{ccc|c} 1-2 & 1 & -1 & 0 \\ -1 & 3-2 & -1 & 0 \\ 1 & -1 & 3-2 & 0 \end{array}\right] &= \left[\begin{array}{ccc|c} -1 & 1 & -1 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 0 \end{array}\right] \\
&\to \left[\begin{array}{ccc|c} 1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 0 \\ -1 & 1 & -1 & 0 \end{array}\right] && R_1 \leftrightarrow R_3 \\
&\to \left[\begin{array}{ccc|c} 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] && \begin{array}{l} R_2 + R_1 \to R_2 \\ R_3 + R_1 \to R_3 \end{array}
\end{aligned}
$$

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 is equal to the algebraic multiplicity. On the other hand, for $\lambda = 3$, we have

$$
\left[\begin{array}{ccc|c} -2 & 1 & -1 & 0 \\ -1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 1 & -1 & 0 \end{array}\right] \qquad R_1 \leftrightarrow R_3
$$

[3]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.

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

$$\rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 \end{array}\right] \qquad -R_2 \rightarrow R_2$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_3 + R_2 \rightarrow R_3$$

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 $\lambda = 3$. □

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

$$p(\lambda) = \det(A - \lambda I) = \begin{vmatrix} -\lambda & -1 \\ 1 & -\lambda \end{vmatrix} = 0$$

$$(-\lambda)^2 - (1)(-1) = \lambda^2 + 1 = 0$$

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 λ such that $A\vec{v}_\lambda = \lambda\vec{v}_\lambda$, as the scalar multiplication on R.H.S. only allows λ 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°. 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 solving

$$\left[\begin{array}{cc|c} -i & -1 & 0 \\ 1 & -i & 0 \end{array}\right]$$

We can use a trick to immediately ignore the second row since the algebraic (and also geometric) 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. □

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 $\vec{v}^{(2)}$ be two eigenvectors corresponding to two different eigenvalues λ_1 and λ_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

$$\begin{aligned} A(c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)}) &= \mathbf{0} \\ c_1(A\vec{v}^{(1)}) + c_2(A\vec{v}^{(2)}) &= \mathbf{0} \\ c_1\lambda_1\vec{v}^{(1)} + c_2\lambda_2\vec{v}^{(2)} &= \mathbf{0} \qquad \text{(Definition 9.1.1)} \end{aligned}$$

But multiplying the same equation of $c_1\vec{v}^{(1)} + c_2\vec{v}^{(2)} = \mathbf{0}$ by λ_1 gives $c_1\lambda_1\vec{v}^{(1)} + c_2\lambda_1\vec{v}^{(2)} = \mathbf{0}$ instead. Subtracting this from the other equation leads to

$$(c_1\lambda_1\vec{v}^{(1)} + c_2\lambda_2\vec{v}^{(2)}) - (c_1\lambda_1\vec{v}^{(1)} + c_2\lambda_1\vec{v}^{(2)}) = \mathbf{0} - \mathbf{0}$$

$$c_2(\lambda_2 - \lambda_1)\vec{v}^{(2)} = \mathbf{0}$$

Since λ_1 and λ_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. □

Properties 9.1.5. For a square matrix A,

1. A^T shares the same eigenvalues, but for each eigenvalue, the eigenvector(s) is/are not guaranteed to be the same. However, the dimensions (geometric multiplicity) of the two eigenspaces will be equal.

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

$$\begin{aligned}
\det\left(A^T - \lambda I\right) &= \det\left(A^T - \lambda I^T\right) \\
&= \det\left((A - \lambda I)^T\right) && \text{(Properties 2.1.4)} \\
&= \det(A - \lambda I) && \text{(Properties 2.3.9)}
\end{aligned}$$

So the characteristic polynomials of A and A^T coincide, and the eigenvalues (if exist) will be the same. The geometric multiplicity can be shown to be equal by noting that transpose does not change the rank of a matrix, and thus $\operatorname{rank}(A^T - \lambda I) = \operatorname{rank}((A - \lambda I)^T) = \operatorname{rank}(A - \lambda I)$. Subsequently, by using the Rank-nullity Theorem 6.3.9 twice, we have

$$\begin{aligned}
\dim(\mathcal{N}(A^T - \lambda_j I)) &= n - \operatorname{rank}(A^T - \lambda I) \\
&= n - \operatorname{rank}(A - \lambda I) = \dim(\mathcal{N}(A - \lambda_j I))
\end{aligned}$$

For the second statement, multiplying both sides of (9.2) by A^{-1} to obtain

$$A\vec{v}_\lambda = \lambda\vec{v}_\lambda$$

$$A^{-1}A\vec{v}_\lambda = I\vec{v}_\lambda = \vec{v}_\lambda = \lambda A^{-1}\vec{v}_\lambda \qquad \text{(Definition 2.2.1)}$$
$$\frac{1}{\lambda}\vec{v}_\lambda = A^{-1}\vec{v}_\lambda$$

which explicitly shows that A^{-1} has $\frac{1}{\lambda}$ and $\vec{v}_\lambda$ as its eigenvalue and eigenvector. □

Theorem 9.1.6. A square matrix is singular if and only if it has zero as one of its eigenvalues.

Proof. Exercise.[4] Note that this is also equivalent to A having a non-trivial null space of dimension ≥ 1 due to Theorem 9.1.3. □

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 $T : \mathcal{V} \to \mathcal{V}$, 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 $T : \mathcal{V} \to \mathcal{V}$, 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 is T-invariant.

[4] 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 imply A is singular. The converse is established by the same argument in the reverse direction.

Properties 9.1.8. For a linear endomorphism $T : \mathcal{V} \to \mathcal{V}$, the eigenspace $\mathcal{E}_{J_i} \subseteq \mathcal{V}$ associated with a specific eigenvalue λ_{J_i} is a T-invariant subspace.

Proof. Exercise.[5] □

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 $T_{J_i} : \mathcal{W}_{J_i} \to \mathcal{W}_{J_i}$ (when they happen to be endomorphisms themselves) 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 λ 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

$$p(\lambda) = \det(A - \lambda I) = (-1)^n \sum_{k=0}^{n} c_k \lambda^k \tag{9.5}$$

where $c_n = 1$ is fixed, then we have

$$p(A) = (-1)^n \sum_{k=0}^{n} c_k A^k = [\mathbf{0}] \tag{9.6}$$

as an $n \times n$ zero matrix.

[5] 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)

One may be tempted to substitute $\lambda = A$ into $\det(A - \lambda I)$ to prove the Cayley-Hamilton Theorem. However, since λ (and $\det(A - \lambda I)$) is a scalar but A (and $[\mathbf{0}]$) is a matrix, it is not a rigorous proof. A correct proof requires advanced knowledge, which is presented in Appendix B.1. 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 λ^k terms in the characteristic polynomial with A^k (Notice that the constant term c_0 becomes $c_0 I$), 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

$$A^2 - 7A + 6I = (A^2 - 7A + 6I) - (A^2 - 6A + 8I)$$

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

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

$$P^{-1}AP = D \tag{9.7}$$

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 A. 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 to

diagonalize A, is formed by combining all linearly independent eigenvectors of A column by column. This is only possible if the number 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 $\mathbb{C}$)[6] used in the vector space, and, for every eigenvalue, its geometric multiplicity is equal to the algebraic multiplicity.[7] 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 $\mathbb{R}^n$ (see part (b) of Properties 6.2.7) when the underlying scalar of vector space is taken to be $\mathbb{R}$ [a basis for $\mathbb{C}^n$ if working over $\mathbb{C}$]. The equivalent conditions are that the characteristic polynomial of A splits over $\mathbb{R}$ [$\mathbb{C}$] and the geometric multiplicities of all distinct eigenvalues are equal to the respective algebraic multiplicities. Then, the columns of P will consist of those eigenvectors $\vec{v}_\lambda^{(j)}$, $j = 1, 2, \ldots, n$. The diagonal entries of $P^{-1}AP = D$, are subsequently each of the eigenvalues λ_j (counting repeated ones) corresponding to the eigenvector $\vec{v}_\lambda^{(j)}$ in the same column position of P.

[6] 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 λ_j are complex [real] constants. Also, there is the *Fundamental Theorem of Algebra* which states that every polynomial splits over $\mathbb{C}$ so the first part of the condition is automatically satisfied for the complex case.

[7] Note that the sum of geometric multiplicities for all eigenvalues is 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 the geometric multiplicity of any eigenvalue is strictly equal to the algebraic multiplicity (Theorem 9.1.3 again) while $p(\lambda)$ splits. (Properties 9.1.4 ensures linear independence over eigenspaces of different eigenvalues.)

Proof. We will explicitly construct the desired form. Consider two matrix products AP and PD, where $P = [\vec{v}_\lambda^{(1)}|\vec{v}_\lambda^{(2)}|\cdots|\vec{v}_\lambda^{(n)}]$ contains 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 9 of Chapter 6. Also

$$\begin{aligned} PD &= [\vec{v}_\lambda^{(1)}|\vec{v}_\lambda^{(2)}|\cdots|\vec{v}_\lambda^{(n)}]\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \\ &= [\lambda_1\vec{v}_\lambda^{(1)}|\lambda_2\vec{v}_\lambda^{(2)}|\cdots|\lambda_n\vec{v}_\lambda^{(n)}] \end{aligned}$$

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 by requirement, and thus $P^{-1}AP = D$. □

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 straightforward by the use of Properties 9.2.2. Below is 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

$$\begin{aligned} \left[\begin{array}{ccc|c} 1-(-1) & -2 & 1 & 0 \\ -\frac{2}{7} & 1-(-1) & -1 & 0 \\ -\frac{4}{7} & -2 & -(-1) & 0 \end{array}\right] &= \left[\begin{array}{ccc|c} 2 & -2 & 1 & 0 \\ -\frac{2}{7} & 2 & -1 & 0 \\ -\frac{4}{7} & -2 & 1 & 0 \end{array}\right] \\ &\to \left[\begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ -\frac{2}{7} & 2 & -1 & 0 \\ -\frac{4}{7} & -2 & 1 & 0 \end{array}\right] && \frac{1}{2}R_1 \to R_1 \\ &\to \left[\begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ 0 & \frac{12}{7} & -\frac{6}{7} & 0 \\ 0 & -\frac{18}{7} & \frac{9}{7} & 0 \end{array}\right] && \begin{array}{r} R_2 + \frac{2}{7}R_1 \to R_2 \\ R_3 + \frac{4}{7}R_1 \to R_3 \end{array} \\ &\to \left[\begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & -\frac{18}{7} & \frac{9}{7} & 0 \end{array}\right] && \frac{7}{12}R_2 \to R_2 \\ &\to \left[\begin{array}{ccc|c} 1 & -1 & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] && R_3 + \frac{18}{7}R_2 \to R_3 \end{aligned}$$

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

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 it 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.[8] We

[8]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.

therefore say that diagonalization is *unique up to a permutation*. □

► Short Exercise: Reverse the diagonalization to recover the original matrix.[9]

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

$$P = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \text{ and}$$

$$\begin{aligned} D &= P^{-1}AP \\ &= \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & -1 \\ -1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & -3 \\ 2 & 0 & -3 \\ 0 & 2 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \end{aligned}$$

The first and second eigenvector can be interchanged without modifying the form of the diagonalized matrix as they correspond to the same eigenvalue of $\lambda = 2$. □

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 $T : \mathcal{V} \to \mathcal{V}$ (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 λ_{J_i}[10] 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 $T_j(\vec{v}_\lambda^{(j)}) : \mathcal{E}_j \to \mathcal{E}_j = T|_{\mathcal{E}_j}(\vec{v}_\lambda^{(j)}) = T(\vec{v}_\lambda^{(j)}) = \lambda_j \vec{v}_\lambda^{(j)} = \lambda_j(\mathrm{id}(\vec{v}_\lambda^{(j)}))$. In matrix representation, it becomes $A = \bigoplus_{j=1}^n [\lambda_j]$ where each of the $[\lambda_j]$ is a singleton (1×1) 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 $T_{J_i} = \lambda_{J_i}(\mathrm{id}[\,]) : \mathcal{E}_{J_i} \to \mathcal{E}_{J_i}$ 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 λ_{J_i} in the characteristic polynomial (here 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 $T : \mathcal{V} \to \mathcal{V}$ becomes a matrix direct

[9] Simply compute PDP^{-1} which should be equal to A.

[10] J_i refers to the set of indices those are associated to a common eigenvalue, denoted by λ_{J_i}.

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 and algebraic multiplicities;
4. same eigenspace dimension (geometric multiplicity) for any fixed eigenvalue (but the eigenvectors are probably different);
5. same trace, a.k.a. the sum of diagonal entries; and
6. same rank.

Proof. We will briefly prove (1) here. Similar matrices are related by $A' = P^{-1}AP$ for any invertible matrix P. As a result, $\det(A') = \det\left(P^{-1}AP\right) = \det\left(P^{-1}\right)\det(A)\det(P) = \frac{1}{\det(P)}\det(A)\det(P) = \det(A)$ due to Properties 2.3.9. ($\det(P) \neq 0$ by Properties 2.3.8.) □

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.[11]

With all of these, we can finally now prove Theorem 9.1.3. Denote the geometric multiplicity of A by n_{J_i} for a fixed eigenvalue λ_{J_i}. Then the n_{J_i}

linearly independent eigenvectors $\vec{v}^{(1)}_{\lambda_{J_i}}, \vec{v}^{(2)}_{\lambda_{J_i}}, \ldots, \vec{v}^{(n_{J_i})}_{\lambda_{J_i}}$ accordingly will 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 become (Properties 7.2.2 again)

$$A' = P^{-1}AP$$

where

$$P = \left[\vec{v}^{(1)}_{\lambda_{J_i}} | \vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | \vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | \text{other vectors used to fill the basis for } \mathbb{R}^n\right]$$

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

$$A' = \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n - n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix}$$

[12] where the top left block represents a rescaling of the eigenvectors (now as

[11] $\det(A' - \lambda I) = \det\left(P^{-1}AP - \lambda I\right) = \det\left(P^{-1}AP - \lambda P^{-1}P\right) = \det\left(P^{-1}AP - \lambda P^{-1}IP\right) = \det\left(P^{-1}(A - \lambda I)P\right) = \det\left(P^{-1}\right)\det(A - \lambda I)\det(P) = \det(A - \lambda I)$ (Definition 2.2.1, Properties 2.1.2 and 2.3.9). (3) follows from (2) immediately.

[12] 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.

$$\begin{aligned} PA' &= \left[\vec{v}^{(1)}_{\lambda_{J_i}} | \vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | \vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | *\right] \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n - n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix} \\ &= \left[\lambda_{J_i}\vec{v}^{(1)}_{\lambda_{J_i}} | \lambda_{J_i}\vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | \lambda_{J_i}\vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | *\right] \end{aligned}$$

and

$$\begin{aligned} AP &= A\left[\vec{v}^{(1)}_{\lambda_{J_i}} | \vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | \vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | *\right] \\ &= \left[A\vec{v}^{(1)}_{\lambda_{J_i}} | A\vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | A\vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | *\right] \\ &= \left[\lambda_{J_i}\vec{v}^{(1)}_{\lambda_{J_i}} | \lambda_{J_i}\vec{v}^{(2)}_{\lambda_{J_i}} | \ldots | \lambda_{J_i}\vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | *\right] \end{aligned}$$

the last step is by the definition of eigenvectors (Definition 9.1.1).

a part of the new basis) by the eigenvalue, and the $*$ parts can be anything (notice the zero block at the bottom left, since $\mathcal{E}_{J_i}$ is an invariant subspace). 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

$$p(\lambda) = (\lambda_{J_i} - \lambda)^{n_{J_i}} \det\left(*_{(n-n_{J_i}) \times (n-n_{J_i})} - \lambda I_{(n-n_{J_i})} \right)$$

which shows that the characteristic polynomial has λ_{J_i} as its roots occurring for at least n_{J_i} times (the determinant that follows 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 λ_{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}.

Diagonalization also enables the efficient computation of matrix powers, as demonstrated in the following example.

Example 9.2.3. Show that for

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

its successive power is

$$A^n = \begin{bmatrix} \frac{1}{2}(2^n + 1) & \frac{1}{2}(2^n - 1) \\ \frac{1}{2}(2^n - 1) & \frac{1}{2}(2^n + 1) \end{bmatrix}$$

Solution. We leave it to the readers to show that the diagonalization of A is done by

$$D = P^{-1} A P$$

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

Hence

$$\begin{aligned}
A^n &= (PDP^{-1})^n \\
&= (PDP^{-1})(PDP^{-1}) \cdots (PDP^{-1})(PDP^{-1}) \\
&= PD(P^{-1}P)D \cdots D(P^{-1}P)P^{-1} \\
&= PDID \cdots DIP^{-1} && \text{(Definition 2.2.1)} \\
&= PD^nP^{-1} \\
&= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2^n & 0 \\ 0 & 1^n \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2^n & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{2}(2^n+1) & \frac{1}{2}(2^n-1) \\ \frac{1}{2}(2^n-1) & \frac{1}{2}(2^n+1) \end{bmatrix}
\end{aligned}$$

where the successive powers of a diagonal matrix can be simply computed over each of the diagonal entries. □

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 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 downside is that D (as well as the P matrix that is made up of the eigenvectors as its columns) is comprised of complex entries, despite A being a real matrix. There is, indeed, another method that uses the

property of complex numbers introduced in the last chapter, which 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 the 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 $\overline{z_0} = a - bi$ 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

$$\begin{aligned}
\overline{p(z_0)} &= \overline{\sum_{k=0}^{n} c_k z_0^k} = \overline{0} \\
&= \sum_{k=0}^{n} \overline{c_k}\,\overline{z_0^k} = 0 \\
p(\overline{z_0}) &= \sum_{k=0}^{n} c_k \overline{z_0}^k = 0
\end{aligned}$$

$\overline{c_k} = c_k$ 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 λ_0 is a complex eigenvalue for a real matrix A with an eigenvector of $\vec{v}_{\lambda_0}$, then A also has $\overline{\lambda_0}$ and $\overline{\vec{v}_{\lambda_0}}$ 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. □

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 $\lambda_0 = \mathrm{Re}\{\lambda_0\} + i\,\mathrm{Im}\{\lambda_0\}$, the corresponding eigenvector $\vec{v}_{\lambda_0} = \mathrm{Re}\{\vec{v}_{\lambda_0}\} + i\,\mathrm{Im}\{\vec{v}_{\lambda_0}\}$, as well as their complex conjugates, by replacing the corresponding columns in

$$P = [\cdots|\vec{v_{\lambda_0}}|\overline{\vec{v_{\lambda_0}}}|\cdots] \qquad D = \begin{bmatrix} \ddots & 0 & 0 & \\ & \lambda_0 & 0 & \\ & 0 & \overline{\lambda_0} & \\ & 0 & 0 & \ddots \end{bmatrix}$$

by

$$P = [\cdots | \operatorname{Re}\{\vec{v}_{\lambda_0}\} | \operatorname{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \quad D = \begin{bmatrix} \ddots & 0 & 0 & \\ & \operatorname{Re}\{\lambda_0\} & \operatorname{Im}\{\lambda_0\} & \\ & -\operatorname{Im}\{\lambda_0\} & \operatorname{Re}\{\lambda_0\} & \\ & 0 & 0 & \ddots \end{bmatrix} \tag{9.8}$$

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 $\operatorname{Im}\{\lambda_0\} = b$. It is not hard to see that

$$\begin{aligned} PD &= [\cdots | \operatorname{Re}\{\vec{v}_{\lambda_0}\} | \operatorname{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \begin{bmatrix} \ddots & 0 & 0 & \\ & a & b & \\ & -b & a & \\ & 0 & 0 & \ddots \end{bmatrix} \\ &= [\cdots | a \operatorname{Re}\{\vec{v}_{\lambda_0}\} - b \operatorname{Im}\{\vec{v}_{\lambda_0}\} | b \operatorname{Re}\{\lambda_0\} + a \operatorname{Im}\{\vec{v}_{\lambda_0}\} | \cdots] \end{aligned}$$

(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

$$\begin{aligned} A\vec{v}_{\lambda_0} &= \lambda_0 \vec{v}_{\lambda_0} \\ A(\operatorname{Re}\{\vec{v}_{\lambda_0}\} + i \operatorname{Im}\{\vec{v}_{\lambda_0}\}) &= (a + bi)(\operatorname{Re}\{\vec{v}_{\lambda_0}\} + i \operatorname{Im}\{\vec{v}_{\lambda_0}\}) \\ A \operatorname{Re}\{\vec{v}_{\lambda_0}\} + iA \operatorname{Im}\{\vec{v}_{\lambda_0}\} &= \begin{aligned} &(a \operatorname{Re}\{\vec{v}_{\lambda_0}\} - b \operatorname{Im}\{\vec{v}_{\lambda_0}\}) \\ &+ i(b \operatorname{Re}\{\vec{v}_{\lambda_0}\} + a \operatorname{Im}\{\vec{v}_{\lambda_0}\}) \end{aligned} \end{aligned}$$

where $\operatorname{Re}\{\vec{v}_{\lambda_0}\}$ and $\operatorname{Im}\{\vec{v}_{\lambda_0}\}$ themselves are real vectors and A is a real matrix. By equating the real and imaginary parts, we have

$$A \operatorname{Re}\{\vec{v}_{\lambda_0}\} = a \operatorname{Re}\{\vec{v}_{\lambda_0}\} - b \operatorname{Im}\{\vec{v}_{\lambda_0}\} \tag{9.9a}$$

$$A \operatorname{Im}\{\vec{v}_{\lambda_0}\} = b \operatorname{Re}\{\vec{v}_{\lambda_0}\} + a \operatorname{Im}\{\vec{v}_{\lambda_0}\} \tag{9.9b}$$

Hence

$$\begin{aligned} AP &= A[\cdots|\operatorname{Re}\{\vec{v}_{\lambda_0}\}|\operatorname{Im}\{\vec{v}_{\lambda_0}\}|\cdots] \\ &= [\cdots|A\operatorname{Re}\{\vec{v}_{\lambda_0}\}|A\operatorname{Im}\{\vec{v}_{\lambda_0}\}|\cdots] \\ &= [\cdots|a\operatorname{Re}\{\vec{v}_{\lambda_0}\} - b\operatorname{Im}\{\vec{v}_{\lambda_0}\}|b\operatorname{Re}\{\lambda_0\} + a\operatorname{Im}\{\vec{v}_{\lambda_0}\}|\cdots] \\ &= PD \end{aligned}$$

The only caveat is to prove that P is invertible, or equivalently $\operatorname{Re}\{\vec{v}_{\lambda_0}\}$ and $\operatorname{Im}\{\vec{v}_{\lambda_0}\}$ are linearly independent (assuming there is no problem with other columns). We can prove that by assuming they are linearly dependent, so $\operatorname{Re}\{\vec{v}_{\lambda_0}\} = k\operatorname{Im}\{\vec{v}_{\lambda_0}\}$ for some *real* constant k, and plugging this into the expressions of $A\operatorname{Re}\{\vec{v}_{\lambda_0}\}$ and $A\operatorname{Im}\{\vec{v}_{\lambda_0}\}$ to derive a contradiction.[13] □

The block in the form

$$\begin{bmatrix} \ddots & 0 & 0 & \\ & a & b & \\ & -b & a & \\ & 0 & 0 & \ddots \end{bmatrix}$$

[13] Substitute $\operatorname{Re}\{\vec{v}_{\lambda_0}\} = k\operatorname{Im}\{\vec{v}_{\lambda_0}\}$ into Equation (9.9a)

$$\begin{aligned} A\operatorname{Re}\{\vec{v}_{\lambda_0}\} &= a\operatorname{Re}\{\vec{v}_{\lambda_0}\} - b\operatorname{Im}\{\vec{v}_{\lambda_0}\} \\ Ak\operatorname{Im}\{\vec{v}_{\lambda_0}\} &= ak\operatorname{Im}\{\vec{v}_{\lambda_0}\} - b\operatorname{Im}\{\vec{v}_{\lambda_0}\} = (ak-b)\operatorname{Im}\{\vec{v}_{\lambda_0}\} \end{aligned}$$

and multiplying Equation (9.9b) 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.

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 9.2.4. 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

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

The roots are $\lambda = 1$ and

$$\begin{aligned} \lambda &= \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(3)}}{2} \\ &= 1 \pm \sqrt{2}i \end{aligned}$$

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

$$\left[\begin{array}{ccc|c} \sqrt{2}i & 0 & 1 & 0 \\ 0 & \sqrt{2}i & 1 & 0 \\ 0 & -2 & \sqrt{2}i & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & -\frac{1}{\sqrt{2}}i & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}}i & 0 \\ 0 & -2 & \sqrt{2}i & 0 \end{array}\right] \qquad \begin{aligned} -\frac{1}{\sqrt{2}}iR_1 &\to R_1 \\ -\frac{1}{\sqrt{2}}iR_2 &\to R_2 \end{aligned}$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -\frac{1}{\sqrt{2}}i & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}}i & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_3 + 2R_1 \rightarrow R_3$$

So the eigenvector for $\lambda = 1 - \sqrt{2}i$ is $(i, i, \sqrt{2})^T$, and the eigenvector for $\lambda = 1 + \sqrt{2}i$ is $(-i, -i, \sqrt{2})^T$. Applying (9.8) in Properties 9.2.6, with $a = 1$, $b = \sqrt{2}$, $\mathrm{Re}\{\vec{v}_{\lambda_0}\} = (0, 0, \sqrt{2})^T$, $\mathrm{Im}\{\vec{v}_{\lambda_0}\} = (-1, -1, 0)^T$, 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.[14]

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

$$\frac{dy}{dx} + P(x)y = Q(x) \tag{9.10}$$

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 cross-product[15] 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 the 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

$$\frac{dy}{dx} = \alpha y \tag{9.11}$$

where α is a constant, the general solution is

$$y(x) = ce^{\alpha x} \tag{9.12}$$

with c as an integration constant to be determined.

A direct substitution can verify the solution given above.[16] A famous example

14

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+\sqrt{2}i & 0 \\ 0 & 0 & 1-\sqrt{2}i \end{bmatrix} = \begin{bmatrix} 1 & -i & i \\ 0 & -i & i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -i & i \\ 0 & -i & i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}$$

[15] This is not the vector cross product.

[16] L.H.S. $= \frac{d}{dx}(ce^{\alpha x}) = \alpha ce^{\alpha x} =$ R.H.S.

will be the first-order decay of radioactive isotopes whose concentration is N:

$$\frac{dN}{dt} = -kN$$

where $\alpha = -k$ 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 that all of the ODEs are 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

$$\begin{cases} dy_1/dx = \alpha_1^{(1)} y_1 + \alpha_2^{(1)} y_2 + \alpha_3^{(1)} y_3 + \cdots + \alpha_n^{(1)} y_n \\ dy_2/dx = \alpha_1^{(2)} y_1 + \alpha_2^{(2)} y_2 + \alpha_3^{(2)} y_3 + \cdots + \alpha_n^{(2)} y_n \\ dy_3/dx = \alpha_1^{(3)} y_1 + \alpha_2^{(3)} y_2 + \alpha_3^{(3)} y_3 + \cdots + \alpha_n^{(3)} y_n \\ \vdots \qquad\quad = \vdots \\ dy_n/dx = \alpha_1^{(n)} y_1 + \alpha_2^{(n)} y_2 + \alpha_3^{(n)} y_3 + \cdots + \alpha_n^{(n)} y_n \end{cases} \tag{9.13}$$

where $\alpha_j^{(i)}$ are the constant coefficients, or more compactly

$$\frac{dy_i}{dx} = \alpha_1^{(i)} y_1 + \alpha_2^{(i)} y_2 + \alpha_3^{(i)} y_3 + \cdots + \alpha_n^{(i)} y_n = \sum_{j=1}^{n} \alpha_j^{(i)} y_j \tag{9.14}$$

or in matrix notation

$$\mathbf{y}' = A\mathbf{y} \tag{9.15}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{y}' = \begin{bmatrix} dy_1/dx \\ dy_2/dx \\ dy_3/dx \\ \vdots \\ dy_n/dx \end{bmatrix} \tag{9.16a}$$

the superscript $'$ represents differentiation, and

$$A = \begin{bmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} & \alpha_3^{(1)} & \cdots & \alpha_n^{(1)} \\ \alpha_1^{(2)} & \alpha_2^{(2)} & \alpha_3^{(2)} & & \alpha_n^{(2)} \\ \alpha_1^{(3)} & \alpha_2^{(3)} & \alpha_3^{(3)} & & \alpha_n^{(3)} \\ \vdots & & & \ddots & \vdots \\ \alpha_1^{(n)} & \alpha_2^{(n)} & \alpha_3^{(n)} & \cdots & \alpha_n^{(n)} \end{bmatrix} \tag{9.16b}$$

is an $n \times n$ matrix holding the coefficients, $A_{ij} = \alpha_j^{(i)}$.

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 on R.H.S., where $j = 1, 2, 3, \ldots$, 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

$$\begin{cases} dz_1/dx = D_{11}z_1 \\ dz_2/dx = D_{22}z_2 \\ dz_3/dx = D_{33}z_3 \\ \vdots \qquad\quad = \vdots \\ dz_n/dx = D_{nn}z_n \end{cases} \tag{9.17}$$

which are all solvable on 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 (9.15)

$$\mathbf{y}' = A\mathbf{y}$$

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

$$\mathbf{y} = P\mathbf{z} \tag{9.18}$$

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

$$\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z} \tag{9.19}$$

and each of the z_i in $\mathbf{z}$, $i = 1, 2, 3, \ldots, n$ 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 it 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

$$\begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} z_1(0) \\ z_2(0) \\ z_3(0) \end{bmatrix}$$

$$\begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

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

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 & -5 & -4 \\ 0 & 1 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} -10e^x \\ 13e^{2x} \\ -12e^{4x} \end{bmatrix}$$

$$\Longrightarrow \begin{cases} y_1 = 20e^x - 65e^{2x} + 48e^{4x} \\ y_2 = 13e^{2x} - 12e^{4x} \\ y_3 = -10e^x + 39e^{2x} - 24e^{4x} \end{cases}$$

or in vector notation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = e^x \begin{bmatrix} 20 \\ 0 \\ -10 \end{bmatrix} + e^{2x} \begin{bmatrix} -65 \\ 13 \\ 39 \end{bmatrix} + e^{4x} \begin{bmatrix} 48 \\ -12 \\ -24 \end{bmatrix}$$

□

► 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.[17]

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 9.2.4.

Solution. From the previous work we know that the eigenvectors are $(1, 0, 0)^T$, $(i, i, \sqrt{2})^T$ and $(-i, -i, \sqrt{2})^T$ for $\lambda = 1, 1 - \sqrt{2}i, 1 + \sqrt{2}i$ 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

$$P = \begin{bmatrix} 1 & i & -i \\ 0 & i & -i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}$$

then we have

$$\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z}$$

with

$$\begin{aligned} D = P^{-1}AP &= \begin{bmatrix} 1 & i & -i \\ 0 & i & -i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & i & -i \\ 0 & i & -i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - \sqrt{2}i & 0 \\ 0 & 0 & 1 + \sqrt{2}i \end{bmatrix} \end{aligned}$$

[17]The corresponding integration constants are $c_1 = -1, c_2 = -6, c_3 = 8$. So $z_1 = -e^x$, $z_2 = -6e^{2x}$, $c_3 = 8e^{4x}$, and $y_1 = 2e^x + 30e^{2x} - 32e^{4x}$, $y_2 = -6e^{2x} + 8e^{4x}$, $y_3 = -e^x - 18e^{2x} + 16e^{4x}$.

The solution in Properties 9.3.2 is valid even when α is complex. Hence the solution for $\mathbf{z}$ will be

$$\begin{cases} z_1 = c_1 e^x \\ z_2 = c_2 e^{(1-\sqrt{2}i)x} = c_2 e^x e^{-\sqrt{2}ix} = c_2 e^x (\cos\left(\sqrt{2}x\right) - i\sin\left(\sqrt{2}x\right)) \\ z_3 = c_3 e^{(1+\sqrt{2}i)x} = c_3 e^x e^{\sqrt{2}ix} = c_3 e^x (\cos\left(\sqrt{2}x\right) + i\sin\left(\sqrt{2}x\right)) \end{cases}$$

where Euler's formula (Definition 8.1.6) is applied. Hence

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} 1 & i & -i \\ 0 & i & -i \\ 0 & \sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} c_1 e^x \\ c_2 e^x (\cos\left(\sqrt{2}x\right) - i\sin\left(\sqrt{2}x\right)) \\ c_3 e^x (\cos\left(\sqrt{2}x\right) + i\sin\left(\sqrt{2}x\right)) \end{bmatrix} \\ &= \begin{bmatrix} c_1 e^x + (c_2 + c_3) e^x \sin\left(\sqrt{2}x\right) + i(c_2 - c_3) e^x \cos\left(\sqrt{2}x\right) \\ (c_2 + c_3) e^x \sin\left(\sqrt{2}x\right) + i(c_2 - c_3) e^x \cos\left(\sqrt{2}x\right) \\ \sqrt{2}(c_2 + c_3) e^x \cos\left(\sqrt{2}x\right) - i\sqrt{2}(c_2 - c_3) e^x \sin\left(\sqrt{2}x\right) \end{bmatrix} \\ &= \begin{bmatrix} c_1 e^x + C_2 e^x \sin\left(\sqrt{2}x\right) + iC_3 e^x \cos\left(\sqrt{2}x\right) \\ C_2 e^x \sin\left(\sqrt{2}x\right) + iC_3 e^x \cos\left(\sqrt{2}x\right) \\ \sqrt{2}C_2 e^x \cos\left(\sqrt{2}x\right) - i\sqrt{2}C_3 e^x \sin\left(\sqrt{2}x\right) \end{bmatrix} \end{aligned}$$

where we set $C_2 = c_2 + c_3$, $C_3 = c_2 - c_3$. Both the real and imaginary parts 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

$$\mathbf{y}'_{\mathrm{Re}} = A\mathbf{y}_{\mathrm{Re}}$$

$$\mathbf{y}'_{\text{Im}} = A\mathbf{y}_{\text{Im}}$$

So the final answer, expressed in real values, is

$$\begin{cases} y_1 = c_1 e^x + C_2 e^x \sin\left(\sqrt{2}x\right) + C_3 e^x \cos\left(\sqrt{2}x\right) \\ y_2 = C_2 e^x \sin\left(\sqrt{2}x\right) + C_3 e^x \cos\left(\sqrt{2}x\right) \\ y_3 = \sqrt{2}C_2 e^x \cos\left(\sqrt{2}x\right) - \sqrt{2}C_3 e^x \sin\left(\sqrt{2}x\right) \end{cases}$$

where c_1, C_2, C_3 are to be decided by initial condition. The same solution basis of

$$\begin{bmatrix} e^x \sin\left(\sqrt{2}x\right) \\ e^x \sin\left(\sqrt{2}x\right) \\ \sqrt{2}e^x \cos\left(\sqrt{2}x\right) \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} e^x \cos\left(\sqrt{2}x\right) \\ e^x \cos\left(\sqrt{2}x\right) \\ -\sqrt{2}e^x \sin\left(\sqrt{2}x\right) \end{bmatrix}$$

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. □

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 the Jordan Normal Form. This is an advanced topic and interested readers can go to Appendix B.2 and B.3 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

$$P \xrightarrow{k_1} Q \xrightarrow{k_2} R$$

and hence have rate laws in the form of

$$\begin{cases} d[P]/dt = -k_1[P] \\ d[Q]/dt = k_1[P] - k_2[Q] \\ d[R]/dt = k_2[Q] \end{cases}$$

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 its 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 $k_1 \neq k_2$.[18] The eigenvector for $\lambda = 0$ is clearly $(0, 0, 1)^T$, while that for $\lambda = -k_1$ can be

[18] 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.

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 matrix for the change of variables 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

$$\begin{aligned} \mathbf{z}' &= (P^{-1}AP)\mathbf{z} = D\mathbf{z} \\ &= \begin{bmatrix} -k_1 & 0 & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{z} \end{aligned}$$

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

$$\mathbf{y} = \begin{bmatrix} k_1 - k_2 & 0 & 0 \\ -k_1 & -1 & 0 \\ k_2 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{k_1 - k_2} e^{-k_1 t} \\ -\frac{k_1}{k_1 - k_2} e^{-k_2 t} \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} e^{-k_1 t} \\ -\frac{k_1}{k_1-k_2}e^{-k_1 t} + \frac{k_1}{k_1-k_2}e^{-k_2 t} \\ \frac{k_2}{k_1-k_2}e^{-k_1 t} - \frac{k_1}{k_1-k_2}e^{-k_2 t} + 1 \end{bmatrix}$$

Hence

$$\begin{cases} [P] &= e^{-k_1 t} \\ [Q] &= -\dfrac{k_1}{k_1-k_2}e^{-k_1 t} + \dfrac{k_1}{k_1-k_2}e^{-k_2 t} \\ [R] &= \dfrac{k_2}{k_1-k_2}e^{-k_1 t} - \dfrac{k_1}{k_1-k_2}e^{-k_2 t} + 1 \end{cases}$$

The time t_Q where $[Q]$ reaches maximum occurs when $\frac{d[Q]}{dt} = 0$, which leads to

$$\begin{aligned} \frac{d[Q]}{dt} &= k_1[P] - k_2[Q] = 0 \\ k_1 e^{-k_1 t} - k_2\left(-\frac{k_1}{(k_1-k_2)}e^{-k_1 t} + \frac{k_1}{(k_1-k_2)}e^{-k_2 t}\right) &= 0 \\ k_1 e^{-k_1 t} + \frac{k_1 k_2}{(k_1-k_2)}e^{-k_1 t} - \frac{k_1 k_2}{(k_1-k_2)}e^{-k_2 t} &= 0 \\ \frac{k_1^2}{(k_1-k_2)}e^{-k_1 t} - \frac{k_1 k_2}{(k_1-k_2)}e^{-k_2 t} &= 0 \\ \frac{k_1^2}{(k_1-k_2)}e^{-k_1 t} &= \frac{k_1 k_2}{(k_1-k_2)}e^{-k_2 t} \\ e^{-(k_1-k_2)t} &= \frac{k_2}{k_1} \\ -(k_1-k_2)t &= \ln\left(\frac{k_2}{k_1}\right) \\ \therefore\ t_Q = -\frac{\ln\left(\frac{k_2}{k_1}\right)}{k_1-k_2} &= \frac{\ln k_1 - \ln k_2}{k_1-k_2} \end{aligned}$$

□

Example 9.4.2. From Example 4.3.1, we know that the horizontal acceleration of a mass due to Coriolis Force is

$$\begin{cases} \dfrac{du}{dt} = fv \\ \dfrac{dv}{dt} = -fu \end{cases} \tag{9.20}$$

where $f = 2\Omega \sin\varphi$, with φ being the latitude. Assume that some sea current in an inland sea (let's say, the Baltic Sea, $\varphi \approx 58\,°\text{N}$) 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 at 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:

$$\left[\begin{array}{cc|c} -fi & f & 0 \\ -f & -fi & 0 \end{array}\right] \to \left[\begin{array}{cc|c} -fi & f & 0 \\ 0 & 0 & 0 \end{array}\right] \qquad R_2 + iR_1 \to R_2$$

which gives $\vec{v}_{\lambda_+} = (-i, 1)^T$. As a result,

$$\vec{v} = \vec{v}_{\lambda_+} e^{\lambda_+ t}$$

$$
\begin{aligned}
&= \begin{bmatrix} -i \\ 1 \end{bmatrix} e^{ift} \\
&= \begin{bmatrix} -i \\ 1 \end{bmatrix} (\cos(ft) + i\sin(ft)) && \text{(Definition 8.1.6)} \\
&= \begin{bmatrix} -i(\cos(ft) + i\sin(ft)) \\ \cos(ft) + i\sin(ft) \end{bmatrix} \\
&= \begin{bmatrix} -i\cos(ft) + \sin(ft) \\ \cos(ft) + i\sin(ft) \end{bmatrix} \\
&= \begin{bmatrix} \sin(ft) \\ \cos(ft) \end{bmatrix} + i \begin{bmatrix} -\cos(ft) \\ \sin(ft) \end{bmatrix}
\end{aligned}
$$

Considering the real and imaginary parts as two linearly independent solutions, we have

$$
\vec{v} = c_1 \begin{bmatrix} \sin(ft) \\ \cos(ft) \end{bmatrix} + c_2 \begin{bmatrix} -\cos(ft) \\ \sin(ft) \end{bmatrix}
$$

which translates to $u = c_1 \sin(ft) - c_2 \cos(ft)$, $v = c_1 \cos(ft) + c_2 \sin(ft)$, 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 $T = \frac{2\pi}{f} = \frac{2\pi}{2\Omega \sin\varphi} = \frac{1}{2\sin\varphi}(\frac{2\pi}{\Omega}) = \frac{1}{2\sin\varphi}(24\,\text{h})$. In the case of the Baltic Sea, the oscillation period will then be about $T = \frac{1}{2\sin(58°)}(24\,\text{h}) = 14.1\,\text{h}$. □

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 matrices 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 (and the characteristic polynomial of A splits), 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.

$$A = \begin{bmatrix} 1 & 3 & 3 \\ 4 & 0 & 4 \\ 0 & 1 & 2 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

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 Show that for any $n \times n$ square matrix A, if its characteristic polynomial is given by

$$p(\lambda) = \det(A - \lambda I) = (-1)^n(\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_0)$$

then $\mathrm{tr}(A) = -c_{n-1}$ and $\det(A) = (-1)^n c_0$ where $\mathrm{tr}()$ is the trace. You may assume for simplicity, that the matrix A is diagonalizable. However, this result actually holds for all square matrices regardless of the diagonalizability.

Exercise 9.7 For the matrix

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

compute its characteristic polynomial. By applying the 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 9.8 Apply diagonalization for the following matrix, and show that the characteristic polynomial remains unchanged under diagonalization, if possible.

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

$$C = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 5 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Exercise 9.9 Solve the following system of ODEs.

$$\begin{cases} y_1' = -y_1 + y_2 + 3y_3 \\ y_2' = -2y_1 + 3y_2 + 2y_3 \\ y_3' = -2y_1 + y_2 + 4y_3 \end{cases}$$

with the initial condition $y_1(0) = 3$, $y_2(0) = 2$, $y_3(0) = 9$.

Exercise 9.10 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 $P \to Q \to R \to P$ 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 reaction constants. Derive the time evolution of the concentrations of the three gases. What happens when $t \to \infty$?

Chapter 10

Orthogonal and Normal Matrices

We have discussed 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 real 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 to be n too so that they form an *orthonormal basis* by Properties 6.3.12 and part (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

$$P = \left[\vec{v}^{(1)} | \cdots | \vec{v}^{(j)} | \cdots | \vec{v}^{(n)}\right]$$

is said to be an orthogonal matrix, if all of its columns $\vec{v}^{(j)} \in \mathbb{R}^n$ satisfy the relationship:

$$\vec{v}^{(p)} \cdot \vec{v}^{(q)} = \begin{cases} 1 & \text{if } p = q \quad \text{i.e. } \left\|\vec{v}^{(p)}\right\| = 1 \\ 0 & \text{if } p \neq q \end{cases} \tag{10.1}$$

so that the column vectors $\vec{v}^{(j)}$, $j = 1, 2, \ldots, n$, form 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.[1]

Due to its definition, an orthogonal matrix P has the property $P^T P = I$, since the resulting entries of this matrix product $P^T P$ are basically dot products of the column vectors of P (see the explanation below Definition 4.2.1), where explicitly

$$\begin{aligned}
P^T P &= \left[\vec{v}^{(1)} | \cdots | \vec{v}^{(j)} | \cdots | \vec{v}^{(n)}\right]^T \left[\vec{v}^{(1)} | \cdots | \vec{v}^{(j)} | \cdots | \vec{v}^{(n)}\right] \\
&= \begin{bmatrix} \vec{v}^{(1)T} \\ \hline \vdots \\ \hline \vec{v}^{(j)T} \\ \hline \vdots \\ \hline \vec{v}^{(n)T} \end{bmatrix} \left[\vec{v}^{(1)} | \cdots | \vec{v}^{(j)} | \cdots | \vec{v}^{(n)}\right] \\
&= \begin{bmatrix} \vec{v}^{(1)} \cdot \vec{v}^{(1)} & \cdots & \vec{v}^{(1)} \cdot \vec{v}^{(j)} & \cdots & \vec{v}^{(1)} \cdot \vec{v}^{(n)} \\ \vdots & \ddots & \vdots & & \vdots \\ \vec{v}^{(j)} \cdot \vec{v}^{(1)} & \cdots & \vec{v}^{(j)} \cdot \vec{v}^{(j)} & \cdots & \vec{v}^{(j)} \cdot \vec{v}^{(n)} \\ \vdots & & \vdots & \ddots & \vdots \\ \vec{v}^{(n)} \cdot \vec{v}^{(1)} & \cdots & \vec{v}^{(n)} \cdot \vec{v}^{(j)} & \cdots & \vec{v}^{(n)} \cdot \vec{v}^{(n)} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} = I_n
\end{aligned}$$

By Definition 2.2.1, $P^T P = I$ means that $PP^T = P^T P = I$ where the transpose of the orthogonal matrix $P^T = P^{-1}$ is exactly its inverse. The argument works in both directions such that $PP^T = I$ also means that P is orthogonal.

Properties 10.1.2. P is an orthogonal matrix if and only if

$$PP^T = P^T P = I \tag{10.2}$$

so that 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 latter equality indicates that the rows of P also have to form an orthonormal basis. So, a parallel definition of orthogonal matrices is

Definition 10.1.3. An $n \times n$ square matrix

$$P = \begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vdots \\ \hline \vec{w}^{(i)T} \\ \hline \vdots \\ \hline \vec{w}^{(n)T} \end{bmatrix}$$

is an orthogonal matrix, also if all the rows $\vec{w}^{(i)} \in \mathbb{R}^n$ satisfy the similar relationship:

$$\vec{w}^{(p)} \cdot \vec{w}^{(q)} = \begin{cases} 1 & \text{if } p = q \quad \text{i.e. } \left\|\vec{w}^{(p)}\right\| = 1 \\ 0 & \text{if } p \neq q \end{cases} \tag{10.3}$$

so that the row vectors $\vec{w}^{(i)}$, $i = 1, 2, \ldots, n$, constitute an orthonormal basis too.

► Short Exercise: Confirm Properties 10.1.2 for the matrix in the last short exercise.[2]

[1]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 it to the readers to check $\vec{v}^{(2)} \cdot \vec{v}^{(3)} = 0$ as well. Also, $\left\|\vec{v}^{(1)}\right\| = \sqrt{(\frac{1}{\sqrt{2}})^2 + (\frac{1}{\sqrt{2}})^2 + (0)^2} = 1$ and $\left\|\vec{v}^{(2)}\right\| = \sqrt{(\frac{1}{\sqrt{3}})^2 + (-\frac{1}{\sqrt{3}})^2 + (\frac{1}{\sqrt{3}})^2} = 1$. Again we let the readers to check $\left\|\vec{v}^{(3)}\right\| = 1$ too.

10.1.2. Geometric Implications of Orthogonal Matrices

All invertible matrices represent some sort of coordinate transformation as highlighted in Section 7.2.1. Orthogonal matrices further belong to 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. Using the relationship (10.2) from Properties 10.1.2 and taking determinants on both sides of it, we have

$$\begin{aligned} \det\left(P^T P\right) &= \det(I) \\ \det\left(P^T\right)\det(P) &= (\det(P))^2 = 1 \\ \det(P) &= \pm 1 \end{aligned}$$

where Properties 2.3.9 is used. □

The 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).

---

[2]It is a direct computation that shows

$$\begin{aligned} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

and $PP^T = I$ similarly.

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 the vector $\vec{v}$ will be $P^{-1}\vec{v} = P^T\vec{v}$ (Properties 10.1.2). The length of this vector in the new coordinate frame is

$$\begin{aligned}
\left\|P^T\vec{v}\right\| &= \sqrt{(P^T\vec{v})\cdot(P^T\vec{v})} && \text{(Properties 4.2.2)}\\
&= \sqrt{(P^T\mathbf{v})^T(P^T\mathbf{v})} && \text{(Properties 4.2.3)}\\
&= \sqrt{(\mathbf{v}^TPP^T\mathbf{v})} && \text{(Properties 4.2.3)}\\
&= \sqrt{(\mathbf{v}^TI\mathbf{v})} && \text{(Properties 10.1.2)}\\
&= \sqrt{(\mathbf{v}^T\mathbf{v})} = \|\vec{v}\| && \text{(Properties 4.2.2 again)}
\end{aligned}$$

So the length of the vector remains the same even when the coordinate system is changed. For active transformation that actually changes the vector in space instead of just 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 the scaling factor for a coordinate transformation in Section 7.2.1 will be helpful). □

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 physically while keeping the coordinate frame unchanged (active transformation), or (b) rotation and reflection of the coordinate system, while keeping the vector in the original place (passive

transformation). (See Figure 10.1) By the previous length-preserving properties, there will be no stretching/compression or shearing, and the rotation/reflection will be pure.

Figure 10.1.: *Left: case (a), active transformation, and right: case (b), passive transformation, for a 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 7.1.14, 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 (b) of the figure above, we have

$$\begin{cases} \hat{x}' = (\cos\theta)\hat{x} + (\sin\theta)\hat{y} & \text{(10.4a)} \\ \hat{y}' = (-\sin\theta)\hat{x} + (\cos\theta)\hat{y} & \text{(10.4b)} \end{cases}$$

The corresponding transition matrix is then

$$\begin{aligned} P^S_\beta &= \left[[\hat{x}']_S | [\hat{y}']_S \right] \\ &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \end{aligned} \tag{10.5}$$

where S represents the standard basis, and β indicates the rotated coordinates system. Alternatively,

$$P^\beta_S = (P^S_\beta)^{-1}$$

$$= (P_\beta^S)^T \qquad \text{(Properties 10.1.2)}$$
$$= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

These two matrices can be compared to the one we saw in the last chapter when we studied 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 produce $\vec{v}_n$ while staying in the same coordinate frame. This is the case (a) active transformation we have discussed above. For case (b) passive transformation, 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 a situation, the vector is still the same one in a physical sense, but represented in the new coordinate system. Most situations belong to case (b), on which we will focus our discussion thereafter, and thus we will assume all coordinate transformations will be the passive type unless otherwise specified.

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 thus $P_f^T = P_n^T \cdots P_3^T P_2^T P_1^T$. 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?[3]

Example 10.1.2. For a vector $\vec{v}_0$ in the three-dimensional standard basis S, if the coordinate system undergoes a positive rotation about the straight line $x = 0, y = z$ by a degree of θ, find its representation $\vec{v}_n$ in the new system β.

Solution. The common way to construct a 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° clockwise so that the intermediate z' axis is oriented along the concerned straight 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 via multiplying by its inverse P_1^{-1} which is equal to P_1^T due to orthogonality (Properties 10.1.2). Hence the net

[3]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.

Figure 10.2.: *Flowchart for Example 10.1.2.*

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 are then

$$
\begin{aligned}
\vec{v}_0 &= P_f \vec{v}_n \\
\vec{v}_n &= (P_f)^T \vec{v}_0 \\
&= \begin{bmatrix} \cos\theta & \frac{\sin\theta}{\sqrt{2}} & -\frac{\sin\theta}{\sqrt{2}} \\ -\frac{\sin\theta}{\sqrt{2}} & \frac{(\cos\theta)+1}{2} & \frac{-(\cos\theta)+1}{2} \\ \frac{\sin\theta}{\sqrt{2}} & \frac{-(\cos\theta)+1}{2} & \frac{(\cos\theta)+1}{2} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \\
&= \begin{bmatrix} (\cos\theta)x_0 + \left(\frac{\sin\theta}{\sqrt{2}}\right) y_0 + \left(-\frac{\sin\theta}{\sqrt{2}}\right) z_0 \\ \left(-\frac{\sin\theta}{\sqrt{2}}\right) x_0 + \left(\frac{(\cos\theta)+1}{2}\right) y_0 + \left(\frac{-(\cos\theta)+1}{2}\right) z_0 \\ \left(\frac{\sin\theta}{\sqrt{2}}\right) x_0 + \left(\frac{-(\cos\theta)+1}{2}\right) y_0 + \left(\frac{(\cos\theta)+1}{2}\right) z_0 \end{bmatrix}
\end{aligned}
$$

□

► Short Exercise: For the example above, if $(x_0, y_0, z_0)^T = (0, 1, 1)^T$ and $\theta = \frac{\pi}{6}$, find $\vec{v}_n = (x_n, y_n, z_n)^T$.[4]

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, comprised of the eigenvectors of A, is orthonormal, i.e. P is an orthogonal matrix. We will discuss the case where A is a real matrix first.

Definition 10.2.1 (Orthogonal Diagonalization)**.** Orthogonal diagonalization on a real $n \times n$ square matrix A is to find an orthogonal matrix P that consists of columns being the n orthonormal $\mathbb{R}^n$ eigenvectors of A (Definition 10.1.1) such that

$$P^{-1}AP = P^TAP = D \tag{10.6}$$

is a real diagonal matrix (Properties 9.2.2 and 10.1.2).

Note that the requirements of eigenvectors being $\mathbb{R}^n$ and hence orthogonality defined by the real dot product implicitly constrain us to work over $\mathbb{R}$ and only real eigenvalues are allowed. Not all real square matrices can be orthogonally diagonalized because orthogonal diagonalization is a stronger version of ordinary diagonalization, and 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.

As a corollary, all symmetric real matrices have only real eigenvalues. The "only if" part is very easy to show.[5] On the other hand, the "if" part is much more difficult to derive and requires us to build some intermediate results first,

---

[4]It is

$$\begin{bmatrix} \cos\frac{\pi}{6} & \frac{\sin\frac{\pi}{6}}{\sqrt{2}} & -\frac{\sin\frac{\pi}{6}}{\sqrt{2}} \\ -\frac{\sin\frac{\pi}{6}}{\sqrt{2}} & \frac{(\cos\frac{\pi}{6})+1}{2} & \frac{-(\cos\frac{\pi}{6})+1}{2} \\ \frac{\sin\frac{\pi}{6}}{\sqrt{2}} & \frac{-(\cos\frac{\pi}{6})+1}{2} & \frac{(\cos\frac{\pi}{6})+1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} (\frac{\sqrt{3}}{2})(0) + (\frac{1}{2\sqrt{2}})(1) + (-\frac{1}{2\sqrt{2}})(1) \\ (-\frac{1}{2\sqrt{2}})(0) + (\frac{\frac{1}{2}+1}{2})(1) + (\frac{-\frac{1}{2}+1}{2})(1) \\ (\frac{1}{2\sqrt{2}})(0) + (\frac{-\frac{1}{2}+1}{2})(1) + (\frac{\frac{1}{2}+1}{2})(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

In fact, all vectors in the form of $s(0, 1, 1)^T$ where s is any number will keep the same coordinates after transformation, no matter what the value θ takes, because the vector is oriented exactly along the rotational axis in question.

[5]Note that if A is orthogonally diagonalizable then $A = PDP^{-1} = PDP^T$. Then $A^T = (PDP^T)^T = PD^TP^T = PDP^T = A$ since the transpose of any diagonal matrix $D^T = D$ is equal to itself.

which will be done sequentially in the remainder of this section. Since the transformation matrix P used to diagonalize a given matrix A is formed from 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 $\vec{v}^{(1)}$ and $\vec{v}^{(2)}$ which have two different eigenvalues λ_1 and λ_2. Then consider the quantity $\vec{v}^{(1)} \cdot (A\vec{v}^{(2)})$ where A is the real symmetric matrix. By Definition 9.1.1, it is equal to $\vec{v}^{(1)} \cdot (\lambda_2\vec{v}^{(2)})$ $= \lambda_2(\vec{v}^{(1)} \cdot \vec{v}^{(2)})$. At the same time, using Properties 4.2.3 too, we have

$$\begin{aligned}
\vec{v}^{(1)} \cdot (A\vec{v}^{(2)}) &= (A^T\vec{v}^{(1)}) \cdot \vec{v}^{(2)} \\
&= (A\vec{v}^{(1)}) \cdot \vec{v}^{(2)} && (A \text{ is symmetric}) \\
&= (\lambda_1\vec{v}^{(1)}) \cdot \vec{v}^{(2)} = \lambda_1(\vec{v}^{(1)} \cdot \vec{v}^{(2)}) && (\text{Definition 9.1.1 again})
\end{aligned}$$

Therefore

$$\begin{aligned}
\vec{v}^{(1)} \cdot (A\vec{v}^{(2)}) = \lambda_1(\vec{v}^{(1)} \cdot \vec{v}^{(2)}) &= \lambda_2(\vec{v}^{(1)} \cdot \vec{v}^{(2)}) \\
(\lambda_1 - \lambda_2)(\vec{v}^{(1)} \cdot \vec{v}^{(2)}) &= 0
\end{aligned}$$

Since the two eigenvalues are required to be distinct, $\lambda_1 \neq \lambda_2$, and $\vec{v}^{(1)} \cdot \vec{v}^{(2)}$ must be 0. By Properties 4.2.5, the two eigenvectors $\vec{v}^{(1)}$ and $\vec{v}^{(2)}$ are orthogonal to each other. □

Even when there are multiple eigenvectors (a geometric multiplicity/an eigenspace of dimension ≥ 2) associated with 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 multiplicities of all their eigenvalues are equal to the corresponding algebraic multiplicities as suggested by Properties 9.2.2[6], or in other words, there is no *"deficient"* eigenvalue.

Properties 10.2.4. The geometric multiplicity of any distinct eigenvalue of a real symmetric matrix is always strictly equal to its algebraic multiplicity.

Proof. Let A be some $n \times n$ real symmetric matrix. Denote the geometric multiplicity of any distinct eigenvalue λ_{J_i} of A by 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 can be n_{J_i} orthonormal (see the discussion above) eigenvectors $\vec{v}^{(1)}_{\lambda_{J_i}}, \vec{v}^{(2)}_{\lambda_{J_i}}, \ldots, \vec{v}^{(n_{J_i})}_{\lambda_{J_i}} \in \mathbb{R}^n$ corresponding to λ_{J_i} in the eigenspace $\mathcal{E}_{J_i}$. By part (c) of Properties 6.2.7, we can complete a basis $\mathcal{E}_{J_i} \oplus (\mathcal{E}_{J_i})^C$ for $\mathbb{R}^n$ by filling them with some other orthonormal (by Gram-Schmidt) $n - n_{J_i}$ vectors. Let P be the orthogonal coordinate matrix that represents this basis in its columns, such that

$$P = \left[\vec{v}^{(1)}_{\lambda_{J_i}} | \vec{v}^{(2)}_{\lambda_{J_i}} | \cdots | \vec{v}^{(n_{J_i})}_{\lambda_{J_i}} | \text{other vectors used to fill the basis for } \mathbb{R}^n\right]$$

similar to the derivation of Theorem 9.1.3. With the same logic, we can apply a change of coordinates for A with P, such that

$$\begin{aligned} A' &= P^{-1}AP \\ &= \begin{bmatrix} \lambda_{J_i} I_{n_{J_i}} & *_{n_{J_i} \times (n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i}) \times n_{J_i}} & *_{(n-n_{J_i}) \times (n-n_{J_i})} \end{bmatrix} \end{aligned}$$

[6]The another condition requires the characteristic polynomial to split over $\mathbb{R}$, but it is immediately satisfied due to the previous corollary to Theorem 10.2.2 which points out that the eigenvalues of any symmetric matrix are always real (see Footnote 6 of Chapter 9).

However, this time, as P has been declared 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.[7] Therefore, the upper-right block of A' also has to be a zero submatrix (as the transpose of the zero submatrix to the lower left):

$$A' = \begin{bmatrix} \lambda_j I_{n_{J_i}} & [\mathbf{0}]_{n_{J_i}\times(n-n_{J_i})} \\ [\mathbf{0}]_{(n-n_{J_i})\times n_{J_i}} & *_{(n-n_{J_i})\times(n-n_{J_i})} \end{bmatrix}$$

which is now in the form of a matrix direct sum (see Definition 8.3.7). This shows that we can treat it as $A' = \lambda_j I_{n_{J_i}} \oplus A^C$ 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 that makes up the bottom right asterisked block. From the perspective of a linear transformation, the restriction of A' to $(\mathcal{E}_{J_i})^C$ is simply $A^C : (\mathcal{E}_{J_i})^C \to (\mathcal{E}_{J_i})^C$ which is a linear operator by itself. Now if the algebraic multiplicity of λ_{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 some other polynomial. By Properties 9.2.3, A' will also have this same characteristic polynomial $p_{A'}(\lambda) = (\lambda_{J_i} - \lambda)^{k_{J_i}} p^-(\lambda)$. At the same time, from the structure of A' derived above, we know that

$$p_{A'}(\lambda) = (\lambda_{J_i} - \lambda)^{n_{J_i}} p^C(\lambda)$$

by repeated cofactor expansions along the first n_{J_i} columns, with $p^C(\lambda)$ denoting the characteristic polynomial for A^C. 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 $n_{J_i} < k_{J_i}$, we know that $k_{J_i} - n_{J_i} \geq 1$, and $p^C(\lambda)$ will then contain some $(\lambda_j - \lambda)$ factor. Hence A^C, as a square matrix in its own right,

[7] $(P^TAP)^T = P^TA^TP = P^TAP$.

must have an eigenvector $\vec{v}_{\lambda_{J_i}}^{(n_{J_i}+1)}$ in the $(\mathcal{E}_{J_i})^C$ subspace corresponding to λ_{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 matrix A and will be linearly independent of $\vec{v}_{\lambda_{J_i}}^{(1)}, \vec{v}_{\lambda_{J_i}}^{(2)}, \ldots, \vec{v}_{\lambda_{J_i}}^{(n_{J_i})}$ because $\mathcal{E}_{J_i} \oplus (\mathcal{E}_{J_i})^C$ is a direct sum (Definition 6.2.9 this time). Therefore, this shows that the geometric multiplicity of λ_{J_i} in A is actually $n_{J_i} + 1$ which contradicts our hypothesis that it is n_{J_i}, whenever $n_{J_i} < k_{J_i}$. (We can also inductively use this argument until the geometric multiplicity adds up to k_{J_i}.) Hence the only reasonable conclusion is $n_{J_i} = k_{J_i}$. □

Properties 10.2.3 and 10.2.4 together imply the "if" part of Theorem 10.2.2, and the proof for the orthogonal diagonalization theorem 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

$$\vec{v}_\lambda = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad \text{for } \lambda = 1$$

$$\vec{v}_\lambda = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad \text{for } \lambda = 3$$

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

$$\begin{aligned} P^T AP &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} = D \end{aligned}$$

□

► Short Exercise: Confirm that P is orthogonal.[8]

Remark

From Properties 7.2.2, we know that orthogonal diagonalization $D = P^{-1}AP = P^T AP$ is simply a special case of coordinate transformation for a square matrix, where P is now orthogonal and only pure rotations/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 vector $\vec{u}$, 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 first generalize the notion of projections involving multi-dimensional subspaces. 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 $T : \mathcal{V} \to \mathcal{V}$ such that for any $\vec{v} = \vec{w}_1 + \vec{w}_2$[9] 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.

---

[8]It is simply checking

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

[9]This decomposition of $\vec{v}$ into $\vec{w}_1$ and $\vec{w}_2$ is unique as $\mathcal{W}_1 \oplus \mathcal{W}_2$ is required to be a direct sum. (see Section 6.2.3)

Properties 10.3.1 (Projection)**.** A linear operator $T : \mathcal{V} \to \mathcal{V}$ 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 demand $\mathcal{W}_2$ such that we can write $\vec{v} = \vec{w}_1 + (\vec{v} - \vec{w}_1) = \vec{w}_1 + \vec{w}_2$ for any $\vec{v} \in \mathcal{V}$, where $\vec{v} - \vec{w}_1 = \vec{w}_2 \in \mathcal{W}_2$ and $\mathcal{W}_1 \oplus \mathcal{W}_2 = \mathcal{V}$[10]. Applying T on both sides gives

$$\begin{aligned}
T(\vec{v}) &= T(\vec{w}_1 + \vec{v} - \vec{w}_1) \\
T(\vec{v}) &= T(\vec{w}_1) + T(\vec{v} - \vec{w}_1) \quad \text{(Linearity from Definition 7.1.1)} \\
\vec{w}_1 &= \vec{w}_1 + T(\vec{v} - \vec{w}_1) \\
\implies T(\vec{v} - \vec{w}_1) &= \mathbf{0}
\end{aligned}$$

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 beginning, for any arbitrary $\vec{v}$, $T^2(\vec{v}) = T(T(\vec{v})) = T(\vec{w}_1) = \vec{w}_1 = T(\vec{v})$, so by Properties 7.1.3 it must be that $T^2 = T$. □

A linear operator/matrix T satisfying the condition $T^2 = T$ is known as ***idempotent***, and we have $T^n = T$ for any positive integer n.[11] So in other words, a linear operator is a projection if and only if it is idempotent.

Example 10.3.1. Show that the linear operator T which has a matrix representa-

[10] $\mathcal{W}_2$ 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.

[11] For $n \geq 2$, $T^n = (T^2)(T^{n-2}) = (T)(T^{n-2}) = T^{n-1}$ which can be recursively used.

tion 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

$$\begin{aligned} [T]^2 &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} (0)(0) + (1)(0) & (0)(1) + (1)(1) \\ (0)(0) + (1)(0) & (0)(1) + (1)(1) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} = [T] \end{aligned}$$

So it is indeed a projection. The subspace onto which the linear operator projects, is simply equal to its range, or equivalently the column space of $[T]$, which can be immediately identified as $\operatorname{span}(\{(1,1)^T\})$, that is, the straight line $y = x$. Similarly, the projection is along its kernel/null space, which is also easily seen to be $\operatorname{span}(\{(1,0)^T\})$, 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 $\mathcal{W}_2$

Figure 10.3.: *Effects of projections (red to blue) in Example 10.3.1.*

to be an orthogonal projection, these 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 $\mathcal{N}(T)^\perp = \mathcal{R}(T)$ now. An equivalent condition of an orthogonal projection is that $[T] = [T]^T$ is symmetric given we are working over $\mathbb{R}$.[12]

Properties 10.3.2 (Orthogonal Projection)**.** A *real*, *finite-dimensional* linear projection operator $T : \mathcal{V} \to \mathcal{V}$ is an orthogonal projection (with respect to the usual real dot product) if and only if $[T] = [T]^T$ in terms of its matrix

[12]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 12.

representation.

Proof. The "if" part: We need to show that $[T] = [T]^T$ implies $\mathcal{R}(T)^\perp = \mathcal{N}(T)$ (and $\mathcal{N}(T)^\perp = \mathcal{R}(T)$)[13]. 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 $\vec{v} = \vec{w}_1^{(2)} + \vec{w}_2^{(2)}$ 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

$$\begin{aligned}
\vec{u} \cdot T(\vec{v}) = (\vec{w}_1^{(1)} + \vec{w}_2^{(1)}) \cdot (\vec{w}_1^{(2)}) &= \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} + \vec{w}_2^{(1)} \cdot \vec{w}_1^{(2)} \\
&= \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)}
\end{aligned}$$

[13]This two are equivalent as $\mathcal{V}$ is finite-dimensional.

where $\vec{w}_2^{(1)} \cdot \vec{w}_1^{(2)} = 0$ because $\mathcal{N}(T) = \mathcal{R}(T)^{\perp}$. Similarly we have $[T]\vec{u} \cdot \vec{v} \equiv T(\vec{u}) \cdot \vec{v} = \vec{w}_1^{(1)} \cdot (\vec{w}_1^{(2)} + \vec{w}_2^{(2)}) = \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} + \vec{w}_1^{(1)} \cdot \vec{w}_2^{(2)} = \vec{w}_1^{(1)} \cdot \vec{w}_1^{(2)} = \vec{u} \cdot T(\vec{v})$. At the same time

$$\begin{aligned} \vec{u} \cdot T(\vec{v}) &= \vec{u} \cdot ([T]\vec{v}) \\ &= ([T]^T \vec{u}) \cdot \vec{v} \qquad \text{(Properties 4.2.3)} \end{aligned}$$

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 $[T] = [T]^T$. □

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:

$$[T] = \left[T(\hat{e}^{(1)}) | T(\hat{e}^{(2)})\right]$$

Its columns are the outputs of applying the orthogonal projection on the standard unit vectors. Geometrically (see Figure 10.4 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

$$T(\vec{v}) \equiv [T](1, -2)^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$

□

Figure 10.4.: *Illustration for Example 10.3.2.*

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 according to 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 it to the readers to check that it is $\{(\frac{4}{\sqrt{17}}, 0, \frac{1}{\sqrt{17}})^T, (-\frac{2}{\sqrt{357}}, \frac{17}{\sqrt{357}}, \frac{8}{\sqrt{357}})^T\}$. Motivated by Properties 7.2.7, we will derive the third vector to complete an orthonormal basis for $\mathbb{R}^3$ and it 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 $\beta = \{(\frac{4}{\sqrt{17}}, 0, \frac{1}{\sqrt{17}})^T, (-\frac{2}{\sqrt{357}}, \frac{17}{\sqrt{357}}, \frac{8}{\sqrt{357}})^T, (-\frac{1}{\sqrt{21}}, -\frac{2}{\sqrt{21}}, \frac{4}{\sqrt{21}})^T\}$ will have a simple matrix representation of

$$[T]_\beta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

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 (7.13) of Properties 7.2.2, where

$$P^S_\beta = \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix}$$

and

$$\begin{aligned} [T]_S &= (P^\beta_S)^{-1}[T]_\beta P^\beta_S \\ &= P^S_\beta [T]_\beta (P^S_\beta)^T \quad \text{(Properties 10.1.2 for the orthogonal } P \text{ matrix)} \\ &= \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{4}{\sqrt{17}} & -\frac{2}{\sqrt{357}} & -\frac{1}{\sqrt{21}} \\ 0 & \frac{17}{\sqrt{357}} & -\frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{17}} & \frac{8}{\sqrt{357}} & \frac{4}{\sqrt{21}} \end{bmatrix}^T \\ &= \begin{bmatrix} \frac{20}{21} & -\frac{2}{21} & \frac{4}{21} \\ -\frac{2}{21} & \frac{17}{21} & \frac{8}{21} \\ \frac{4}{21} & \frac{8}{21} & \frac{5}{21} \end{bmatrix} \end{aligned}$$

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]. □

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

$$\begin{aligned} T(\vec{v}^T) &\equiv \frac{1}{\|\mathbf{u}\|^2}(\mathbf{v}^T\mathbf{u})\mathbf{u}^T \\ \implies T(\vec{v}) &\equiv \frac{1}{\|\mathbf{u}\|^2}(\mathbf{v}^T\mathbf{u}\mathbf{u}^T)^T \\ &= \left(\frac{1}{\|\mathbf{u}\|^2}\mathbf{u}\mathbf{u}^T\right)\mathbf{v} \\ &= \left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)^T\right)\mathbf{v} = [T]\mathbf{v} \end{aligned}$$

where the orthogonal projection matrix is

$$[T] = \frac{\mathbf{u}}{\|\mathbf{u}\|}\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)^T = \hat{\mathbf{u}}\hat{\mathbf{u}}^T \tag{10.7}$$

which is clearly symmetric and can be shown that $[T]^2 = [T]$.[14] 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 $\{\vec{w}^{(1)}, \vec{w}^{(2)}, \ldots, \vec{w}^{(r)}\}$, complete this orthonormal basis for $\mathbb{R}^n$ by appending another orthonormal basis $\{\vec{w}^{(r+1)}, \ldots, \vec{w}^{(n)}\}$ of its kernel $\mathcal{N}(T) = \mathcal{R}(T)^\perp$ (possible due to Definition 7.2.4 and Properties 7.2.7) just like Example 10.3.3 above. By the same argument in

[14] $(\hat{\mathbf{u}}\hat{\mathbf{u}}^T)(\hat{\mathbf{u}}\hat{\mathbf{u}}^T) = \hat{\mathbf{u}}(\hat{\mathbf{u}}^T\hat{\mathbf{u}})\hat{\mathbf{u}}^T = \hat{\mathbf{u}}(1)\hat{\mathbf{u}}^T = \hat{\mathbf{u}}\hat{\mathbf{u}}^T$ 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 which simply equals 1.

that example, we know that

$$[T] = \left[\vec{w}^{(1)}|\cdots|\vec{w}^{(r)}|\vec{w}^{(r+1)}|\cdots|\vec{w}^{(n)}\right]\begin{bmatrix} I_r & [\mathbf{0}]_{r\times(n-r)} \\ [\mathbf{0}]_{(n-r)\times r} & [\mathbf{0}]_{(n-r)\times(n-r)} \end{bmatrix}\begin{bmatrix} \vec{w}^{(1)T} \\ \hline \vdots \\ \hline \vec{w}^{(r)T} \\ \hline \vec{w}^{(r+1)T} \\ \hline \vdots \\ \hline \vec{w}^{(n)T} \end{bmatrix} \tag{10.8a}$$

$$= \hat{\mathbf{w}}^{(1)}\hat{\mathbf{w}}^{(1)T} + \hat{\mathbf{w}}^{(2)}\hat{\mathbf{w}}^{(2)T} + \cdots + \hat{\mathbf{w}}^{(r)}\hat{\mathbf{w}}^{(r)T} \tag{10.8b}$$

which generalizes the form of an orthogonal projection matrix above 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.

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 a projection onto $\mathcal{W} = \mathcal{R}(T)$. Therefore, $(\vec{v} - T(\vec{v})) \cdot (T(\vec{v}) - \vec{w}) = 0$, and

$$\|\vec{v} - \vec{w}\|^2 = \|\vec{v} - T(\vec{v})\|^2 + \|T(\vec{v}) - \vec{w}\|^2 > \|T(\vec{v}) - \vec{w}\|^2 \tag{10.9a}$$

$$> \|\vec{v} - T(\vec{v})\|^2 \tag{10.9b}$$

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 $T : \mathcal{V} \to \mathcal{V}$ 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 λ_j, $j = 1, 2, \ldots, n$ (counting repeated ones). Assume there are k distinct eigenvalues out of them. For every such distinct eigenvalue, collect all the indices j where λ_j points to that same eigenvalue, let's say, λ_{J_i}, and group them into an index set J_i, $i = 1, 2, \ldots, k$. 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) $\mathcal{V} = \mathcal{E}_{J_1} \oplus \mathcal{E}_{J_2} \oplus \cdots \oplus \mathcal{E}_{J_k} = \bigoplus_{i=1}^{k} \mathcal{E}_{J_i}$;

(b) $(\bigoplus_{i\in\{I\}} \mathcal{E}_{J_i})^{\perp} = \bigoplus_{i\notin\{I\}} \mathcal{E}_{J_i}$ where I is an index set containing some integers 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 λ_{J_i}.

(b) This is a consequence of applying Properties 7.2.7 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 $\mathcal{V}$ that contains the n orthonormal eigenvectors of the symmetric $[T]$, available due to the implication of Theorem 10.2.5.

(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) where $\vec{v}_{J_i} \in \mathcal{E}_{J_i}$. Then $T_{J_{i'}}(\vec{v}) = \vec{v}_{J_{i'}}$ due to the definition of an orthogonal projection and (b) where $\mathcal{E}_{J_{i'}}^{\perp} = \bigoplus_{i \neq i'} \mathcal{E}_{J_i}$. Using the same argument again 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 $\vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k}$, $\vec{v}_{J_i} \in \mathcal{E}_{J_i}$. Since T_{J_i} is an orthogonal projection onto $\mathcal{E}_{J_i}$, and also $\mathcal{E}_{J_i}^{\perp} = \bigoplus_{i \neq i'} \mathcal{E}_{J_i'} = \mathcal{N}(T_{J_i})$, we have $T_{J_i}(\vec{v}) = \vec{v}_{J_i}$. Thus $I(\vec{v}) = \vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k} = T_{J_1}(\vec{v}) + T_{J_2}(\vec{v}) + \cdots + T_{J_k}(\vec{v}) = (T_{J_1} + T_{J_2} + \cdots + T_{J_k})(\vec{v})$ and it must be that $I = T_{J_1} + T_{J_2} + \cdots + T_{J_k}$ by Properties 7.1.3.

(e) As usual, $\vec{v} = \vec{v}_{J_1} + \vec{v}_{J_2} + \cdots + \vec{v}_{J_k}$ and

$$\begin{aligned}
T(\vec{v}) &= T(\vec{v}_{J_1}) + T\vec{v}_{J_2} + \cdots + T(\vec{v}_{J_k}) \\
&= \lambda_{J_1}\vec{v}_{J_1} + \lambda_{J_2}\vec{v}_{J_2} + \cdots + \lambda_{J_k}\vec{v}_{J_k} && \text{(Definition 9.1.1 for eigenvalues/eigenvectors)} \\
&= \lambda_{J_1}T_{J_1}(\vec{v}) + \lambda_{J_2}T_{J_2}(\vec{v}) + \cdots + \lambda_{J_k}T_{J_k}(\vec{v}) && \text{(Just as (d))} \\
&= (\lambda_{J_1}T_{J_1} + \lambda_{J_2}T_{J_2} + \cdots + \lambda_{J_k}T_{J_k})(\vec{v})
\end{aligned}$$

and thus the desired relation follows, again by Properties 7.1.3.

□

The set of eigenvalues $\{\lambda_{J_i}\}$, $i = 1, 2, \ldots, k$, 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 a symmetric T, where the corresponding coordinates are given by simply projecting it onto each of these eigenspaces; The spectral decomposition in (e) means that any symmetric linear operator T, when applied on a vector $\vec{v}$, can be treated as equivalent to first decomposing $\vec{v}$ into components in the orthogonal coordinate system as suggested by (d), and then rescaling each of these components according to the spectrum (eigenvalues), which represents the "intensity" of T in the respective eigenspace. Since only projections and rescalings are involved, it implies that the symmetric T constitutes no rotation. Finally, notice that an orthogonal projection is a special case of (e) where the spectrum consists 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.[15]

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 is known as ***unitary matrices***.

Definition 10.4.1 (Unitary Matrix)**.** A complex matrix A is said to be unitary if

$$A^*A = AA^* = I \tag{10.10}$$

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 (Unitary Diagonalization)**.** 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 diagonal 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 matrix. However, the criterion is actually looser where the matrix only needs to be ***normal***.

---

[15] We will just show (d) here and leave (e) to the readers. By Equation (10.8b), R.H.S. of (d) reads

$$\begin{aligned}
T_{J_1} + T_{J_2} &= (\mathbf{v}_{J_1}^{(1)T}\mathbf{v}_{J_1}^{(1)T} + \mathbf{v}_{J_1}^{(2)T}\mathbf{v}_{J_1}^{(2)T}) + \mathbf{v}_{J_2}^{(1)T}\mathbf{v}_{J_2}^{(1)T} \\
&= \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix} + \begin{bmatrix}0\\-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} + \begin{bmatrix}0\\\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} \\
&= \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0 & 0\\0 & \frac{1}{2} & -\frac{1}{2}\\0 & -\frac{1}{2} & \frac{1}{2}\end{bmatrix} + \begin{bmatrix}0 & 0 & 0\\0 & \frac{1}{2} & \frac{1}{2}\\0 & \frac{1}{2} & \frac{1}{2}\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}
\end{aligned}$$

where the set indices J_1 and J_2 indicate the eigenvectors for the eigenvalues of $\lambda = 1$ and 3 respectively.

Definition 10.4.3 (Normal Matrix)**.** A matrix A is known as normal whenever

$$A^*A = AA^* \tag{10.11}$$

which may or may not be equal to the identity.

So normal is a less strict condition than 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 can happen on a normal matrix.

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 the eigenvectors are now $\mathbb{C}^n$, and orthogonality is defined with respect to the complex dot product (Definition 8.2.1), which has a complex conjugate applied on the second input vector. To show the equivalence between a matrix being unitarily diagonalizable and normal, 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 multiplicities, are located along the main diagonal of S, if the characteristic polynomial of A splits over $\mathbb{C}$ [$\mathbb{R}$].

Notice the requirement at the end of the theorem (see Footnote 6 of Chapter 9) and recall that every polynomial splits over $\mathbb{C}$. So in other words, every square matrix is *unitariliy similar* to an upper-triangular matrix. (But not always

orthogonally similar to a real diagonal matrix as the characteristic polynomial may not split over $\mathbb{R}$.[16])

Proof. We will use mathematical induction on the size of A to show the theorem for the complex case. If A is 1×1 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 λ_1, and a corresponding eigenvector $\vec{v}_\lambda^{(1)}$. Use the Gram-Schmidt process (with respect to the complex dot product now) to generate another set of $n - 1$ vectors $\vec{w}^{(2)}, \ldots, \vec{w}^{(n)}$ to complete an orthonormal basis for $\mathbb{C}^n$. The matrix formed by arranging these orthonormal basis vectors (normalized into unit length) in columns,

$$Q = \left[\vec{v}_\lambda^{(1)} | \vec{w}^{(2)} | \cdots | \vec{w}^{(n)}\right] = \left[\vec{v}_\lambda^{(1)} | W\right]$$

satisfies $Q^*Q = QQ^* = I$ and is unitary, so $Q^{-1} = Q^*$. A change of coordinates on A by Q as in Properties 7.2.2 (adapted for complex matrices) is

$$\begin{aligned}
Q^{-1}AQ = Q^*AQ &= Q^*A\left[\vec{v}_\lambda^{(1)} | W\right] \\
&= Q^*\left[A\vec{v}_\lambda^{(1)} | AW\right] \\
&= Q^*\left[\lambda_1\vec{v}_\lambda^{(1)} | AW\right] \qquad \text{(Definition 9.1.1 for an eigenvalue/eigenvector)} \\
&= \left[\frac{\vec{v}_\lambda^{(1)*}}{W^*}\right]\left[\lambda_1\vec{v}_\lambda^{(1)} | AW\right] \\
&= \left[\left[\frac{\vec{v}_\lambda^{(1)*}}{W^*}\right]\lambda_1\vec{v}_\lambda^{(1)} \;\middle|\; \left[\frac{\vec{v}_\lambda^{(1)*}}{W^*}\right]AW\right] \\
&= \begin{bmatrix} \lambda_1(\vec{v}_\lambda^{(1)*} \cdot \vec{v}_\lambda^{(1)}) & \vec{v}_\lambda^{(1)*}AW \\ \mathbf{0} & W^*AW \end{bmatrix}
\end{aligned}$$

[16]Its splitting over $\mathbb{R}$ is equivalent to all of the roots being real.

$$= \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} AW \\ \mathbf{0} & W' \end{bmatrix} \tag{10.12}$$

which is expressed in the form of a 2×2 block matrix. The leftmost column takes its form because $\vec{v}_\lambda^{(1)}$ is a unit vector and hence $\mathbf{v}_\lambda^{(1)*}\mathbf{v}_\lambda^{(1)} = 1$, and it is made orthogonal to other vectors in W by construct so the bottom left block is a zero matrix. Now we can use the induction hypothesis on the bottom right block $W' = W^*AW$ which is an $(n-1) \times (n-1)$ complex matrix, to claim that there exists a unitary matrix R so that $R^*W'R$ is upper-triangular. Then, consider

$$U = Q \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \tag{10.13}$$

This matrix is unitary (Definition 10.4.1), because

$$\begin{aligned}
U^*U &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} Q^*Q \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\
&= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} (I) \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} && \text{(Q is unitary, Definition 10.4.1)} \\
&= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\
&= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^*R \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & I_{n-1} \end{bmatrix} = I && \text{(By hypothesis, R is unitary)}
\end{aligned}$$

Subsequently,

$$\begin{aligned}
S = U^*AU &= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} Q^*AQ \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} \\
&= \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R^* \end{bmatrix} \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} AW \\ \mathbf{0} & W' \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix} && \text{(from (10.12) above)} \\
&= \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} AW \\ \mathbf{0} & R^*W' \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & R \end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} \lambda_1 & \vec{v}_\lambda^{(1)*} AWR \\ \mathbf{0} & R^* W' R \end{bmatrix} \tag{10.14}$$

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 that the diagonal elements of S are exactly its eigenvalues, which will be the same as those of A since they are unitarily similar (Properties 9.2.3 can be adapted to complex matrices too). □

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 such that $U^*AU = D$ 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

$$AA^* = (UDU^*)(UDU^*)^* = UD(U^*U)DU^* = UDD^*U^*$$

$U^*U = I$ by Definition 10.4.1 as U is unitary. But also $D^*D = DD^*$ since they are diagonal and commute, thus $A^*A = AA^*$. 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

$$\begin{aligned} S^*S = U^*A^*UU^*AU &= U^*A^*AU && (U \text{ is unitary}) \\ &= U^*AA^*U && (A \text{ is normal}) \\ &= U^*A(UU^*)A^*U && (\text{again, } U \text{ is unitary}) \end{aligned}$$

$$= (U^*AU)(U^*A^*U) = SS^*$$

With

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ 0 & s_{22} & s_{23} & \cdots & s_{2n} \\ 0 & 0 & s_{33} & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_{nn} \end{bmatrix} \quad \text{and} \quad S^* = \begin{bmatrix} \overline{s_{11}} & 0 & 0 & \cdots & 0 \\ \overline{s_{12}} & \overline{s_{22}} & 0 & \cdots & 0 \\ \overline{s_{13}} & \overline{s_{23}} & \overline{s_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{s_{1n}} & \overline{s_{2n}} & \overline{s_{3n}} & \cdots & \overline{s_{nn}} \end{bmatrix}$$

by considering

$$\begin{aligned} (S^*S)_{11} &= (SS^*)_{11} \\ \overline{s_{11}}s_{11} &= s_{11}\overline{s_{11}} + s_{12}\overline{s_{12}} + \cdots + s_{1n}\overline{s_{1n}} \\ 0 &= |s_{12}|^2 + \cdots + |s_{1n}|^2 \geq 0 \end{aligned}$$

we must have $s_{12} = \cdots = s_{1n} = 0$. Similarly, consider $(S^*S)_{22} = (SS^*)_{22}, \ldots,$ $(S^*S)_{nn} = (SS^*)_{nn}$ in sequential order, we conclude that $s_{ij} = 0$ when $i < j$, so all off-diagonal entries 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. The matrix A can be seen to be Hermitian, 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 a Hermitian matrix is not a mere coincidence and will be proved afterward. In fact, it is an extension of the fact that the eigenvalues of any symmetric matrix are all real. The eigenvectors can be found 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

$$U = \begin{bmatrix} \frac{-1-i}{\sqrt{3}} & \frac{1+i}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}$$

(The readers should check if U is really unitary.) Then the unitary diagonalization has the form of

$$U^* A U = D$$

$$\begin{bmatrix} \frac{-1+i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1-i}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 1 & 1+i \\ 1-i & 2 \end{bmatrix} \begin{bmatrix} \frac{-1-i}{\sqrt{3}} & \frac{1+i}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix}$$

□

Properties 10.4.6. The eigenvalues of any Hermitian (including symmetric) matrix $A = A^*$ must be real.

Proof. Consider $\vec{v}_\lambda \cdot (A\vec{v}_\lambda)$ where $\vec{v}_\lambda$ is a complex eigenvector of A and the complex dot product is used in place. Then

$$\begin{aligned} \vec{v}_\lambda \cdot (A\vec{v}_\lambda) &= \vec{v}_\lambda \cdot (\lambda\vec{v}_\lambda) && \text{(Definition 9.1.1)} \\ &= \overline{\lambda}(\vec{v}_\lambda \cdot \vec{v}_\lambda) && \text{(Properties 8.2.3)} \\ &= \overline{\lambda}\|\vec{v}_\lambda\|^2 \end{aligned}$$

But also

$$\begin{aligned} \vec{v}_\lambda \cdot (A\vec{v}_\lambda) &= (A^*\vec{v}_\lambda) \cdot \vec{v}_\lambda && \text{(Properties 8.2.7)} \\ &= (A\vec{v}_\lambda) \cdot \vec{v}_\lambda && (A \text{ is Hermitian, Definition 8.2.5}) \end{aligned}$$

$$
\begin{aligned}
&= (\lambda\vec{v}_\lambda)\cdot\vec{v}_\lambda && \text{(Definition 9.1.1)}\\
&= \lambda(\vec{v}_\lambda\cdot\vec{v}_\lambda) && \text{(Properties 8.2.3)}\\
&= \lambda\|\vec{v}_\lambda\|^2
\end{aligned}
$$

Therefore, $\overline{\lambda}\|\vec{v}_\lambda\|^2 = \lambda\|\vec{v}_\lambda\|^2$, and since $\vec{v}_\lambda \neq \mathbf{0}$, $\|\vec{v}_\lambda\| \neq 0$, $\overline{\lambda} = \lambda$ and the eigenvalue λ has to be real. □

Example 10.4.2. Check that

$$A = \begin{bmatrix} \frac{5}{3} - \frac{1}{3}i & \frac{1}{3} + i \\ -1 + \frac{1}{3}i & \frac{4}{3} + \frac{1}{3}i \end{bmatrix}$$

is normal and carry out unitary diagonalization on it.

Solution. First,

$$A^*A = \begin{bmatrix} \frac{5}{3} + \frac{1}{3}i & -1 - \frac{1}{3}i \\ \frac{1}{3} - i & \frac{4}{3} - \frac{1}{3}i \end{bmatrix} \begin{bmatrix} \frac{5}{3} - \frac{1}{3}i & \frac{1}{3} + i \\ -1 + \frac{1}{3}i & \frac{4}{3} + \frac{1}{3}i \end{bmatrix} = \begin{bmatrix} 4 & -1+i \\ -1-i & 3 \end{bmatrix}$$

and

$$AA^* = \begin{bmatrix} \frac{5}{3} - \frac{1}{3}i & \frac{1}{3} + i \\ -1 + \frac{1}{3}i & \frac{4}{3} + \frac{1}{3}i \end{bmatrix} \begin{bmatrix} \frac{5}{3} + \frac{1}{3}i & -1 - \frac{1}{3}i \\ \frac{1}{3} - i & \frac{4}{3} - \frac{1}{3}i \end{bmatrix} = \begin{bmatrix} 4 & -1+i \\ -1-i & 3 \end{bmatrix} = A^*A$$

So A is indeed normal. Now the characteristic equation is

$$
\begin{aligned}
(\frac{5}{3} - \frac{1}{3}i - \lambda)(\frac{4}{3} + \frac{1}{3}i - \lambda) - (-1 + \frac{1}{3}i)(\frac{1}{3} + i) &= 0\\
\lambda^2 - 3\lambda + (3+i) &= 0
\end{aligned}
$$

which can be checked to have the complex roots of $\lambda_1 = 1 + i$ and $\lambda_2 = 2 - i$ as the two eigenvalues. The corresponding orthonormal eigenvectors are $\vec{v}_{\lambda_1} = (\frac{1}{\sqrt{3}}, \frac{1+i}{\sqrt{3}})^T$ and $\vec{v}_{\lambda_2} = (\frac{-1+i}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$, and thus the unitary diagonalization reads

$$D = U^*AU$$

$$
\begin{aligned}
&= \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1-i}{\sqrt{3}} \\ \frac{-1-i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{5}{3} - \frac{1}{3}i & \frac{1}{3} + i \\ -1 + \frac{1}{3}i & \frac{4}{3} + \frac{1}{3}i \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{-1+i}{\sqrt{3}} \\ \frac{1+i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \\
&= \begin{bmatrix} 1+i & 0 \\ 0 & 2-i \end{bmatrix}
\end{aligned}
$$

where

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

is unitary. □

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

A1 = np.array([[1., 0., 0.],
               [0., 2., 1.],
               [0., 1., 2.]])
A1_sympy = sympy.Matrix(A1)
print(A1_sympy.is_symmetric())
P, D1 = A1_sympy.diagonalize()
print(P, D1)
```

which returns `true` for the symmetricity checking 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 slightly 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 the exact but close equality against I over all entries. The same idea applies to unitary diagonalization and let's use Example 10.4.2 to demonstrate. First, we check if the matrix is normal by

```
A2 = np.array([[5/3 - (1/3)*1j, 1/3 + 1j],
               [-1 + (1/3)*1j, 4/3 + (1/3)*1j]])

print(np.allclose(np.conjugate(A2).T @ A2, A2 @ np.conjugate(
    A2).T))
```

which outputs `true`. Next,

```
A2_sympy = sympy.Matrix(A2)

U, D = A2_sympy.diagonalize()
print(U, D)
print(np.allclose(np.array(U.H @ U, dtype=complex), np.
    identity(2)))
```

gives the matrices U and D:

```
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 returns `true` as we check if U is unitary. The form of U may seem to be different from Example 10.4.2 but they are actually equivalent, only differed by a complex multiplicative factor. To see this, we can do

```
U_np = np.array(U, dtype=complex)
U_np[:,0] = U_np[:,0] * np.conjugate(U_np[1,0])
U_np[:,1] = U_np[:,1] * np.conjugate(U_np[0,1])
print(U_np)
```

which produces

```
[[-0.33333333+3.33333333e-01j  0.33333333-2.58092536e-18j]
 [ 0.33333333-3.97842146e-19j  0.33333333+3.33333333e-01j]]
```

which is essentially the same as in Example 10.4.2 but with the two columns of eigenvectors interchanged plus some very small round-off errors.

10.6. Exercises

Exercise 10.1 Determine if the following matrices are orthogonal. If so, also determine if they represent a rotation or reflection.

(a) $\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{3}{2\sqrt{6}} & \frac{1}{2} & -\frac{3}{2\sqrt{6}} \\ -\frac{1}{2\sqrt{2}} & \frac{\sqrt{3}}{2} & \frac{1}{2\sqrt{2}} \end{bmatrix}$

(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 the old and new coordinate systems ($\vec{v}_0$ 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 the other two eigenvalues is 18. Using the knowledge that the trace and characteristic polynomial are invariants, i.e. remain 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 produce 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 $x + 2y - 2z + 4w = 0$ and $3x - y + z - 2w = 0$.

Exercise 10.8 Show that every projection/idempotent matrix A satisfies the following condition:

$$\mathrm{tr}(A) = \mathrm{rank}(A)$$

where tr() stands for the trace.

Exercise 10.9 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 10.10 Show that the matrix below is Hermitian and unitarily diagonalize it.

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

Exercise 10.11 *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 the 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, θ are some arbitrary constants to be determined from the initial condition. Consider the situation shown in the given figure. 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 rewrite them into matrix form. Carry out orthogonal diagonalization to simplify the equations and find the general solutions for the motions.

Figure 10.5.: *The scenario of springs and masses in Exercise 10.11.*

Chapter 11

Quadratic Forms

Symmetric matrices often see their appearance in the areas of Geometry, Statistics, Physics, and more, as *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 Science will be obtaining a *covariance matrix* between different variables, eventually leading to *Principal Component Analysis (PCA)*, which breaks down the variables into uncorrelated modes that explain the spread of their distribution. 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 with *quadratic equations* in the form of $y = ax^2 + bx + c$. ***Quadratic forms*** are the generalization of quadratic

equations when there are multiple variables $x_1, x_2, x_3, \ldots$: The possible quadratic forms will be made up of the usual quadratic terms $x_1^2, x_2^2, x_3^2, \ldots$, as well as the *cross-product terms* (again, not to be confused with the vector cross product in Section 4.2.2) $x_p x_q$, $p \neq q$. 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, finite-dimensional vector spaces first.

Definition 11.1.1 (Quadratic Form)**.** Real quadratic forms in multiple variables $\vec{x} = (x_1, x_2, x_3, \ldots, x_n)^T$ has a structure of

$$Q(\vec{x}) = \sum_{p=1}^{n} \sum_{q=1}^{n} b_{pq} x_p x_q \tag{11.1}$$

Note that b_{pq} are real and they produce a real scalar.

For example, in a two-variable situation, $x^2+3xy+y^2$ and $3x^2-4xy$ are quadratic forms, while $x^2 + y$ and $xy + xy^2$ are not. Notice that $x_p x_q$ and $x_q x_p$ are actually the same term, and we will replace b_{pq} or b_{qp} by $\frac{1}{2}(b_{pq} + b_{qp})$ as a single coefficient for both of them.

Properties 11.1.2 (Matrix Representation of a Quadratic Form)**.** All quadratic forms as given by (11.1) in Definition 11.1.1 for any real vector $\vec{x} = (x_1, x_2, x_3, \ldots, x_n)^T \in \mathcal{V}$ can be expressed as

$$Q : \mathcal{V} \to \mathbb{R},\ Q(\vec{x}) = \mathbf{x}^T B \mathbf{x} \tag{11.2}$$

where B is real symmetric and has the form of

$$\begin{bmatrix} b_{11} & \frac{1}{2}(b_{12}+b_{21}) & \frac{1}{2}(b_{13}+b_{31}) & \cdots & \frac{1}{2}(b_{1n}+b_{n1}) \\ \frac{1}{2}(b_{12}+b_{21}) & b_{22} & \frac{1}{2}(b_{23}+b_{32}) & & \frac{1}{2}(b_{2n}+b_{n2}) \\ \frac{1}{2}(b_{13}+b_{31}) & \frac{1}{2}(b_{23}+b_{32}) & b_{33} & \cdots & \frac{1}{2}(b_{3n}+b_{n3}) \\ \vdots & & \vdots & \ddots & \vdots \\ \frac{1}{2}(b_{1n}+b_{n1}) & \frac{1}{2}(b_{2n}+b_{n2}) & \frac{1}{2}(b_{3n}+b_{n3}) & \cdots & b_{nn} \end{bmatrix} \tag{11.3}$$

The readers can check that the above matrix expression indeed leads to the desired quadratic form by a direct expansion.[1] 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.[2]

Even if we are given $\mathbf{x}^T A\mathbf{x}$ where A is not symmetric to begin with, we can extract the symmetric part of A (see Exercise 2.8), that is, $B = \frac{1}{2}(A + A^T)$. Reconstruction of the quadratic form by $\mathbf{x}^T B\mathbf{x}$ will still be equivalent to the original $\mathbf{x}^T A\mathbf{x}$:

$$\begin{aligned} \mathbf{x}^T B\mathbf{x} &= \frac{1}{2}\mathbf{x}^T (A + A^T)\mathbf{x} \\ &= \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{1}{2}\mathbf{x}^T A^T\mathbf{x} \end{aligned}$$

1

$$\begin{aligned}
&\begin{bmatrix} x_1 & x_2 & x_3 & \cdots \end{bmatrix} \begin{bmatrix} b_{11} & \frac{1}{2}(b_{12}+b_{21}) & \frac{1}{2}(b_{13}+b_{31}) & \cdots \\ \frac{1}{2}(b_{12}+b_{21}) & b_{22} & \frac{1}{2}(b_{23}+b_{32}) & \\ \frac{1}{2}(b_{13}+b_{31}) & \frac{1}{2}(b_{23}+b_{32}) & b_{33} & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & x_3 & \cdots \end{bmatrix} \begin{bmatrix} b_{11}x_1 + \frac{1}{2}(b_{12}+b_{21})x_2 + \frac{1}{2}(b_{13}+b_{31})x_3 + \cdots \\ \frac{1}{2}(b_{12}+b_{21})x_1 + b_{22}x_2 + \frac{1}{2}(b_{23}+b_{32})x_3 + \cdots \\ \frac{1}{2}(b13+b_{31})x_1 + \frac{1}{2}(b_{23}+b_{32})x_2 + b_{33}x_3 + \cdots \\ \vdots \end{bmatrix} \\
&= x_1\left(b_{11}x_1 + \frac{1}{2}(b_{12}+b_{21})x_2 + \frac{1}{2}(b_{13}+b_{31})x_3 + \cdots\right) \quad = b_{11}x_1^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + \cdots \\
&\quad + x_2\left(\frac{1}{2}(b_{12}+b_{21})x_1 + b_{22}x_2 + \frac{1}{2}(b_{23}+b_{32})x_3 + \cdots\right) \quad + b_{21}x_2x_1 + b_{22}x_2^2 + b_{23}x_2x_3 + \cdots \\
&\quad + x_3\left(\frac{1}{2}(b_{13}+b_{31})x_1 + \frac{1}{2}(b_{23}+b_{32})x_2 + b_{33}x_3 + \cdots\right) \quad + b_{31}x_3x_1 + b_{32}x_3x_2 + b_{33}x_3^2 + \cdots
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{1}{2}(\mathbf{x}^T A^T \mathbf{x})^T && (\mathbf{x}^T A^T \mathbf{x} = (\mathbf{x}^T A^T \mathbf{x})^T \text{ since it is just} \\
& && \text{a scalar in a } 1 \times 1 \text{ singleton block}) \\
&= \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \frac{1}{2}\mathbf{x}^T A\mathbf{x} && \text{(Properties 2.1.4)} \\
&= \mathbf{x}^T A\mathbf{x}
\end{aligned}
$$

Similarly the skew-symmetric part does not contribute anything to the quadratic form[3], and we will characterize all quadratic forms using symmetric matrices henceforth. Such a symmetric matrix B is also frequently considered to behave as a ***symmetric bilinear form*** in a more general setting.

Definition 11.1.3 ((Symmetric) Bilinear Form)**.** A real bilinear form $B(\vec{x}, \vec{y}) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ over a real vector space $\mathcal{V}$ takes two vectors $\vec{x}, \vec{y} \in \mathcal{V}$ from it as inputs and returns a real scalar. A symmetric bilinear form is one that satisfies

$$B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x}) \tag{11.4}$$

for any pair of $\vec{x}$ and $\vec{y}$. For finite-dimensional cases, the expression becomes

$$B(\vec{x}, \vec{y}) = \mathbf{x}^T B\mathbf{y} \tag{11.5}$$

where $B^T = B$ is an $n \times n$ real symmetric matrix and n is the dimension of $\mathcal{V}$.

It is not hard to see that the form $\mathbf{x}^T B\mathbf{y}$ fulfills the requirement of $B(\vec{x}, \vec{y}) =$

---

[2]

$$
\begin{aligned}
\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} x - y \\ -x + 3y \end{bmatrix} \\
&= x(x - y) + y(-x + 3y) \\
&= x^2 - xy - xy + 3y^2 = x^2 - 2xy + 3y^2
\end{aligned}
$$

[3] $\frac{1}{2}\mathbf{x}^T(A - A^T)\mathbf{x} = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \frac{1}{2}\mathbf{x}^T A^T\mathbf{x} = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \frac{1}{2}(\mathbf{x}^T A^T\mathbf{x})^T = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \frac{1}{2}\mathbf{x}^T A\mathbf{x} = 0$

$B(\vec{y}, \vec{x})$.[4] It is also not hard to see that the pattern of quadratic forms proposed in Properties 11.1.2 can be derived from a symmetric bilinear form as we set $\vec{y} = \vec{x}$. In this sense, we say that a quadratic form is characterized by the symmetric bilinear form induced by the appropriate symmetric matrix B.

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 outputs 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 relaxes the restriction such that the quadratic form can also return 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 that characterizes the quadratic form.

Definition 11.1.4. For any real quadratic form $Q(x) = \mathbf{x}^T B\mathbf{x}$, Q (or B) is called

(a) positive-definite, if for any $\vec{x} \neq \vec{0}$, $\mathbf{x}^T B\mathbf{x} > 0$ (positive-semidefinite if $\mathbf{x}^T B\mathbf{x} \geq 0$);

(b) negative-definite, if for any $\vec{x} \neq \vec{0}$, $\mathbf{x}^T B\mathbf{x} < 0$ (negative-semidefinite if $\mathbf{x}^T B\mathbf{x} \leq 0$);

(c) indefinite if $\mathbf{x}^T B\mathbf{x}$ can take both positive and negative values.

Theorem 11.1.5. The quadratic form $Q(x) = \mathbf{x}^T B\mathbf{x}$ where B is real symmetric, is

(a) positive-definite, if and only if all eigenvalues of B are positive (positive-

[4] $B(\vec{x}, \vec{y}) = \mathbf{x}^T B\mathbf{y} = (\mathbf{x}^T B\mathbf{y})^T = \mathbf{y}^T B^T \mathbf{x} = \mathbf{y}^T B\mathbf{x} = B(\vec{y}, \vec{x})$, where $\mathbf{x}^T B\mathbf{y} = (\mathbf{x}^T B\mathbf{y})^T$ as it is just a single real number.

semidefinite 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-semidefinite 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 10.3.4 naturally comes off as handy. Assume we are working in an 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, we have

$$B\mathbf{x} = \lambda_{J_1}\vec{x}_{J_1} + \lambda_{J_2}\vec{x}_{J_2} + \cdots + \lambda_{J_k}\vec{x}_{J_k}$$

Subsequently,

$$\begin{aligned}
Q(x) &= \mathbf{x}^T B\mathbf{x} \\
&= (\vec{x}_{J_1} + \vec{x}_{J_2} + \cdots + \vec{x}_{J_k}) \cdot (\lambda_{J_1}\vec{x}_{J_1} + \lambda_{J_2}\vec{x}_{J_2} + \cdots + \lambda_{J_k}\vec{x}_{J_k}) \\
&= \lambda_{J_1}(\vec{x}_{J_1} \cdot \vec{x}_{J_1}) + \lambda_{J_2}(\vec{x}_{J_2} \cdot \vec{x}_{J_2}) + \cdots + \lambda_{J_k}(\vec{x}_{J_k} \cdot \vec{x}_{J_k}) \\
&= \lambda_{J_1}\|\vec{x}_{J_1}\|^2 + \lambda_{J_2}\|\vec{x}_{J_2}\|^2 + \cdots + \lambda_{J_k}\|\vec{x}_{J_k}\|^2
\end{aligned}$$

where $\vec{x}_{J_i} \cdot \vec{x}_{J_i'} = 0$ 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. □

Example 11.1.1. Show that the symmetric matrix

$$B = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$$

is positive-definite.

Solution. By Theorem 11.1.5, we simply check whether all 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:

$$\begin{aligned} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= 4x^2 - 2xy + 4y^2 \\ &= (x^2 - 2xy + y^2) + 3x^2 + 3y^2 \\ &= (x-y)^2 + 3x^2 + 3y^2 > 0 \end{aligned}$$

which is always positive as long as x and y are not both zeros. □

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 a coordinate transformation works for a quadratic/symmetric bilinear form. 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 hence it is reasonable that the rule of coordinate transformation will be

somehow different when the matrix acts like a quadratic form instead. Let's take a step back and consider the change of coordinates for a vector in (7.7) under Theorem 7.1.14: $[\vec{x}]_\beta = P^\beta_{\beta'}[\vec{x}']_{\beta'}$, hence

$$\begin{aligned} Q(\vec{x}) &\equiv [\vec{x}]_\beta^T B[\vec{x}]_\beta \\ &= (P^\beta_{\beta'}[\vec{x}']_{\beta'})^T B(P^\beta_{\beta'}[\vec{x}']_{\beta'}) \\ &= [\vec{x}']^T_{\beta'}((P^\beta_{\beta'})^T BP^\beta_{\beta'})[\vec{x}']_{\beta'} = [\vec{x}']^T_{\beta'}B'[\vec{x}']_{\beta'} \end{aligned}$$

so we identify the coordinate transformation of a quadratic form by $B' = P^T BP$ where P is some invertible coordinate basis matrix. In this case, B' and B are known as ***congruent***.

Definition 11.1.6. The coordinate transformation of a symmetric matrix B as a quadratic form follows

$$B' = P^T BP \tag{11.6}$$

where P is invertible and consists of 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, previously during orthogonal diagonalization, we have $P^{-1} = P^T$, and hence the corresponding coordinate transformation of a symmetric matrix to its orthonormal eigenbasis 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 $D = P^T BP$ where the columns of P are all the orthonormal eigenvectors of B (Properties 10.2.5)[5]. Furthermore, such a diagonal matrix D contains the eigenvalues of B and let's say r of them are positive: $\lambda_1^+, \lambda_2^+, \cdots, \lambda_r^+$, s of them are negative $\lambda_1^-, \lambda_2^-, \cdots, \lambda_s^-$, and the

[5]Notice that this is not the only way to make a quadratic form diagonal (unlike a linear transformation, dictated by its eigenvectors) and there exist many other P that can do it. (However, an important observation about them is Theorem 11.1.8, to be introduced soon.)

remaining eigenvalues are zeros. Arranging them in the order of positive-negative-zero, and via an extra real diagonal factor matrix F, where

$$F_{kk} = \begin{cases} \frac{1}{\sqrt{\lambda_k^+}} & 1 \le k \le r \\ \frac{1}{\sqrt{-\lambda_k^-}} & r+1 \le k \le r+s \\ 1 & r+s+1 \le k \le n \end{cases} \tag{11.7}$$

it is easy to see that F is invertible, and we can further transform the quadratic form into

$$\begin{aligned} C &= F^T D F \\ &= \begin{bmatrix} \frac{1}{\sqrt{\lambda_1^+}} & & & & & & & & \\ & \ddots & & 0 & & & & & \\ & & \frac{1}{\sqrt{\lambda_r^+}} & & & & & & \\ & 0 & & \frac{1}{\sqrt{-\lambda_1^-}} & & & 0 & & \\ & & & & \ddots & & & & \\ & & & 0 & & \frac{1}{\sqrt{-\lambda_s^-}} & & & \\ & & & & & & 1 & & \\ & & & & & 0 & & \ddots & \end{bmatrix}^T \begin{bmatrix} \lambda_1^+ & & & & & & & & \\ & \ddots & & 0 & & & & & \\ & & \lambda_r^+ & & & & & & \\ & 0 & & \lambda_1^- & & & 0 & & \\ & & & & \ddots & & & & \\ & & & 0 & & \lambda_s^- & & & \\ & & & & & & 0 & & \\ & & & & & 0 & & \ddots & \end{bmatrix} \\ &\quad \begin{bmatrix} \frac{1}{\sqrt{\lambda_1^+}} & & & & & & & & \\ & \ddots & & 0 & & & & & \\ & & \frac{1}{\sqrt{\lambda_r^+}} & & & & & & \\ & 0 & & \frac{1}{\sqrt{-\lambda_1^-}} & & & 0 & & \\ & & & & \ddots & & & & \\ & & & 0 & & \frac{1}{\sqrt{-\lambda_s^-}} & & & \\ & & & & & & 1 & & \\ & & & & & 0 & & \ddots & \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 1 & & 0 & & & & & \\ & \ddots & & & & & & \\ 0 & & 1 & 0 & & & & \\ & & 0 & -1 & & 0 & & \\ & & & & \ddots & & & \\ & & & 0 & & -1 & 0 & \\ & & & & & 0 & 0 & 0 \\ & & & & & & 0 & \ddots \end{bmatrix} = \begin{bmatrix} I_r & & \\ & -I_s & \\ & & [\mathbf{0}] \end{bmatrix}$$

which is known as the ***canonical quadratic form*** for B. To summarize, the matrix product PF converts B into such a form by $(PF)^T B(PF) = F^T(P^T BP)F = F^T DF = C$, and therefore B is congruent to its canonical quadratic form. The following theorem shows that two different canonical quadratic forms cannot be congruent, and hence the canonical quadratic form of any matrix is unique.

Theorem 11.1.7. Two canonical quadratic forms of the same extent n

$$C = \begin{bmatrix} I_r & & \\ & -I_s & \\ & & [\mathbf{0}] \end{bmatrix} \qquad C' = \begin{bmatrix} I_{r'} & & \\ & -I_{s'} & \\ & & [\mathbf{0}] \end{bmatrix}$$

are congruent if and only if $r = r'$ and $s = s'$. As a corollary, this shows the uniqueness of the canonical quadratic form for any quadratic form.

Proof. The "if" part is trivial. For the "only if" part, without loss of generality, assume $r' > r$. If the congruence relation $C' = P^T CP$ has to hold where P is some matrix, then consider

$$\mathbf{e}^{(j)T} C' \mathbf{e}^{(j)} = 1 > 0 \qquad \text{where } 1 \leq j \leq r'$$

which is also equal to

$$\mathbf{e}^{(j)T} P^T CP \mathbf{e}^{(j)} = \mathbf{p}^{(j)T} C \mathbf{p}^{(j)}$$

where $P = \left[\mathbf{p}^{(1)} | \cdots | \mathbf{p}^{(n)}\right]$ is consisted of n column vectors $\mathbf{p}^{(j)}$ and $P\mathbf{e}^{(j)} = \mathbf{p}^{(j)}$. We claim that there exists a non-trivial linear combination of $\mathbf{p}^{(j)}$, $1 \le j \le r'$, i.e. $\mathbf{q} = \sum_{j=1}^{r'} c_j \mathbf{p}^{(j)}$, such that $\mathbf{q}_i = 0$ for $1 \le i \le r$.[6] Subsequently, consider $\mathbf{x} = \sum_{j=1}^{r'} c_j \mathbf{e}^{(j)}$, and

$$\begin{aligned}
\mathbf{x}^T C' \mathbf{x} &= \begin{bmatrix} c_1 & \cdots & c'_r & 0 & \cdots \end{bmatrix} \begin{bmatrix} I_{r'} & & \\ & -I_{s'} & \\ & & [\mathbf{0}] \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c'_r \\ 0 \\ \vdots \end{bmatrix} \\
&= c_1^2 + \cdots + c_r'^2 = \sum_{j=1}^{r'} c_j^2 > 0
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{x}^T P^T C P \mathbf{x} &= (P\mathbf{x})^T C P \mathbf{x} \\
&= \mathbf{q}^T C \mathbf{q} \\
&= \begin{bmatrix} 0 & \cdots & 0 \text{ (the } r\text{-th entry)} & * \end{bmatrix} \begin{bmatrix} I_r & & \\ & -I_s & \\ & & [\mathbf{0}] \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \text{ (the } r\text{-th entry)} \\ * \end{bmatrix} \le 0
\end{aligned}$$

[6]The corresponding system is

$$\begin{bmatrix} \mathbf{p}_1^{(1)} & \cdots & \mathbf{p}_1^{(r)} & \cdots & \mathbf{p}_1^{(r')} \\ \vdots & & \vdots & & \vdots \\ \mathbf{p}_r^{(1)} & \cdots & \mathbf{p}_r^{(r)} & \cdots & \mathbf{p}_r^{(r')} \\ \vdots & & \vdots & & \vdots \\ \mathbf{p}_n^{(1)} & \cdots & \mathbf{p}_n^{(r)} & \cdots & \mathbf{p}_n^{(r')} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ \vdots \\ c_{r'} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \text{ (the } r\text{-th entry)} \\ * \end{bmatrix}$$

The part below the r-th row does not matter since the constraints are for the first r rows and so it is effective a $r \times r'$ 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, $1 \le j \le r'$.

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 $r' = r$.[7] By the same logic, we have $s = s'$ as well. □

An immediate result from this is ***Sylvester's Law of Inertia***.

Theorem 11.1.8 (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, and vice versa.

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 11.1.7, 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 and both are congruent to it. □

Example 11.1.2. Show that

$$B = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix} \qquad \text{and} \qquad B' = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$$

are congruent.

Solution. By Sylvester's Law of Inertia above, we simply count if the two (symmetric) quadratic forms have the same numbers of positive/negative/zero

[7] The same argument in opposite direction will show that it is also not possible to have $r' < r$.

eigenvalues as they will be the diagonal entries when converted via orthogonal diagonalization. The eigenvalues of B are 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 solution is

$$\begin{aligned} \lambda &= \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(-9)}}{2} \\ &= 1 \pm \sqrt{10} \end{aligned}$$

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.[8] □

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* (see Figure 11.1). They can be described by the general equation form as below.

Definition 11.1.9 (Conic Sections)**.** Conic Sections (circles, ellipses, parabola, hyperbola) are the curves generated by a second-degree polynomial in two variables (x, y) that take the general form of

$$ax^2 + bxy + cy^2 + mx + ny - h = 0 \tag{11.8}$$

[8] One possible choice of P as in $B' = P^T BP$ is $P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -1 \\ 0 & \frac{2}{3} \end{bmatrix}$.

Figure 11.1.: *The creation of conic sections. (Adapted from the code of Ridlo W. Wibowo)*

where a, b, c, m, n and h are all constants. It can be expressed as a quadratic form:

$$\mathbf{x}^T B\mathbf{x} = \begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} a & \frac{b}{2} & \frac{m}{2} \\ \frac{b}{2} & c & \frac{n}{2} \\ \frac{m}{2} & \frac{n}{2} & -h \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0 \tag{11.9}$$

where $\mathbf{x}^T = (x, y, 1)^T$.

To see what type of conic sections a quadratic form represents, we can examine the determinants of B and its minor B_{33}. We simply state the results below.

Properties 11.1.10. The quadratic form constructed in Definition 11.1.9 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 a parabola, only keeping the ellipses and hyperbola. The quadratic form can then be simplified as follows.

Properties 11.1.11 (Central Conics)**.** Ellipses (plus circles) and hyperbola, centered at the origin, are called *central conics* and have the form of

$$ax^2 + bxy + cy^2 = h \tag{11.10}$$

or can be expressed as a quadratic form of

$$\mathbf{x}^T B_{33}\mathbf{x} = h \tag{11.11}$$

where now $\mathbf{x} = (x, y)^T$ only and B_{33} is as defined in Properties 11.1.10.

By Properties 11.1.10, 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 $x^2 + xy - y^2 = 1$.[9]

Notice that sometimes an "ellipse" where $\det(B_{33}) > 0$ may not produce a 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 11.1.12. Given $\mathbf{x}^T B_{33}\mathbf{x} = h$ as of (11.11) in Properties 11.1.11, where h is chosen to be 1 for the scaling, then it represents

- an ellipse if B_{33} is positive-definite,
- a hyperbola if B_{33} is indefinite,
- 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 viewpoint of orthogonal diagonalization the 2×2 B_{33} matrix must have one positive and one negative eigenvalue, which by Theorem 11.1.5 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. 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 just $x^2 - 4xy + 3y^2 = 1$. 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 11.1.12.

Figure 11.2 shows that hyperbola and ellipses can be rotated from their *standard position*. The effect on a quadratic equation resulting from the coordinate

Figure 11.2.: *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.*

transformation 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 θ, we can make a rotation by the same angle θ but in an opposite direction to recover the standard position. It is equivalent to rotating the coordinate system by an angle of θ in the same sense as the initial tilting. The readers can refer back to Section 10.1.2 and Definition 11.1.6 about the rotation of a coordinate system for a quadratic form.

[9]The first one is an ellipse ($\Delta = (-1)^2 - 4(1)(2) = -7 < 0$) and the second one is a hyperbola ($\Delta = (1)^2 - 4(1)(-1) = 5 > 0$).

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, and 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 11.1.12, 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

$$(P\mathbf{x}')^T B(P\mathbf{x}') = \mathbf{x}'^T (P^T BP)\mathbf{x}'$$

We immediately identify $P^T BP$ as a rotation of the coordinate system for the matrix B, as noted by Definition 11.1.6. 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

$$\vec{v}_\lambda = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ for } \lambda = \frac{1}{2} \qquad \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ for } \lambda = \frac{3}{2}$$

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

$$P^T BP = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} = D$$

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 graphs 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 $\pi/4 = 45°$,

(a) *Before Rotation.* (b) *After Rotation.*

Figure 11.3.: *Illustration for Example 11.1.3.*

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 11.1.13 (Principal Axes Theorem)**.** For a quadratic form $\mathbf{x}^T B\mathbf{x}$, where B is a real symmetric matrix, we can always make an orthogonal change of variables $\mathbf{x}' = P^T\mathbf{x}$ (or equivalently $\mathbf{x} = P\mathbf{x}'$) such that it turns into $\mathbf{x}'^T D\mathbf{x}' = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \cdots$ 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 θ such that

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix} \tag{11.12}$$

the off-diagonal elements become zero. The required θ is found by expanding the L.H.S. and equating both sides, which gives

$$\begin{aligned} -\sin\theta(a\cos\theta + \frac{b}{2}\sin\theta) + \cos\theta(\frac{b}{2}\cos\theta + c\sin\theta) &= 0 \\ \frac{c-a}{2}\sin(2\theta) + \frac{b}{2}\cos(2\theta) &= 0 \\ \cot(2\theta) &= \frac{a-c}{b} \end{aligned} \tag{11.13}$$

where we have applied the familiar double-angle formulae from the first 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 3×3 symmetric matrix. Now the quadratic shapes include *ellipsoids* and *hyperboloids*, and the change of 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 11.3.

11.1.4. Hermitian Forms

The concept of quadratic forms/symmetric bilinear forms can be readily promoted to complex vector spaces. Since Hermiticity is the complex equivalent of symmetricity, it is not surprising that we have the ***Hermitian forms*** as the complex counterpart of symmetric bilinear forms.

Definition 11.1.14 (Hermitian Form)**.** A (complex) Hermitian form $H(\vec{x}, \vec{y}) : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ takes two vectors $\vec{x}, \vec{y} \in \mathcal{V}$ from a complex vector space and returns a complex scalar, that satisfies

$$H(\vec{x}, \vec{y}) = \overline{H(\vec{y}, \vec{x})} \tag{11.14}$$

for any pair of $\vec{x}$ and $\vec{y}$ (aware of the conjugation!). For finite-dimensional cases, it takes the general form of

$$H(\vec{x}, \vec{y}) = \mathbf{x}^T \overline{H\mathbf{y}} \tag{11.15}$$

where $H^* = H$ is an $n \times n$ Hermitian matrix and n is the dimension of $\mathcal{V}$.

Note that the second argument is conjugated just like the complex dot product given in Definition 8.2.1. Due to Properties 10.4.6, the eigenvalues of a

Hermitian form/matrix H are always real, hence the logic of Theorem 11.1.5 is still valid and it can be applied when $\vec{x} = \vec{y}$. We simply note the transferred results below.

Theorem 11.1.15. The Hermitian form $H(\vec{x}) = \mathbf{x}^T \overline{H\mathbf{x}}$ by (11.15) where the two input complex vectors are now identical and H is Hermitian, is

(a) positive-definite, if and only if all eigenvalues of H are positive (positive-semidefinite if and only if all eigenvalues of H are non-negative;

(b) negative-definite, if and only if all eigenvalues of H are negative (negative-semidefinite if and only if all eigenvalues of H are non-positive);

(c) indefinite when there are both positive and negative eigenvalues for H.

Example 11.1.4. Show that the Hermitian form characterized by

$$H = \begin{bmatrix} -3 & 1+\sqrt{3}i \\ 1-\sqrt{3}i & -3 \end{bmatrix}$$

is negative-definite.

Solution. The readers should check that H is indeed Hermitian. By Theorem 11.1.15, we just need to show that the eigenvalues of H are all negative. So we now solve the characteristic polynomial

$$\begin{aligned} &\begin{vmatrix} -3-\lambda & 1+\sqrt{3}i \\ 1-\sqrt{3}i & -3-\lambda \end{vmatrix} \\ &= (-3-\lambda)(-3-\lambda) - (1-\sqrt{3}i)(1+\sqrt{3}i) \\ &= (9+6\lambda+\lambda^2) - 4 \\ &= \lambda^2 + 6\lambda + 5 = (\lambda+1)(\lambda+5) \end{aligned}$$

which yields two negative roots $\lambda = -1, -5$ and we are done. □

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. The larger the variance, the more dispersed the data points are. In Earth Science, 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 m data $x_1, x_2, x_3, \ldots, x_m$, its ***population variance*** is

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X) \\ &= \frac{1}{m}((x_1-\mu)^2 + (x_2-\mu)^2 + (x_3-\mu)^2 + \cdots + (x_m-\mu)^2) && (11.16a) \\ &= \frac{1}{m}\sum_{k=1}^{m}(x_k-\mu)^2 && (11.16b)\end{aligned}$$

where μ is the ***mean***, or ***expected value*** of X, and is computed by

$$\mu = E[X] = \frac{1}{m}(x_1 + x_2 + x_3 + \cdots + x_m) \tag{11.17}$$

that is, the average of all data. Hence, variance is the average of squares of differences between the data and their mean, equivalent to $E[(X-\mu)^2]$.

A simpler formula for computing the population variance is

$$\begin{aligned}
\sigma^2 &= E[(X-\mu)^2] \\
&= E[(X-E[X])^2] \\
&= E[X^2 - 2XE[X] + (E[X])^2] \\
&= E[X^2] - 2E[X]E[X] + (E[X])^2 && \text{(Note that } E[X] \text{ is a constant and } E[E[X]] = E[X], \; E[XE[X]] = E[X]E[X].\text{)} \\
&= E[X^2] - (E[X])^2 = E[X^2] - \mu^2 && (11.18)
\end{aligned}$$

As said before, we always have a finite sample size. To account for this, we have to use the sample variance s^2, which is the same as population variance but with the $\frac{1}{m}$ factor replaced by $\frac{1}{m-1}$.[10] Hence $s^2 = \frac{m}{m-1}\sigma^2$. 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 aforementioned short-cut Formula (11.18).

$$\begin{aligned}
\sigma^2 &= E[X^2] - \mu^2 \\
&= \frac{1}{5}(1^2 + 3^2 + 6^2 + 9^2 + 11^2) - 6^2 \\
&= 49.6 - 36
\end{aligned}$$

[10] It is because we have used one degree of freedom in estimating the mean.

$$= 13.6$$

► Short Exercise: Find the sample variance of X.[11]

Note that Formula (11.16) for variance in Definition 11.2.1 can be written as a dot product shown below.

Properties 11.2.2. Given a distribution X, with m data $\vec{x} = (x_1, x_2, x_3, \cdots, x_m)^T$, and a mean of μ, the population variance can be written as

$$\frac{1}{m}(\vec{x}' \cdot \vec{x}') = \frac{1}{m}\mathbf{x}'^T\mathbf{x}' \tag{11.19}$$

where $\mathbf{x}' = \vec{x}' = \vec{x} - \mu$ is the centered distribution with the mean μ removed.

It is simply a matter of observing that

$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{m}((x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_m-\mu)^2) \\ &= \frac{1}{m}(x_1-\mu, x_2-\mu, \ldots, x_n-\mu)^T \cdot (x_1-\mu, x_2-\mu, \ldots, x_m-\mu)^T\end{aligned}$$

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.[12]

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 (n) distributions, like $Z = c_1X^{(1)} + c_2X^{(2)} + \cdots + c_nX^{(n)}$, we may want to know how to compute its mean and variance. The mean will be simply

$$\mu_Z = E[c_1X^{(1)} + c_2X^{(2)} + \cdots + c_nX^{(n)}]$$

$$= c_1 E[X^{(1)}] + c_2 E[X^{(2)}] + \cdots + c_n E[X^{(n)}]$$
$$= c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n \tag{11.20}$$

where E is linear and $E[X^{(j)}] = \mu_j$ is the mean of $X^{(j)}$. The variance $\mathrm{Var}(Z)$ 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 m pairs of data, their ***population covariance*** is

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \frac{1}{m}((x_1-\mu_x)(y_1-\mu_y) + (x_2-\mu_x)(y_2-\mu_y)) \\ &\quad + \cdots + (x_m-\mu_x)(y_m-\mu_y)) \\ &= \frac{1}{m}\sum_{k=1}^{m}(x_k-\mu_x)(y_k-\mu_y)\end{aligned} \tag{11.21}$$

where μ_x and μ_y are the population means of X and Y respectively. It can be easily seen that $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ 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

$$\frac{1}{m}\vec{x}' \cdot \vec{y}' = \frac{1}{m}\mathbf{x}'^T\mathbf{y}' \tag{11.22}$$

For ***sample covariance***, it is

$$q_{xy} = \frac{1}{m-1}\sum_{k=1}^{m}(x_k-\bar{x})(y_k-\bar{y}) \tag{11.23}$$

where the $\frac{1}{m-1}$ factor replaces $\frac{1}{m}$ as for sample variance, $\bar{x}$ and $\bar{y}$ are the ***sample means*** of X and Y which happen to have the same values as μ_x and μ_y.

[11] 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 $\frac{5}{5-1}\sigma^2$.)

[12] When all data are equal (to the mean).

There is also a short-cut formula very similar to that for variance:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{11.24}$$

In general, if $\mathrm{Cov}(X,Y)$ is positive (negative), it means that when X increases, Y tends to increase (decrease) together. Finally, a direct comparison reveals that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ for any distribution X.

Example 11.2.1. Two time series of measured zonal and meridional wind speeds U and V at a weather station are shown in the table below.

(in 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

$$\begin{aligned}\mathrm{Cov}(U,V) = \frac{1}{7}[&(4.4-3.0)((-3.5)-(-1.9)) + (3.8-3.0)((-2.6)-(-1.9)) \\ &+ (3.3-3.0)((-2.7)-(-1.9)) + (2.8-3.0)((-1.4)-(-1.9)) \\ &+ (2.9-3.0)((-1.2)-(-1.9)) + (1.7-3.0)((-0.8)-(-1.9)) \\ &+ (2.1-3.0)((-1.1)-(-1.9))]\end{aligned}$$

$$= \frac{-5.36}{7} = -0.77\,\mathrm{m}^2\,\mathrm{s}^{-2}$$

Alternatively, the short-cut Formula (11.24) gives

$$\begin{aligned}
\mathrm{Cov}(U,V) &= E[UV] - \mu_U\mu_V \\
&= \frac{1}{7}[(4.4)(-3.5) + (3.8)(-2.6) + (3.3)(-2.7) + (2.8)(-1.4) \\
&\quad + (2.9)(-1.2) + (1.7)(-0.8) + (2.1)(-1.1)] - (3.0)(-1.9) \\
&= (-6.466) - (-5.7) = -0.77\,\mathrm{m}^2\,\mathrm{s}^{-2}
\end{aligned}$$

□

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 distributions 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

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{11.25a}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X,X)\mathrm{Cov}(Y,Y)}} \tag{11.25b}$$

where Var and 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, (11.22) in Definition 11.2.3, which gives

$$\begin{aligned}\rho_{xy} &= \frac{(\vec{x}' \cdot \vec{y}')}{\sqrt{(\vec{x}' \cdot \vec{x}')(\vec{y}' \cdot \vec{y}')}} \\ &= \frac{(\vec{x}' \cdot \vec{y}')}{\sqrt{\|\vec{x}'\|\|\vec{y}'\|}}\end{aligned}$$

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 $\left|\rho_{xy}\right| \leq 1$. □

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. The higher the magnitude of correlation, the 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, $\mathrm{Cov}(U, V) = -0.77\,\mathrm{m}^2\,\mathrm{s}^{-2}$, $\mathrm{Var}(U) = 0.75\,\mathrm{m}^2\,\mathrm{s}^{-2}$, $\mathrm{Var}(V) = 0.90\,\mathrm{m}^2\,\mathrm{s}^{-2}$, 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 ρ_{xy} take the value of 1 (or -1)?[13]

We are now prepared to derive the variance formula for linear combinations of multiple variables.

Properties 11.2.6. For a distribution constructed by a linear combination of multiple random variables, in the form of $Z = c_1X^{(1)} + c_2X^{(2)} + \cdots + c_nX^{(n)}$, where the coefficients $\vec{c} = (c_1, c_2, \cdots, c_n)^T$ are all constants, the variance $\mathrm{Var}(Z)$ can be expressed as a quadratic form $\mathbf{c}^TQ\mathbf{c}$, where

$$Q = \begin{bmatrix} \mathrm{Cov}(X^{(1)}, X^{(1)}) & \mathrm{Cov}(X^{(1)}, X^{(2)}) & \cdots & \mathrm{Cov}(X^{(1)}, X^{(n)}) \\ \mathrm{Cov}(X^{(2)}, X^{(1)}) & \mathrm{Cov}(X^{(2)}, X^{(2)}) & & \mathrm{Cov}(X^{(2)}, X^{(n)}) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X^{(n)}, X^{(1)}) & \mathrm{Cov}(X^{(n)}, X^{(2)}) & \cdots & \mathrm{Cov}(X^{(n)}, X^{(n)}) \end{bmatrix} \tag{11.26a}$$

$$= \begin{bmatrix} \mathrm{Var}(X^{(1)}) & \mathrm{Cov}(X^{(1)}, X^{(2)}) & \cdots & \mathrm{Cov}(X^{(1)}, X^{(n)}) \\ \mathrm{Cov}(X^{(2)}, X^{(1)}) & \mathrm{Var}(X^{(2)}) & & \mathrm{Cov}(X^{(2)}, X^{(n)}) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X^{(n)}, X^{(1)}) & \mathrm{Cov}(X^{(n)}, X^{(2)}) & \cdots & \mathrm{Var}(X^{(n)}) \end{bmatrix} \tag{11.26b}$$

is the so-called ***covariance matrix*** so that $Q_{ij} = \mathrm{Cov}(X^{(i)}, X^{(j)})$. If $[X'] = [X'^{(1)}|X'^{(2)}|\cdots|X'^{(n)}]$ is the matrix consisted of the centered variables $X'^{(j)} = X^{(j)} - E[X^{(j)}]$ in columns, then we have $Q = \frac{1}{m-1}[X']^T[X']$ where m is the number of data.

Proof. Let's say we have m data for Z: $z_1, z_2, \ldots, z_m$, as well as each of the $X^{(j)}$: $x_1^{(j)}, x_2^{(j)}, \ldots, x_m^{(j)}$. Denote the mean of $X^{(j)}$ by μ_j. Starting from the

Expression (11.16b) in Definition 11.2.1, we have

$$
\begin{aligned}
\mathrm{Var}(Z) &= \frac{1}{m-1}\sum_{k=1}^{m}(z_k-\mu_z)^2 \\
&= \frac{1}{m-1}\sum_{k=1}^{m}\left(\sum_{j=1}^{n}c_j x_k^{(j)} - \sum_{j=1}^{n}c_j\mu_j\right)^2 \\
&= \frac{1}{m-1}\sum_{k=1}^{m}\left(\sum_{j=1}^{n}(c_j x_k^{(j)} - c_j\mu_j)\right)^2 \\
&= \frac{1}{m-1}\sum_{k=1}^{m}\left[\left(\sum_{i=1}^{n}c_i(x_k^{(i)}-\mu_i)\right)\left(\sum_{j=1}^{n}c_j(x_k^{(j)}-\mu_j)\right)\right] \\
&\quad \text{(Changing to a new dummy summation variable)} \\
&= \frac{1}{m-1}\sum_{k=1}^{m}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_j(x_k^{(i)}-\mu_i)(x_k^{(j)}-\mu_j)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_j\left(\frac{1}{m-1}\sum_{k=1}^{m}(x_k^{(i)}-\mu_i)(x_k^{(j)}-\mu_j)\right) \\
&\quad \text{(Switching the order of summation)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_j\mathrm{Cov}(X^{(i)},X^{(j)}) \quad \text{((11.21) in Definition 11.2.3)} \\
&= \mathbf{c}^T Q\mathbf{c}
\end{aligned}
$$

□

[13] 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, $a > 0$ [$a < 0$].

For the two-variable situation, it reduces to

$$\mathbf{c}^T Q\mathbf{c} = \begin{bmatrix} c_1 & c_2 \end{bmatrix} \begin{bmatrix} \mathrm{Cov}(X^{(1)}, X^{(1)}) & \mathrm{Cov}(X^{(1)}, X^{(2)}) \\ \mathrm{Cov}(X^{(2)}, X^{(1)}) & \mathrm{Cov}(X^{(2)}, X^{(2)}) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \tag{11.27}$$

Example 11.2.2. From the previous wind speed observations example, if $W = 0.8U - 0.6V$, find $\mathrm{Var}(W)$.

Solution. Earlier calculations showed that $\mathrm{Var}(U) = 0.75\,\mathrm{m}^2\,\mathrm{s}^{-2}$, $\mathrm{Var}(V) = 0.90\,\mathrm{m}^2\,\mathrm{s}^{-2}$, $\mathrm{Cov}(U, V) = \mathrm{Cov}(V, U) = -0.77\,\mathrm{m}^2\,\mathrm{s}^{-2}$. Inserting the values into Formula (11.27) for Properties 11.2.6, we have

$$\begin{aligned} \mathrm{Var}(W) &= \begin{bmatrix} c_u & c_v \end{bmatrix} \begin{bmatrix} \mathrm{Var}(U) & \mathrm{Cov}(U, V) \\ \mathrm{Cov}(U, V) & \mathrm{Var}(V) \end{bmatrix} \begin{bmatrix} c_u \\ c_v \end{bmatrix} \\ &= \begin{bmatrix} 0.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.75 & -0.77 \\ -0.77 & 0.90 \end{bmatrix} \begin{bmatrix} 0.8 \\ -0.6 \end{bmatrix} \\ &= 1.54\,\mathrm{m}^2\,\mathrm{s}^{-2} \end{aligned}$$

The readers can cross-check by computing all the W data first and finding their variance directly. □

Finally, recall that variance is always a positive-semidefinite quantity, and since it can be calculated as a quadratic form via 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 is also known as ***Empirical Orthogonal Functions (EOFs)*** in Atmospheric Sciences, is the most common technique for this purpose, by finding the mode(s), 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 m pairs of data, from $\mathbf{x}_1 = (x_1, y_1)$ to $\mathbf{x}_m = (x_m, y_m)$. We can compute the covariance matrix

$$Q = \begin{bmatrix} \mathrm{Cov}(X, X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y, Y) \end{bmatrix}$$

which is 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.

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 Multivariable Calculus. We will proceed with the case of two variables for brevity.

Theorem 11.2.7 (Lagrange Multiplier)**.** To find the extreme values attained by a function $f(u, v, \ldots)$, under the constraint $g(u, v, \ldots) = 0$, we consider the expression

$$h(u, v, \ldots) = f(u, v, \ldots) - \lambda g(u, v, \ldots) \tag{11.28}$$

Figure 11.4.: *The two principal axes/directions (PCs) found by Principal Component Analysis for a random set of data (green). The longer, red [shorter, blue] arrow represents the direction of the largest [smallest] variance. The data points can be seen to spread more [less] along that direction.*

where λ is a constant so that the system below has a solution:

$$\begin{cases} \partial h/\partial u = 0 \\ \partial h/\partial v = 0 \\ \vdots \qquad\quad = 0 \end{cases} \tag{11.29}$$

$\partial/\partial u$ ($\partial/\partial v$) 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

$$g(x', y') = x'^2 + y'^2 - 1 = 0 \tag{11.30}$$

and with $f(x', y') = \mathbf{e}^T Q\mathbf{e}$

$$h(x', y') = (\mathbf{e}^T Q\mathbf{e}) - \lambda(x'^2 + y'^2 - 1) \tag{11.31a}$$

$$= \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} \mathrm{Cov}(X, X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y, Y) \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} - \lambda(x'^2 + y'^2 - 1)$$

$$= x'^2\mathrm{Cov}(X, X) + 2x'y'\mathrm{Cov}(X, Y) + y'^2\mathrm{Cov}(Y, Y) - \lambda(x'^2 + y'^2 - 1) \tag{11.31b}$$

according to (11.27). Carrying out the differentiation gives

$$\begin{cases} \partial h/\partial x' = 2x'\mathrm{Cov}(X, X) + 2y'\mathrm{Cov}(X, Y) - 2\lambda x' = 0 & (11.32a) \\ \partial h/\partial y' = 2x'\mathrm{Cov}(X, Y) + 2y'\mathrm{Cov}(Y, Y) - 2\lambda y' = 0 & (11.32b) \end{cases}$$

This system can be immediately simplified and recognized as

$$2\begin{bmatrix} \mathrm{Cov}(X, X) - \lambda & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y, Y) - \lambda \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = 0 \tag{11.33a}$$

$$(Q - \lambda I)\mathbf{e} = 0 \tag{11.33b}$$

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 extreme value when $\mathbf{e} = (x', y')^T$ 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

$$\begin{aligned}\mathbf{e}^{(j)T}(Q\mathbf{e}^{(j)}) &= \mathbf{e}^{(j)T}(\lambda_j\mathbf{e}^{(j)})\\ &= \lambda_j(\mathbf{e}^{(j)T}\mathbf{e}^{(j)})\\ &= \lambda_j\left\|\mathbf{e}^{(j)}\right\|^2\\ &= \lambda_j \qquad (11.34)\end{aligned}$$

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 the eigenvector is exactly equal to the corresponding eigenvalue. Note that orthogonal diagonalization (see the discussion below Definition 11.1.6) transforms the quadratic form to a diagonal matrix consisting of the eigenvalues λ_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 is zero, and hence the transformed variables are made uncorrelated.

Theorem 11.2.8 (Principal Component Analysis)**.** For a covariance matrix $Q = \frac{1}{m-1}[X']^T[X']$ (which happens to be symmetric), the variance $\mathbf{e}^TQ\mathbf{e}$ achieves its maximum value λ_1 along the direction $\mathbf{e}^{(1)}$, the largest eigenvalue of Q and the associated unit eigenvector. Generalizing, as Q has n orthonormal eigenvectors $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \cdots, \mathbf{e}^{(n)}$, arranged by the magnitude of corresponding eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ from largest to smallest, the largest variance will be λ_1 when the direction is along $\mathbf{e}^{(1)}$, the second largest will be λ_2 for $\mathbf{e}^{(2)}$ and so on, with the smallest variance being λ_n for $\mathbf{e}^{(n)}$. This set of orthonormal eigenvectors is 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 λ_j, we can compute the ratio of ***explained variance***, which is the fraction of λ_j over the total variance, the sum of eigenvalues/variances from all eigenvectors for the covariance matrix. This quantity allows us to assess how well the principal direction contributes to the total variance. In the coordinate system constructed by the orthonormal eigenvectors of Q, $[\mathbf{e}] = [\mathbf{e}^{(1)}|\mathbf{e}^{(2)}|\cdots|\mathbf{e}^{(n)}]$, the new coordinates for any (i-th) data point are

$$\vec{u}_i = [\mathbf{e}]^T \vec{x}_i \tag{11.35}$$

(see Section 10.1.2). By the Spectral Theorem 10.3.4, this $\vec{u}_i$ can be regarded to be the projection of $\vec{x}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})^T$, the i-th set 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 X and Y are as follows.

(in °C)	X	Y		X	Y
1st Day	21.0	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	23.0	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 X and Y are 22.5 °C and 22.6 °C respectively. After detrending by subtracting the respective means, the new data are

(in °C)	X'	Y'		X'	Y'
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

$$\begin{aligned} Q &= \frac{1}{14-1}\begin{bmatrix} X'\cdot X' & X'\cdot Y' \\ Y'\cdot X' & Y'\cdot Y' \end{bmatrix} \\ &= \frac{1}{13}\begin{bmatrix} -1.5 & -0.7 & -1.6 & \cdots \\ -0.3 & -1.0 & -1.4 & \cdots \end{bmatrix}\begin{bmatrix} -1.5 & -0.3 \\ -0.7 & -1.0 \\ -1.6 & -1.4 \\ \vdots & \vdots \end{bmatrix} \end{aligned}$$

$$= \frac{1}{13}\begin{bmatrix} 18.18 & 13.66 \\ 13.66 & 13.64 \end{bmatrix} = \begin{bmatrix} 1.398 & 1.051 \\ 1.051 & 1.049 \end{bmatrix}$$

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 $\lambda_1 = 2.289\,(^\circ\mathrm{C})^2$, and $\mathbf{e}^{(2)} = (-0.647, 0.763)^T$ of a smaller variance $\lambda_2 = 0.159\,(^\circ\mathrm{C})^2$. 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 $\vec{x}_i' = (x_i', y_i')^T$ 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 time series for the principal components 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 detail, the two principal components for the first day are 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 originating from the first principal direction only, we can simply remove the 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
X'	−1.021	−0.901	−1.622	−0.968	0.820	2.020	−0.094
Y'	−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
X'	−0.331	−1.026	0.125	−0.668	0.636	2.025	1.003
Y'	−0.281	−0.869	0.106	−0.566	0.539	1.716	0.850

For example, on the third day, PC1 contributes

$$\vec{x}_3' = \begin{bmatrix} x_3' \\ y_3' \end{bmatrix} = \mathbf{e}^{(1)} u_3^{(1)} = \begin{bmatrix} 0.763 \\ 0.647 \end{bmatrix} \begin{bmatrix} -2.126 \end{bmatrix} = \begin{bmatrix} -1.622 \\ -1.374 \end{bmatrix}$$

□

11.3. Python Programming

Since doing Empirical Orthogonal Functions (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],
```

Figure 11.5.: *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, for Example 11.2.3.*

Figure 11.6.: *Comparison between the original time series and the one reconstructed using only the first PC for Y' in Example 11.2.3.*

```
                [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 use, and for the present, 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 percentages 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 following.

```
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.

```
W = X - np.mean(X, axis=0)
lambda_1 = pca.explained_variance_[0]
lambda_2 = pca.explained_variance_[1]

import matplotlib.pyplot as plt

plt.scatter(W[:,0], W[:,1], color="g")
plt.xlim([-3,3])
plt.ylim([-3,3])
plt.xlabel("X' (deg C)")
plt.ylabel("Y' (deg C)")
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.arrow(0,0,lambda_1**0.5,0, color="r", width=0.05)
plt.arrow(0,0,0,lambda_2**0.5, color="b", width=0.05)
plt.grid()
plt.gca().set_aspect("equal")
```

```
plt.show()

plt.plot(np.arange(1,14+1), W[:,1], color="g", label="Original
    Y'")
plt.plot(np.arange(1,14+1), X_inv[:,1] - np.mean(X, axis=0)
   [1], color="r", label="Y' reconstructed from PC1")
plt.legend()
plt.grid()
plt.xticks(np.arange(1,14+1))
plt.xlabel("Day")
plt.ylabel("Temperature (deg C)")
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. 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
```

Load the SST data, work on coarse grids with a $2.5° \times 2.5°$ spatial resolution (optional to reduce the 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_12791_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")[0, ...])
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 over-water 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", dpi=300, bbox_inches="tight")
```

You should be able to get Figure 11.7 as below. This EOF pattern represents the ***El Niño-Southern Oscillation (ENSO)*** phenomenon where the western Pacific gets cooler [warmer] and the eastern Pacific becomes warmer [cooler] during the El Niño [La Niña] phase.

Figure 11.7.: *The first EOF pattern of the Pacific Sea Surface Temperature (SST) anomalies.*

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 an anti-clockwise rotation of 60 degrees for the coordinate frame 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 Prove that for any matrix A, $B = I + (A - A^T)$ is invertible. Hint: Consider $B\vec{x} = \mathbf{0}$ and show that $\vec{x} = \mathbf{0}$ must be the zero vector, inferred from the quadratic form $\mathbf{x}^T B\mathbf{x}$.

Exercise 11.5 Use 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 11.6 Find the covariance matrix for three sea level 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 11.7 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 the largest variance only.

Exercise 11.8 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 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 the 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 $\mathbb{R}$ [$\mathbb{C}$] scalar, 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* (in the first argument): 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 equipped with an inner product is then known as a real [complex] inner product space.

► Short Exercise: Show that $\langle a\vec{u}, b\vec{v} \rangle = a\overline{b}\langle \vec{u}, \vec{v} \rangle$ for any complex inner product space.[1]

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

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 any positive-definite symmetric bilinear form B, it can be verified that $\langle \vec{u}, \vec{v} \rangle = \mathbf{u}^T B \mathbf{v}$ will satisfy the real inner product axioms.[2] Below are some other properties of inner products 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. $\langle \vec{u} \pm \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle \pm \langle \vec{v}, \vec{w} \rangle$ (Distributive Property);
2. $\langle \vec{u}, \vec{v} \pm \vec{w} \rangle = \langle \vec{u}, \vec{v} \rangle \pm \langle \vec{u}, \vec{w} \rangle$ (Distributive Property);
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

[1] $\langle a\vec{u}, b\vec{v} \rangle = a\langle \vec{u}, b\vec{v} \rangle = a\overline{\langle b\vec{v}, \vec{u} \rangle} = ab\overline{\langle \vec{v}, \vec{u} \rangle} = a(\bar{b})\overline{\langle \vec{v}, \vec{u} \rangle} = a\bar{b}\langle \vec{u}, \vec{v} \rangle$ by the first and third axiom.

[2] For (1): $\langle \vec{u}, \vec{v} \rangle = \mathbf{u}^T B \mathbf{v} = (\mathbf{u}^T B \mathbf{v})^T$ since it is only a scalar, and then $(\mathbf{u}^T B \mathbf{v})^T = \mathbf{v}^T B^T \mathbf{u} = \mathbf{v}^T B \mathbf{u} = \langle \vec{v}, \vec{u} \rangle$ due to Properties 2.1.4 and because B is symmetric. (2) and (3) are obvious. For (4), as B is required to be positive-definite, $\langle \vec{v}, \vec{v} \rangle = \mathbf{v}^T B \mathbf{v} > 0$ by Definition 11.1.4 as long as $\vec{v} \neq \mathbf{0}$ and it is apparent that $\langle \vec{v}, \vec{v} \rangle = \mathbf{0}^T B \mathbf{0} = 0$ when $\vec{v} = \mathbf{0}$.

exercise with $a = 1$. For (5):

$$\begin{aligned} \langle \vec{u}, \vec{v} \rangle &= \langle \vec{u}, \vec{w} \rangle \\ \langle \vec{u}, \vec{v} - \vec{w} \rangle &= 0 && \text{(By (1))} \end{aligned}$$

Then by letting $\vec{u} = \vec{v} - \vec{w}$ and using the last axiom we get $\vec{v} - \vec{w} = \mathbf{0}$ and thus $\vec{v} = \vec{w}$. □

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 with 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 (Vector Norm)**.** The norm of a vector in an inner product space is induced by

$$\|\vec{v}\| = \sqrt{\langle \vec{v}, \vec{v} \rangle} \qquad \text{or equivalently} \qquad \|\vec{v}\|^2 = \langle \vec{v}, \vec{v} \rangle \tag{12.1}$$

where $\langle \cdot, \cdot \rangle$ is an inner product following Definition 12.1.1. Unit vectors with respect to a specific inner product are then conveniently created using this definition of norm.

Definition 12.1.4. The unit vector of a non-zero vector $\vec{v} \in \mathcal{V}$ in an inner

product space is denoted as $\hat{v}$ and is given by

$$\hat{v} = \frac{1}{\|\vec{v}\|}\vec{v} \tag{12.2}$$

where the norm $\|\vec{v}\|$ is now defined as in Properties 12.1.3.

The notion of orthogonality (as well as orthonormality) in Properties 4.2.5 is also transferred to an inner product space by the same essence of replacing the dot product with 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

$$|\langle \vec{u}, \vec{v} \rangle| \leq \|\vec{u}\|\|\vec{v}\| = \sqrt{\langle \vec{u}, \vec{u} \rangle}\sqrt{\langle \vec{v}, \vec{v} \rangle} \tag{12.3}$$

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 ones. 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

$$\langle \vec{u}, \vec{v} \rangle = \mathbf{u}^T B \mathbf{v}$$

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 11.1.5, it simply amounts to checking 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

$$\begin{aligned} \langle \vec{u}, \vec{v} \rangle &= \mathbf{u}^T B \mathbf{v} \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = 0 \end{aligned}$$

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. □

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 becomes 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[3], it must be infinite. Recall that from Chapter 7, a general vector space can be generated in terms of functions, and the same can be said for an inner product space. 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

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)}dx \tag{12.4}$$

where $a < b$ are some real constants. Checking the validity of this inner product formulation is not hard.[4] Unfortunately, the above example fails to be a ***Hilbert space*** which we often desire. 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[5], 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[6] equipped with the same inner product in Equation (12.4) is such a Hilbert space. The reason why we care so much about Hilbert spaces, particularly those that are separable, is because they always admit a countable orthonormal basis that is *complete*.

[3]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*.

[4]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 $\langle f, f \rangle \geq \frac{|f(c)|^2}{2}(2\delta) = |f(c)|^2\delta > 0$.

[5]To put it 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.

[6]A square-integrable function f means that $\int |f|^2 dx < \infty$ is finite, and the $L^2[a, b]$ space is actually the "completed" (in the sense of the footnote above) version of $C^0[a, b]$.

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 φ_j denotes the j-th basis vector/function 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 $\tilde{\varphi}$ that 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 infinite-dimensional Hilbert space can be expanded into an infinite sum of orthonormal vectors $f = c_1\varphi_1 + c_2\varphi_2 + \cdots = \lim_{n\to\infty} \sum_{j=1}^{n} c_j\varphi_j$.[7] 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 *Schuader basis*. As noted before, the formal treatment of Hilbert spaces (or infinite-dimensional spaces in general) is out of our reach, and we will invoke the relevant properties as we see fit to avoid going down the rabbit hole.

Back to infinite-dimensional inner product spaces in general, the properties and theorems given earlier in this section still hold for them as we have defined inner products in a way without regard to the (in)finiteness of dimensions.

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 (12.4).

Solution. This is to check

$$|\langle \varphi_1, \varphi_2 \rangle| \leq \|\varphi_1\| \|\varphi_2\|$$

[7](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 $\tilde{\varphi} = c_1\varphi_1 + c_2\varphi_2 + \cdots$. 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.

$$\left|\langle x, x^2\rangle\right| \leq \|x\|\|x^2\|$$

by computing the three quantities in the above inequality:

$$\begin{aligned}
\left|\langle x, x^2\rangle\right| &= \left|\int_0^1 x\overline{(x^2)}dx\right| \\
&= \left|\int_0^1 x^3 dx\right| \\
&= \left|\left[\frac{1}{4}x^4\right]_0^1\right| = \left|\frac{1}{4}\right| = \frac{1}{4} \\
\|x\| &= \sqrt{\int_0^1 x\overline{(x)}dx} \\
&= \sqrt{\int_0^1 x^2 dx} \\
&= \sqrt{\left[\frac{1}{3}x^3\right]_0^1} = \frac{1}{\sqrt{3}}
\end{aligned}$$

similarly $\left\|x^2\right\| = \frac{1}{\sqrt{5}}$. Hence $\left|\langle x, x^2\rangle\right| = \frac{1}{4} = \frac{1}{\sqrt{16}} < \frac{1}{\sqrt{15}} = (\frac{1}{\sqrt{3}})(\frac{1}{\sqrt{5}}) = \|x\|\|x^2\|$ and the inequality holds in this case. □

► Short Exercise: Show that the same[8] 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 (12.4) but integrated from -1 to 1 instead.[9]

[8]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.

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 scenes (see Section 7.1), we can extend this idea for a general vector space $\mathcal{V}$ that has an inner product, where $\vec{u}, \vec{v} \in \mathcal{V}$ and the linear operator is $T : \mathcal{V} \to \mathcal{V}$ now. Subsequently, the adjoint of T will then be another linear operator $T^\dagger$ defined by the relationship $\langle T(\vec{u}), \vec{v}\rangle = \langle \vec{u}, T^\dagger(\vec{v})\rangle$. Again, if the standard inner product is employed, then we know that if the matrix representation of T in the basis β is $[T]_\beta$, then

$$\begin{aligned}
\langle T(\vec{u}), \vec{v}\rangle &\equiv \langle [T]_\beta [\vec{u}]_\beta, [\vec{v}]_\beta\rangle \\
&= ([T]_\beta [\vec{u}]_\beta)^T \overline{[\vec{v}]_\beta} \\
&= [\vec{u}]_\beta^T [T]_\beta^T \overline{[\vec{v}]_\beta} \\
&= [\vec{u}]_\beta^T (\overline{[T]_\beta^* [\vec{v}]_\beta}) \\
&= \langle [\vec{u}]_\beta, ([T]_\beta)^* [\vec{v}]_\beta\rangle \overset{\text{def}}{\equiv} \langle \vec{u}, T^\dagger(\vec{v})\rangle
\end{aligned} \tag{12.5}$$

by Properties 8.2.7, so that we identify the adjoint operator $T^\dagger$ with the conjugate transpose matrix representation of $([T]_\beta)^*$ as expected.

Definition 12.2.1 (Adjoint)**.** The adjoint of a linear operator $T : \mathcal{V} \to \mathcal{V}$ with respect to some inner product $\langle \cdot, \cdot\rangle$ is another linear operator denoted by $T^\dagger$ such

[9] $\int_{-1}^{1} x\overline{(x^2)} = \int_{-1}^{1} x^3 = 0$ as x^3 is an odd function and the integration limits are symmetric.

that the relation

$$\langle T(\vec{u}), \vec{v}\rangle = \langle \vec{u}, T^\dagger(\vec{v})\rangle \tag{12.6}$$

holds for all $\vec{u}, \vec{v} \in \mathcal{V}$. Such an adjoint $T^\dagger$ is always unique for a given T.

It is easy to check the linearity and uniqueness of an adjoint.[10] Now let's look at how adjoint is found when inner products other than the standard one are used. First, if $\mathcal{V}$ is a finite-dimensional complex vector space and a new inner product is defined via a positive-definite Hermitian form (see Section 11.1.4[11]) $\langle \vec{u}, \vec{v}\rangle = \mathbf{u}^T\overline{H\mathbf{v}} = (H^*\vec{u}) \cdot \vec{v}$ where $\cdot$ denotes the usual complex dot product, then for a (complex) linear operator T with a matrix representation of $[T]_\beta$ in the β system, by Definition 8.2.1 and Properties 8.2.7:

$$\begin{aligned}
\langle T(\vec{u}), \vec{v}\rangle &\equiv (H^*[T]_\beta[\vec{u}]_\beta) \cdot ([\vec{v}]_\beta) \\
&= (H^*[T]_\beta(H^*)^{-1}H^*[\vec{u}]_\beta) \cdot ([\vec{v}]_\beta) \\
&= (H^*[\vec{u}]_\beta) \cdot ((H^*[T]_\beta(H^*)^{-1})^*[\vec{v}]_\beta) \\
&\equiv \langle \vec{u}, (H^*[T]_\beta(H^*)^{-1})^*\vec{v}\rangle = \langle \vec{u}, H^{-1}[T]_\beta^* H\vec{v}\rangle
\end{aligned}$$

and thus we identify the matrix representation of the adjoint as

$$T^\dagger \equiv H^{-1}[T]_\beta^* H \tag{12.7}$$

relative 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

[10] Linearity (Definition 7.1.1): Consider $\langle \vec{u}, T^\dagger(a\vec{v} + b\vec{w})\rangle = \langle T(\vec{u}), a\vec{v} + b\vec{w}\rangle = \overline{a}\langle T(\vec{u}), \vec{v}\rangle + \overline{b}\langle T(\vec{u}), \vec{w}\rangle$ by Definitions 12.2.1 and Properties 12.1.2. Subsequently $\overline{a}\langle T(\vec{u}), \vec{v}\rangle + \overline{b}\langle T(\vec{u}), \vec{w}\rangle = \overline{a}\langle \vec{u}, T^\dagger(\vec{v})\rangle + \overline{b}\langle \vec{u}, T^\dagger(\vec{w})\rangle = \langle \vec{u}, aT^\dagger(\vec{v})\rangle + \langle \vec{u}, bT^\dagger(\vec{w})\rangle = \langle \vec{u}, aT^\dagger(\vec{v}) + bT^\dagger(\vec{w})\rangle$ by those definitions again. Since this has to hold for any $\vec{u}$ (plus $\vec{v}$ and $\vec{w}$), $\langle \vec{u}, T^\dagger(a\vec{v} + b\vec{w})\rangle = \langle \vec{u}, aT^\dagger(\vec{v}) + bT^\dagger(\vec{w})\rangle$ means that $T^\dagger(a\vec{v} + b\vec{w}) = aT^\dagger(\vec{v}) + bT^\dagger(\vec{w})$ by Properties 12.1.2. Uniqueness: Assume that there is another adjoint $S^\dagger$ satisfies $\langle T(\vec{u}), \vec{v}\rangle = \langle \vec{u}, S^\dagger(\vec{v})\rangle$ for all $\vec{u}$ and $\vec{v}$, then by Properties 12.1.2 $T^\dagger(\vec{v}) = S^\dagger(\vec{v})$ for all $\vec{v}$. Hence $T^\dagger$ and $S^\dagger$ must be the same operator by Properties 7.1.3.

[11] The readers should check for themselves that the four inner space axioms hold in this case.

Equation (12.4), 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:

$$\langle f, g\rangle = \int_a^b w(x) f(x)\overline{g(x)}dx \tag{12.8}$$

where $w(x) > 0$ is a positive-definite *real* weighting function, the unique adjoint $\mathcal{L}^\dagger$ of a linear operator $\mathcal{L}$ is another one that satisfies

$$\begin{aligned}\langle \mathcal{L}[f], g\rangle &= \int_a^b w(x)\mathcal{L}[f(x)]\overline{g(x)}dx \\ &= \int_a^b w(x) f(x)\overline{(\mathcal{L}^\dagger[g(x)])}dx + \text{boundary terms} \\ &= \langle f, \mathcal{L}^\dagger[g]\rangle + \text{boundary terms} \end{aligned} \tag{12.9}$$

for all functions $f, g \in \mathcal{V}$ and the boundary terms are evaluated at the end-points a and b.

Equation (12.4) is then simply a special case of (12.8) 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 (12.4).

Solution. We start with $\langle \mathcal{L}[f], g\rangle = \int_a^b \left(x\frac{d}{dx}(f(x))\right)\overline{g(x)}dx$ and aim to rewrite it into the form of $\int_a^b f(x)\overline{(\mathcal{L}^\dagger[g(x)])}dx = \langle f, \mathcal{L}^\dagger[g]\rangle$, plus possibly some

boundary term(s). As suggested above we can try to apply integration by parts:

$$\begin{aligned}
\langle \mathcal{L}[f], g\rangle &= \int_a^b \left(x\frac{d}{dx}(f(x))\right)\overline{g(x)}dx \\
&= \int_a^b \frac{d}{dx}(f(x))x\overline{(g(x))}dx \\
&= [xf(x)\overline{g(x)}]_a^b - \int_a^b f(x)\frac{d}{dx}(x\overline{(g(x))})dx \\
&= \int_a^b f(x)\overline{\left(-\frac{d}{dx}(x(g(x)))\right)}dx + [xf(x)\overline{(g(x))}]_a^b
\end{aligned}$$

The boundary term is now put after the integral, from which we deduce that $\mathcal{L}^\dagger[g] = -\frac{d}{dx}(x[g])$ by comparing it with $\int_a^b f(x)\overline{(\mathcal{L}^\dagger[g(x)])}dx$. □

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 $\mathcal{V}$ with adjoints $T^\dagger$ and $U^\dagger$ respectively, we have

1. $(cT)^\dagger = \overline{c}T^\dagger$, where c is any complex scalar;
2. $(T^\dagger)^\dagger = T$ (the adjoint of adjoint is itself);
3. $(T \pm U)^\dagger = T^\dagger \pm U^\dagger$;
4. $(T^\dagger)^{-1} = (T^{-1})^\dagger$ if T (hence $T^\dagger$) is invertible;
5. $(TU)^\dagger = U^\dagger T^\dagger$.

We will briefly show the last item here. Using Definition 12.2.1 twice, we have

$$\begin{aligned}
\langle TU(\vec{u}), \vec{v}\rangle &= \langle U(\vec{u}), T^\dagger(\vec{v})\rangle \\
&= \langle \vec{u}, U^\dagger T^\dagger(\vec{v})\rangle
\end{aligned}$$

so we identify that $(TU)^\dagger = U^\dagger T^\dagger$.

12.2.2. Hermitian Operators

As suggested in 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 $T^\dagger$ which is equal to itself, i.e. $T^\dagger = T$, 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 H that induces its inner product, then according to what we have just derived, we simply need to check if $T^\dagger \equiv H^{-1}[T]_\beta^* H = [T]_\beta \equiv T$. The problem becomes a bit more complicated when we consider the integral inner product in Equation (12.8) 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 operator $\mathcal{L}$ is further known as a ***Hermitian*** (or *truly self-adjoint*, compared to *formally self-adjoint*) operator.

Definition 12.2.4 (Self-adjointness and Hermiticity for Linear Operators)**.** A linear operator T is self-adjoint if its adjoint $T^\dagger = T$ is equal 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 $\mathcal{L}^\dagger$ 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 (12.4) over the entire x-axis. Find the form of its eigenfunctions (treating functions as (eigen)vectors).

Solution.

$$\langle \mathcal{L}[f], g\rangle = \int_{-\infty}^{\infty} \mathcal{L}[f(x)]\overline{g(x)}dx$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \left(-i\hbar \frac{d}{dx}(f(x))\right) \overline{g(x)} dx \\
&= [-i\hbar f(x)\overline{g(x)}]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(x) \left(-i\hbar \frac{d}{dx}(\overline{g(x)})\right) dx \\
&\quad \text{(Integration by parts)} \\
&= \int_{-\infty}^{\infty} f(x) \left(i\hbar \frac{d}{dx}(\overline{g(x)})\right) dx + [-i\hbar f(x)\overline{g(x)}]_{-\infty}^{\infty} \\
&= \int_{-\infty}^{\infty} f(x) \overline{\left(-i\hbar \frac{d}{dx}(g(x))\right)} dx + [-i\hbar f(x)\overline{g(x)}]_{-\infty}^{\infty} \\
&= \int_{-\infty}^{\infty} f(x) \overline{\mathcal{L}^{\dagger}[g(x)]} dx + [-i\hbar f(x)\overline{g(x)}]_{-\infty}^{\infty}
\end{aligned}
$$

So we see that $\mathcal{L}^{\dagger} = \mathcal{L} = -i\hbar \frac{d}{dx}$ is self-adjoint. The eigenfunctions φ_m of $\mathcal{L}$ can be found from solving the ODE:

$$
\begin{aligned}
\mathcal{L}[\varphi_m] = -i\hbar \frac{d\varphi_m}{dx} &= m\varphi_m \\
-\int i\hbar \frac{d\varphi_m}{\varphi_m} &= \int m dx \\
-i\hbar \ln \varphi_m &= mx + C \\
\therefore \varphi_m &= K e^{\frac{i}{\hbar} mx}
\end{aligned}
$$

where $K = e^{\frac{i}{\hbar} C}$ is any scaling constant. Notice that the eigenvalue m can take any value ranging from $-\infty$ to ∞ as there is no boundary condition being enforced. In this case, the eigenvalues form a *continuous spectrum* and we note that the inner product 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. □

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 $f = c_1\varphi_1 + c_2\varphi_2 + \cdots = \sum_{j=1}^{\infty} c_j \varphi_j$ 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 have introduced separable Hilbert spaces and Hermitian operators before and now, as they carry the desirable attributes of possessing countably infinite eigenfunctions that form a complete orthonormal basis and thus a *discrete spectrum*. 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 important 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 $\mathcal{L}^\dagger = \mathcal{L}$ must be real.

Proof. Let φ be an eigenfunction of $\mathcal{L}$ with an eigenvalue of λ. Then consider

$$\begin{aligned}
\langle \mathcal{L}[\varphi], \varphi \rangle &= \langle \lambda\varphi, \varphi \rangle && \text{(Definition 9.1.1)} \\
&= \lambda \langle \varphi, \varphi \rangle && \text{(Definition 12.1.1)} \\
&= \lambda \|\varphi\|^2 && \text{(Properties 12.1.3)}
\end{aligned}$$

but also

$$\begin{aligned}
\langle \mathcal{L}[\varphi], \varphi \rangle &= \langle \varphi, \mathcal{L}^\dagger[\varphi] \rangle + \cdots && \text{(Properties 12.2.2)} \\
&= \langle \varphi, \mathcal{L}[\varphi] \rangle && (\mathcal{L} \text{ is Hermitian, Definition 12.2.4}) \\
&= \langle \varphi, \lambda\varphi \rangle && \text{(Definition 9.1.1)} \\
&= \overline{\lambda} \langle \varphi, \varphi \rangle = \overline{\lambda} \|\varphi\|^2 && \text{(Properties 12.1.2 and 12.1.3)}
\end{aligned}$$

This implies that $\lambda\|\varphi\|^2 = \overline{\lambda}\|\varphi\|^2$. Since the eigenfunction $\varphi \neq 0$ cannot be zero, $\|\varphi\|^2 \neq 0$ and thus it must be $\lambda = \overline{\lambda}$ so the eigenvalue will be real. □

Moreover, we can briefly prove the orthogonality between their eigenfunctions.

Properties 12.2.6. Any two eigenfunctions of a Hermitian operator $\mathcal{L}^\dagger = \mathcal{L}$ 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 φ_1 and φ_2 where the respective eigenvalues are m_1 and m_2. Subsequently, we have

$$\begin{aligned}\langle \mathcal{L}[\varphi_1], \varphi_2\rangle &= \langle m_1\varphi_1, \varphi_2\rangle && \text{(Definition 9.1.1)}\\ &= m_1\langle \varphi_1, \varphi_2\rangle\end{aligned}$$

but also

$$\begin{aligned}\langle \mathcal{L}[\varphi_1], \varphi_2\rangle &= \langle \varphi_1, \mathcal{L}^\dagger[\varphi_2]\rangle + \cdots && \text{(Properties 12.2.2)}\\ &= \langle \varphi_1, \mathcal{L}[\varphi_2]\rangle && \text{(Definition 12.2.4)}\\ &= \langle \varphi_1, m_2\varphi_2\rangle && \text{(Definition 9.1.1)}\\ &= \overline{m_2}\langle \varphi_1, \varphi_2\rangle && \text{(Properties 12.1.2)}\\ &= m_2\langle \varphi_1, \varphi_2\rangle && \text{(Properties 12.2.5)}\end{aligned}$$

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 φ_1, φ_2 are orthogonal with respect to the inner product (Properties 12.1.5). □

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 (12.4) 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

$$\begin{aligned}
\langle \mathcal{L}[f], g\rangle &= \int_{-\pi}^{\pi} \mathcal{L}[f(x)]\overline{g(x)}dx \\
&= \int_{-\pi}^{\pi} \frac{d^2}{dx^2}(f(x))\overline{g(x)}dx \\
&= [\frac{d}{dx}(f(x))\overline{(g(x))}]_{-\pi}^{\pi} - \int_{-\pi}^{\pi} \frac{d}{dx}(f(x))\frac{d}{dx}\overline{(g(x))}dx \\
&= [\frac{d}{dx}(f(x))\overline{(g(x))}]_{-\pi}^{\pi} - [(f(x))\frac{d}{dx}\overline{(g(x))}]_{-\pi}^{\pi} \\
&\quad + \int_{-\pi}^{\pi} f(x)\frac{d^2}{dx^2}\overline{(g(x))}dx \\
&= [\frac{d}{dx}(f(x))\overline{(g(x))}]_{-\pi}^{\pi} - [(f(x))\frac{d}{dx}\overline{(g(x))}]_{-\pi}^{\pi} \\
&\quad + \int_{-\pi}^{\pi} f(x)\overline{\left(\frac{d^2}{dx^2}(g(x))\right)}dx \\
&= [\frac{d}{dx}(f(x))\overline{(g(x))}]_{-\pi}^{\pi} - [(f(x))\frac{d}{dx}\overline{(g(x))}]_{-\pi}^{\pi} \\
&\quad + \int_{-\pi}^{\pi} f(x)\overline{\mathcal{L}[g(x)]}dx
\end{aligned}$$

we thus see $\mathcal{L}^\dagger = \frac{d^2}{dx^2} = \mathcal{L}$. Again from basic ODE, we know that the eigenfunctions φ 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, $m = 0, 1, 2, \ldots$ must be a non-negative integer[12], with the eigenvalues being $-m^2$. Hence the countably infinite, orthogonal[13](Properties 12.1.8, not yet normalized) basis of eigenfunctions for $\mathcal{L}$ are $\{\sin(x), \sin(2x), \sin(3x), \ldots, 1, \cos(x), \cos(2x), \cos(3x), \ldots\}$. This basis, consisting of sines and cosines with discrete, equally spaced frequencies, is famously known as the ***Fourier basis***. □

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 the conjugate transpose is now replaced by its adjoint $T^\dagger$ and the notion of distance is relative to the given inner product.

Definition 12.2.7 (Unitary Operator)**.** A unitary operator T is a linear operator such that its inverse operator is equal to its adjoint $T^{-1} = T^\dagger$ with respect to the inner product used.

[12]Without loss of generality, let $f = \sin(m_1x)$ and $g = \cos(m_2x)$, then

$$\begin{aligned}
&[\frac{d}{dx}(f(x))\overline{(g(x))}]_{-\pi}^{\pi} - [(f(x))\frac{d}{dx}\overline{([g(x)])}]_{-\pi}^{\pi} \\
&= [m_1\cos(m_1x)\cos(m_2x)]_{-\pi}^{\pi} - [-m_2\sin(m_1x)\sin(m_2x)]_{-\pi}^{\pi} \\
&= [\frac{m_1}{2}(\cos((m_1-m_2)x) + \cos((m_1+m_2)x))]_{-\pi}^{\pi} \\
&\quad - [-\frac{m_2}{2}(\cos((m_1-m_2)x) - \cos((m_1+m_2)x))]_{-\pi}^{\pi}
\end{aligned}$$

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 $\sin(-m_1x) = -\sin(m_1x)$ and $\cos(-m_2x) = \cos(m_2x)$, we only need to take the integers that are positive.

With this definition, the distance-preserving property can be readily derived.

Properties 12.2.8. Transformation by a unitary operator $T = T^\dagger$ on a vector (or function) 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,

$$\begin{aligned}
\|T(\vec{v})\|^2 = \langle T(\vec{v}), T(\vec{v})\rangle &= \langle \vec{v}, T^\dagger(T(\vec{v}))\rangle && \text{(Definition 12.2.1)}\\
&= \langle \vec{v}, T^{-1}(T(\vec{v}))\rangle && \text{(Definition 12.2.7)}\\
&= \langle \vec{v}, \mathrm{id}(\vec{v})\rangle && \text{(Definition 7.1.12)}\\
&= \langle \vec{v}, \vec{v}\rangle = \|\vec{v}\|^2
\end{aligned}$$

is shown to be equal throughout the unitary transformation. □

Example 12.2.4. Show that for the $L^2[a, b]$ space with the inner product of Equation (12.4), 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 $(\mathcal{L}^{-1} \circ \mathcal{L})[f] = e^{-ikx}(e^{ikx}[f]) = (1)f = \mathrm{id}[f]$, and we have to show that it

[13] 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

$$\begin{aligned}
\int_{-\pi}^{\pi} \sin(m_1 x)\cos(m_2 x)dx &= \int_{-\pi}^{\pi} \frac{1}{2}(\sin((m_1 - m_2)x) + \sin((m_1 + m_2)x))dx\\
&= [\frac{1}{2}(-\frac{\cos((m_1 - m_2)x)}{m_1 - m_2} - \frac{\cos((m_1 + m_2)x)}{m_1 + m_2})]_{-\pi}^{\pi}
\end{aligned}$$

which yields 0 when the end-points $-\pi, \pi$ are substituted into the expression.

equals to the adjoint $\mathcal{L}^\dagger$:

$$\begin{aligned}
\langle \mathcal{L}[f], g\rangle &= \int_a^b \mathcal{L}[f(x)]\overline{g(x)}dx \\
&= \int_a^b (e^{ikx} f(x))\overline{g(x)}dx \\
&= \int_a^b f(x)\overline{e^{-ikx} g(x)}dx \qquad\qquad (\overline{e^{ikx}} = e^{-ikx}) \\
&= \int_a^b f(x)\overline{\mathcal{L}^{-1}[g(x)]}dx
\end{aligned}$$

Comparing with $\langle f, \mathcal{L}^\dagger[g]\rangle$, from this we readily infer that $\mathcal{L}^\dagger = \mathcal{L}^{-1}$ and thus $\mathcal{L}$ is unitary. □

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 the Spectral Theorem. Remember in Section 5.2.1, we derived 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

$$\overrightarrow{\text{proj}}_u v = \frac{\langle \vec{v}, \vec{u}\rangle}{\|\vec{u}\|^2}\vec{u} \tag{12.10}$$

This definition is consistent in the sense that $\vec{u}$ and the component of $\vec{v}$ normal to $\vec{u}$ are orthogonal:

$$\begin{aligned}
\langle \vec{u}, \vec{v} - \overrightarrow{\text{proj}}_u v \rangle &= \langle \vec{u}, \vec{v} - \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{u}\|^2} \vec{u} \rangle \\
&= \langle \vec{u}, \vec{v} \rangle - \langle \vec{u}, \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{u}\|^2} \vec{u} \rangle \\
&= \langle \vec{u}, \vec{v} \rangle - \frac{\overline{\langle \vec{v}, \vec{u} \rangle}}{\|\vec{u}\|^2} \langle \vec{u}, \vec{u} \rangle \qquad \text{(Properties 12.1.2)} \\
&= \langle \vec{u}, \vec{v} \rangle - \frac{\langle \vec{u}, \vec{v} \rangle}{\|\vec{u}^2\|} \|\vec{u}\|^2 \\
&= \langle \vec{u}, \vec{v} \rangle - \langle \vec{u}, \vec{v} \rangle = 0
\end{aligned}$$

Similar to Expression (10.8b), the orthogonal projection operator T onto a subspace $\mathcal{W} \subseteq \mathcal{V}$ of an inner product space with an orthonormal basis $\{\vec{w}^{(1)}, \vec{w}^{(2)}, \ldots\}$ 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:

$$T(\vec{v}) = \frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\left\|\vec{w}^{(1)}\right\|^2} \vec{w}^{(1)} + \frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\left\|\vec{w}^{(2)}\right\|^2} \vec{w}^{(2)} + \cdots \tag{12.11}$$

Just as Properties 10.3.1 and 10.3.2, a linear operator T represents an orthogonal projection if and only if it has an adjoint $T^\dagger$ so that $T^2 = T = T^\dagger$.

Properties 12.3.2. A linear operator $T : \mathcal{V} \to \mathcal{V}$ over an inner product space, is an orthogonal projection with respect to the inner product if and only if $T^2 = T = T^\dagger$ where $T^\dagger$ is its adjoint.

The proof for the first equality $T^2 = T$ is the same one in Properties 10.3.1 except that the † adjoint superscript now replaces that of conjugate transpose *, and $\mathcal{V}$, as well as $\mathcal{W}_1$ and $\mathcal{W}_2$ may be infinite-dimensional now. Meanwhile

for the second equality $T = T^\dagger$, 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.[14] We now show that the expression of the orthogonal projector T given in Equation (12.11) satisfies the above requirement $T^2 = T = T^\dagger$: Using (12.11) twice, we have

$$\begin{aligned}
T^2(\vec{v}) = T(T(\vec{v})) &= \frac{\langle T(\vec{v}), \vec{w}^{(1)}\rangle}{\left\|\vec{w}^{(1)}\right\|^2}\vec{w}^{(1)} + \frac{\langle T(\vec{v}), \vec{w}^{(2)}\rangle}{\left\|\vec{w}^{(2)}\right\|^2}\vec{w}^{(2)} + \cdots \\
&= \frac{\left\langle \frac{\langle \vec{v},\vec{w}^{(1)}\rangle}{\|\vec{w}^{(1)}\|^2}\vec{w}^{(1)} + \frac{\langle \vec{v},\vec{w}^{(2)}\rangle}{\|\vec{w}^{(2)}\|^2}\vec{w}^{(2)} + \cdots, \vec{w}^{(1)}\right\rangle}{\left\|\vec{w}^{(1)}\right\|^2}\vec{w}^{(1)} \\
&\quad + \frac{\left\langle \frac{\langle \vec{v},\vec{w}^{(1)}\rangle}{\|\vec{w}^{(1)}\|^2}\vec{w}^{(1)} + \frac{\langle \vec{v},\vec{w}^{(2)}\rangle}{\|\vec{w}^{(2)}\|^2}\vec{w}^{(2)} + \cdots, \vec{w}^{(2)}\right\rangle}{\left\|\vec{w}^{(2)}\right\|^2}\vec{w}^{(2)} + \cdots \\
&= \frac{\frac{\langle \vec{v},\vec{w}^{(1)}\rangle}{\|\vec{w}^{(1)}\|^2}\left\|\vec{w}^{(1)}\right\|^2 + (0)}{\left\|\vec{w}^{(1)}\right\|^2}\vec{w}^{(1)} \\
&\quad + \frac{(0) + \frac{\langle \vec{v},\vec{w}^{(2)}\rangle}{\|\vec{w}^{(2)}\|^2}\left\|\vec{w}^{(2)}\right\|^2 + (0)}{\left\|\vec{w}^{(2)}\right\|^2}\vec{w}^{(2)} + \cdots \\
&\quad (\vec{w}^{(j)} \text{ are orthogonal})
\end{aligned}$$

[14] 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 $\langle T(\vec{w}), \vec{w} - T(\vec{w})\rangle = \langle \vec{w}, T^\dagger(\vec{w} - T(\vec{w}))\rangle = \langle \vec{w}, T(\vec{w} - T(\vec{w}))\rangle = \langle \vec{w}, T(\vec{w}) - T^2(\vec{w})\rangle = \langle \vec{w}, \mathbf{0}\rangle = 0$ since $T^\dagger = T$ 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 $\vec{w} \in \mathcal{R}(T)$.

$$= \frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\left\| \vec{w}^{(1)} \right\|^2} \vec{w}^{(1)} + \frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\left\| \vec{w}^{(2)} \right\|^2} \vec{w}^{(2)} + \cdots = T(\vec{v})$$

and

$$\begin{aligned}
\langle \vec{u}, T(\vec{v}) \rangle &= \langle \vec{u}, \frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\left\| \vec{w}^{(1)} \right\|^2} \vec{w}^{(1)} + \frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\left\| \vec{w}^{(2)} \right\|^2} \vec{w}^{(2)} + \cdots \rangle \\
&= \overline{\frac{\langle \vec{v}, \vec{w}^{(1)} \rangle}{\left\| \vec{w}^{(1)} \right\|^2}} \langle \vec{u}, \vec{w}^{(1)} \rangle + \overline{\frac{\langle \vec{v}, \vec{w}^{(2)} \rangle}{\left\| \vec{w}^{(2)} \right\|^2}} \langle \vec{u}, \vec{w}^{(2)} \rangle + \cdots \\
&= \frac{\langle \vec{w}^{(1)}, \vec{v} \rangle}{\left\| \vec{w}^{(1)} \right\|^2} \langle \vec{u}, \vec{w}^{(1)} \rangle + \frac{\langle \vec{w}^{(2)}, \vec{v} \rangle}{\left\| \vec{w}^{(2)} \right\|^2} \langle \vec{u}, \vec{w}^{(2)} \rangle + \cdots
\end{aligned}$$

but

$$\begin{aligned}
\langle T(\vec{u}), \vec{v} \rangle &= \langle \frac{\langle \vec{u}, \vec{w}^{(1)} \rangle}{\left\| \vec{w}^{(1)} \right\|^2} \vec{w}^{(1)} + \frac{\langle \vec{u}, \vec{w}^{(2)} \rangle}{\left\| \vec{w}^{(2)} \right\|^2} \vec{w}^{(2)} + \cdots, \vec{v} \rangle \\
&= \frac{\langle \vec{u}, \vec{w}^{(1)} \rangle}{\left\| \vec{w}^{(1)} \right\|^2} \langle \vec{w}^{(1)}, \vec{v} \rangle + \frac{\langle \vec{u}, \vec{w}^{(2)} \rangle}{\left\| \vec{w}^{(2)} \right\|^2} \langle \vec{w}^{(2)}, \vec{v} \rangle + \cdots \\
&= \langle \vec{u}, T(\vec{v}) \rangle
\end{aligned}$$

Hence we identify $T^\dagger$ 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 with an appropriate inner product.

Definition 12.3.3. Given a countably infinite basis $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \ldots\}$, where $\vec{u}^{(j)} \in \mathcal{V}$ belongs to an inner product space, it is transformed into an orthogonal basis $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \vec{v}^{(3)}, \ldots\}$, $\vec{v}^{(j)} \in \mathcal{V}$, by Gram-Schmidt Orthogonalization according to the following formulae:

$$\begin{aligned}
\vec{v}^{(1)} &= \vec{u}^{(1)} \\
\vec{v}^{(2)} &= \vec{u}^{(2)} - \operatorname{proj}_{\mathcal{V}^{(1)}} \vec{u}^{(2)} = \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\left\| \vec{v}^{(1)} \right\|^2} \vec{v}^{(1)} \\
\vec{v}^{(3)} &= \vec{u}^{(3)} - \operatorname{proj}_{\mathcal{V}^{(1)}} \vec{u}^{(3)} - \operatorname{proj}_{\mathcal{V}^{(2)}} \vec{u}^{(3)} \\
&= \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\left\| \vec{v}^{(1)} \right\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\left\| \vec{v}^{(2)} \right\|^2} \vec{v}^{(2)} \\
&\vdots \\
\vec{v}^{(n)} &= \vec{u}^{(n)} - \operatorname{proj}_{\mathcal{V}^{(1)}} \vec{u}^{(n)} - \operatorname{proj}_{\mathcal{V}^{(2)}} \vec{u}^{(n)} - \cdots - \operatorname{proj}_{\mathcal{V}^{(n-1)}} \vec{u}^{(n)} \\
&= \vec{u}^{(n)} - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(1)} \rangle}{\left\| \vec{v}^{(1)} \right\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(2)} \rangle}{\left\| \vec{v}^{(2)} \right\|^2} \vec{v}^{(2)} - \cdots - \frac{\langle \vec{u}^{(n)}, \vec{v}^{(n-1)} \rangle}{\left\| \vec{v}^{(n-1)} \right\|^2} \vec{v}^{(n-1)}
\end{aligned} \tag{12.12}$$

Generally, for $j \geq 2$, the j-th new orthogonal basis vector is computed by

$$\vec{v}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \operatorname{proj}_{\mathcal{V}^{(k)}} \vec{u}^{(j)} = \vec{u}^{(j)} - \sum_{k=1}^{j-1} \frac{\langle \vec{u}^{(j)}, \vec{v}^{(k)} \rangle}{\left\| \vec{v}^{(k)} \right\|^2} \vec{v}^{(k)} \tag{12.13}$$

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

$$\langle \vec{u}, \vec{v} \rangle = \mathbf{u}^T B \mathbf{v}$$

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 with respect to this inner product.

Solution. We leave it 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 11.1.5 and the inner product makes sense. We calculate each of the expressions appearing in Definition 12.3.3:

$$\left\| \vec{v}^{(1)} \right\|^2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 2$$

$$\begin{aligned} \vec{v}^{(2)} &= \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\left\| \vec{v}^{(1)} \right\|^2} \vec{v}^{(1)} \\ &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \frac{\left(\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

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

$$\left\| \vec{v}^{(2)} \right\|^2 = \begin{bmatrix} -\frac{1}{2} & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} = \frac{3}{2}$$

$$\vec{v}^{(3)} = \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\left\| \vec{v}^{(1)} \right\|^2} \vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\left\| \vec{v}^{(2)} \right\|^2} \vec{v}^{(2)}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \frac{\left(\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$- \frac{\left(\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} \right)}{\frac{3}{2}} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - (0) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix}$$

$$\left\| \vec{v}^{(3)} \right\|^2 = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix} = \frac{4}{3}$$

Therefore, the required orthonormal basis is

$$\left\{ \left(\frac{1}{\sqrt{2}}, 0, 0 \right)^T, \left(-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, 0 \right)^T, \left(\frac{1}{2\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{3}{2\sqrt{3}} \right)^T \right\}$$

found by dividing each of the $\vec{v}^{(j)}$ by $\left\| \vec{v}^{(j)} \right\|$. 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. □

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 counterparts 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 $T : \mathcal{V} \to \mathcal{V}$ where $T^\dagger = T$ and the inner product space $\mathcal{V}$ is a separable Hilbert space, denote its eigenvalues by λ_j, $j = 1, 2, \ldots$, 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 $\lim_{n\to\infty}$) 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 $T_j(\vec{v}) = \langle \vec{v}, \varphi^{(j)} \rangle$, then we have:

(a) $\mathcal{V} = \mathcal{E}_1 \oplus \mathcal{E}_2 \oplus \cdots = \lim_{n\to\infty} \bigoplus_{j=1}^{n} \mathcal{E}_j$;

(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) $I = T_1 + T_2 + \cdots = \lim_{n\to\infty} \sum_{j=1}^{n} T_j$; and

(e) $T = \lambda_1 T_1 + \lambda_2 T_2 + \cdots = \lim_{n\to\infty} \sum_{j=1}^{n} \lambda_j T_j$, where the eigenvalues λ_j are all real due to Properties 12.2.5.

Note that (a) and subsequently (d) are 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 $\vec{v} = c_1\varphi^{(1)} + c_2\varphi^{(2)} + \cdots = \lim_{n\to\infty} \sum_{j=1}^{n} c_j\varphi^{(j)} \in \mathcal{V}$ in the inner product space can be written as a unique (infinite) sum of the eigenvectors which is exactly what a (Schauder) basis indicates, although it is only in the so-called L^2 sense.[15] We put the justification of (b) in the footnote below[16] 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 $\vec{v} = c_1\varphi^{(1)} + c_2\varphi^{(2)} + \cdots = \lim_{n\to\infty} \bigoplus_{j=1}^{n} c_j\varphi^{(j)}$ and by Definition 12.3.1, we have

$$\begin{aligned}
T_j(\vec{v}) &= \langle \vec{v}, \varphi^{(j)}\rangle \varphi^{(j)} \\
&= \langle \cdots + c_{j-1}\varphi^{(j-1)} + c_j\varphi^{(j)} + c_{j+1}\varphi^{(j+1)} + \cdots, \varphi^{(j)}\rangle \varphi^{(j)} \\
&= (\cdots(0) + c_j(1) + \cdots(0))\varphi^{(j)} = c_j\varphi^{(j)}
\end{aligned}$$

and thus

$$\begin{aligned}
I(\vec{v}) = \vec{v} = c_1\varphi^{(1)} + c_2\varphi^{(2)} + \cdots &= \lim_{n\to\infty} \sum_{j=1}^{n} c_j\varphi^{(j)} \\
= \langle \vec{v}, \varphi^{(1)}\rangle\varphi^{(1)} + \langle \vec{v}, \varphi^{(2)}\rangle\varphi^{(2)} + \cdots &= \lim_{n\to\infty} \sum_{j=1}^{n} \langle \vec{v}, \varphi^{(j)}\rangle\varphi^{(j)} \\
= T_1(\vec{v}) + T_2(\vec{v}) + \cdots = (T_1 + T_2 + \cdots)(\vec{v}) &= \left(\lim_{n\to\infty} \sum_{j=1}^{n} T_j\right)(\vec{v})
\end{aligned}$$

[15] We will demonstrate this using *Bessel's Inequality* introduced in Section 14.1.1 for the Fourier basis. But the L^2 convergence is already good enough in many situations.

[16] 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 $\langle \vec{v}, \vec{w}\rangle = 0$. 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.

So $I = \lim_{n\to\infty} \sum_{j=1}^{n} T_j$. (e) is also shown in a similar fashion:

$$\begin{aligned}
T(\vec{v}) &= T(c_1\varphi^{(1)} + c_2\varphi^{(2)} + \cdots) = T\left(\lim_{n\to\infty}\sum_{j=1}^{n} c_j\varphi^{(j)}\right) \\
&= c_1T(\varphi^{(1)}) + c_2T(\varphi^{(2)}) + \cdots = \lim_{n\to\infty}\sum_{j=1}^{n} c_jT(\varphi^{(j)}) && (T \text{ is linear}^{17}) \\
&= \lambda_1c_1\varphi^{(1)} + \lambda_2c_2\varphi^{(2)} + \cdots = \lim_{n\to\infty}\sum_{j=1}^{n} \lambda_jc_j\varphi^{(j)} \\
&= \lambda_1T_1(\vec{v}) + \lambda_2T_2(\vec{v}) + \cdots && (\text{as in deriving (d)}) \\
&= (\lambda_1T_1 + \lambda_2T_2 + \cdots)\vec{v} = \left(\lim_{n\to\infty}\sum_{j=1}^{n} \lambda_jT_j\right)(\vec{v})
\end{aligned}$$

Be wary that in (e), the linear operator T applied on $\vec{v}$ is required to be Hermitian (truly self-adjoint), not only formally self-adjoint.[18] This means that there will be problems 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 $\{\sin(x), \sin(2x), \sin(3x), \ldots, 1, \cos(x), \cos(2x), \cos(3x), \ldots\}$ for the $L^2[-\pi, \pi]$ Hilbert space with the inner product of Equation (12.4). To normalize it, note that each eigenfunction has a norm of $\frac{1}{\sqrt{\pi}}$[19]($\frac{1}{\sqrt{2\pi}}$ for 1) and hence the orthonormal Fourier basis is

$$\begin{aligned}
&\left\{\frac{1}{\sqrt{\pi}}\sin(x), \frac{1}{\sqrt{\pi}}\sin(2x), \frac{1}{\sqrt{\pi}}\sin(3x), \ldots,\right. \\
&\left.\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos(x), \frac{1}{\sqrt{\pi}}\cos(2x), \frac{1}{\sqrt{\pi}}\cos(3x), \ldots\right\}
\end{aligned} \tag{12.14}$$

[17]Notice that in such an infinite series, linearity is actually not sufficient, see the caution just below the derivation.

[18]Curious readers may like to search about what a compact operator is.

According to part (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

$$\begin{aligned} f &= \frac{a_0}{2} + a_1 \cos(x) + a_2 \cos(2x) + a_3 \cos(3x) + \cdots \\ &\quad + b_1 \sin(x) + b_2 \sin(2x) + b_3 \sin(3x) + \cdots \qquad (12.15a) \\ &= \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx) \qquad (12.15b) \end{aligned}$$

an infinite sum of these sinusoidal eigenfunctions (omitted 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 (12.11), leading to

$$\begin{aligned} f &= \sum_j \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \\ &= \left\langle f, \frac{1}{\sqrt{2\pi}} \right\rangle \left(\frac{1}{\sqrt{2\pi}} \right) \\ &\quad + \left\langle f, \frac{1}{\sqrt{\pi}} \cos(x) \right\rangle \left(\frac{1}{\sqrt{\pi}} \cos(x) \right) + \left\langle f, \frac{1}{\sqrt{\pi}} \cos(2x) \right\rangle \left(\frac{1}{\sqrt{\pi}} \cos(2x) \right) + \cdots \\ &\quad + \left\langle f, \frac{1}{\sqrt{\pi}} \sin(x) \right\rangle \left(\frac{1}{\sqrt{\pi}} \sin(x) \right) + \left\langle f, \frac{1}{\sqrt{\pi}} \sin(2x) \right\rangle \left(\frac{1}{\sqrt{\pi}} \sin(2x) \right) + \cdots \end{aligned}$$

[19] We will only show this for the cosines but the calculation is the same for the sines. For any integer m,

$$\begin{aligned} \|\cos(mx)\|^2 &= \int_{-\pi}^{\pi} \cos(mx)\overline{\cos(mx)}dx \\ &= \int_{-\pi}^{\pi} \cos^2(mx)dx \\ &= \int_{-\pi}^{\pi} \frac{1}{2}(1 + \cos(2mx))dx \\ &= [\frac{1}{2}x + \frac{1}{4m}\sin(2mx)]_{-\pi}^{\pi} = \frac{1}{2}(2\pi) + (0) = \frac{1}{\pi} \end{aligned}$$

$$= \frac{1}{2\pi}\langle f, 1\rangle + \frac{1}{\pi}\langle f, \cos(x)\rangle \cos(x) + \frac{1}{\pi}\langle f, \cos(2x)\rangle \cos(2x) + \cdots$$
$$+ \frac{1}{\pi}\langle f, \sin(x)\rangle \sin(x) + \frac{1}{\pi}\langle f, \sin(2x)\rangle \sin(2x) + \cdots$$

Comparing this expression with the form of Fourier series in Equation (12.15), we yield the following formulae for the Fourier 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

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx) \tag{12.16}$$

where the Fourier coefficients are given by

$$a_m = \frac{1}{\pi}\langle f, \cos(mx)\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)dx \tag{12.17a}$$
$$b_n = \frac{1}{\pi}\langle f, \sin(nx)\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx \tag{12.17b}$$

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 (12.17a) and (12.17b). 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.

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin(nx)dx$$
$$= \frac{1}{\pi}[-\frac{1}{n}x\cos(nx)]_{-\pi}^{\pi} - \frac{1}{\pi}\int_{-\pi}^{\pi}(-\frac{1}{n}\cos(nx))dx$$

$$
\begin{aligned}
&= \frac{1}{\pi}[-\frac{1}{n}\pi\cos(n\pi) + \frac{1}{n}(-\pi)\cos(n(-\pi))] + \frac{1}{n\pi}\int_{-\pi}^{\pi}\cos(nx)dx \\
&= -\frac{1}{n}\cos(n\pi) - \frac{1}{n}\cos(-n\pi) + \frac{1}{n^2\pi}[\sin(nx)]_{-\pi}^{\pi} \\
&= -\frac{2(-1)^n}{n} + (0) = -\frac{2(-1)^n}{n}
\end{aligned}
$$

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 the self-adjoint operator $T = \frac{d^2}{dx^2}$ induces the Fourier basis 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?[20] □

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

$$
\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y + \lambda w(x)y = 0 \tag{12.18}
$$

where $p(x)$, $q(x)$ are real functions of x and $w(x)$ is a real positive-definite weighting function in x. It can be written in terms of the ***Sturm-Liouville operator***

$$\mathcal{L}[f] = -\left[\frac{d}{dx}\left(p(x)\frac{d}{dx}\right) + q(x)\right][f] \tag{12.19}$$

as

$$\mathcal{L}[y] = \lambda w(x) y \tag{12.20}$$

which apparently poses an eigenvalue-eigenfunction problem with a weighting $w(x)$ where λ is the eigenvalue. The Sturm-Liouville operator is self-adjoint with respect to the inner product (12.4) and can become Hermitian under suitable boundary conditions as shown below:

$$\begin{aligned}
\langle \mathcal{L}[f], g\rangle &= \int_a^b \mathcal{L}[f(x)]\overline{g(x)}dx \\
&= \int_a^b -\left(\left(\frac{d}{dx}(p(x)\frac{df(x)}{dx}\right) + q(x)f(x)\right)\overline{g(x)}dx \\
&= -\int_a^b \frac{d}{dx}\left(p(x)\frac{df(x)}{dx}\right)\overline{g(x)} - \int_a^b q(x)f(x)\overline{g(x)}dx
\end{aligned}$$

[20] It is because the Hermiticity 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

$$[f(x)\frac{d}{dx}\overline{([g(x)])}]_{-\pi}^{\pi} - [\frac{d}{dx}(f(x))\overline{g(x)}]_{-\pi}^{\pi}$$

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 terms by terms 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 2π, so actually such a periodic extension of the function x will be discontinuous at the boundaries between the repeated graphs.

$$
\begin{aligned}
&= -\left[p(x)\frac{df(x)}{dx}\overline{g(x)}\right]_a^b + \int_a^b \frac{df(x)}{dx}p(x)\frac{d\overline{g(x)}}{dx}dx \\
&\quad - \int_a^b q(x)f(x)\overline{g(x)}dx \\
&= -\left[p(x)\frac{df(x)}{dx}\overline{g(x)}\right]_a^b + \left[f(x)p(x)\frac{d\overline{g(x)}}{dx}\right]_a^b \\
&\quad - \int f(x)\frac{d}{dx}\left(p(x)\frac{d\overline{g(x)}}{dx}\right)dx - \int_a^b q(x)f(x)\overline{g(x)}dx \\
&= -\left[p(x)\frac{df(x)}{dx}\overline{g(x)}\right]_a^b + \left[f(x)p(x)\frac{d\overline{g(x)}}{dx}\right]_a^b \\
&\quad + \int f(x)\left(-\left[\frac{d}{dx}\left(p(x)\frac{d\overline{g(x)}}{dx}\right) + q(x)\overline{g(x)}\right]\right)dx \\
&= -\left[\frac{df(x)}{dx}p(x)\overline{g(x)}\right]_a^b + \left[f(x)p(x)\frac{d\overline{g(x)}}{dx}\right]_a^b \\
&\quad + \int f(x)\overline{\left(-\left[\frac{d}{dx}\left(p(x)\frac{dg(x)}{dx}\right) + q(x)g(x)\right]\right)}dx \\
&\quad (p(x), q(x) \text{ are real}) \\
&= -\left[\frac{df(x)}{dx}p(x)\overline{g(x)}\right]_a^b + \left[f(x)p(x)\frac{d\overline{g(x)}}{dx}\right]_a^b \\
&\quad + \int_a^b f(x)\overline{\mathcal{L}[g(x)]}dx
\end{aligned}
$$

So $\mathcal{L}^\dagger = \mathcal{L}$ 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 $g(x)$[21] so that $\mathcal{L}$

[21]This implies that $[\frac{d\overline{f(x)}}{dx}p(x)g(x)]_a^b = 0$ as well when the roles of $f(x)$ and $g(x)$ are exchanged. Conjugating this relation gives $[\frac{df(x)}{dx}p(x)\overline{g(x)}]_a^b = 0$ so the first boundary term also vanishes.

can be Hermitian. This boundary condition will also have to be obeyed by the eigenfunctions φ_j where $\mathcal{L}[\varphi_j] = \lambda w(x)\varphi_j$. However, sometimes $p(x)$ may take a form such that at the end-points its value may be zero, so that $\mathcal{L}$ is automatically Hermitian over this *natural interval*. The reality of eigenvalues (left as an exercise, see Properties 12.2.5) and the orthogonality of eigenfunctions are enabled by the Hermiticity of the Sturm-Liouville operator as in Properties 12.2.6 but with an extra factor from the weighting function $w(x)$:

$$\begin{aligned}
\int_a^b \mathcal{L}[\varphi_1]\overline{\varphi_2} &= \int_a^b \varphi_1\overline{\mathcal{L}[\varphi_2]} \\
\int_a^b \lambda_1 w(x)\varphi_1\overline{\varphi_2}dx &= \int_a^b \varphi_1\overline{\lambda_2 w(x)\varphi_2}dx \\
\int_a^b \lambda_1 w(x)\varphi_1\overline{\varphi_2}dx &= \int_a^b \lambda_2 w(x)\varphi_1\overline{\varphi_2}dx \\
(\lambda_1 - \lambda_2)\int_a^b w(x)\varphi_1\overline{\varphi_2}dx &= 0
\end{aligned}$$

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 (12.8) assumed that $\lambda_1 \neq \lambda_2$. To transform a general second-order ODE appearing as $P(x)\frac{d^2y}{dx^2} + R(x)\frac{dy}{dx} + Q(x)y + \lambda w(x)y = 0$ into the Sturm-Liouville form, we can multiply it by the integrating factor

$$F(x) = \exp\left(\int \frac{R(x) - P'(x)}{P(x)}dx\right) \tag{12.21}$$

so that it becomes

$$\begin{aligned}
&[F(x)P(x)y']' + F(x)Q(x)y + \lambda F(x)w(x)y \\
&= [p(x)y']' + q(x)y + \lambda F(x)w(x)y = 0
\end{aligned} \tag{12.22}$$

[22] 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***

$$y'' - 2xy' + 2\nu y = 0 \tag{12.23}$$

to the Sturm-Liouville form and find its eigenfunctions.

Solution. The integrating factor, by (12.21), 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

$$\begin{aligned} e^{-x^2} y'' - 2xe^{-x^2} y' + 2\nu e^{-x^2} y &= 0 \\ (e^{-x^2} y')' + 2\nu e^{-x^2} y &= 0 \end{aligned} \tag{12.24}$$

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

$$y = a_0 + a_1 x + a_2 x^2 + \cdots = \sum_{n=0}^{\infty} a_n x^n \tag{12.25}$$

---

[22] $F(x)$ is derived such that $F(x)P(x)y'' + F(x)R(x)y' = [F(x)P(x)y']'$:

$$\begin{aligned} [F(x)P(x)y']' = F(x)P(x)y'' + F(x)R(x)y' &= F(x)P(x)y'' + F(x)P'(x)y' + F'(x)P(x)y' \\ F(x)R(x)y' &= F(x)P'(x)y' + F'(x)P(x)y' \\ F(x)(R(x) - P'(x)) &= F'(x)P(x) \\ \frac{dF(x)}{F(x)} &= \frac{R(x) - P'(x)}{P(x)} \\ \ln F(x) &= \int \frac{R(x) - P'(x)}{P(x)} dx \\ \Rightarrow F(x) &= \exp\left(\int \frac{R(x) - P'(x)}{P(x)} dx\right) \end{aligned}$$

then

$$y' = \sum_{n=1}^{\infty} n a_n x^{n-1} \tag{12.26a}$$

$$y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} \tag{12.26b}$$

Substituting them into the original form of Hermite's equation, we have

$$\sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} - 2x \sum_{n=1}^{\infty} n a_n x^{n-1} + 2\nu \sum_{n=0}^{\infty} a_n x^n = 0$$

$$\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n - 2 \sum_{n=1}^{\infty} n a_n x^n + 2\nu \sum_{n=0}^{\infty} a_n x^n = 0 \tag{12.27}$$

Comparing the coefficients of x^n gives a recurrence relation of

$$(n+2)(n+1) a_{n+2} - 2n a_n + 2\nu a_n = 0 \tag{12.28}$$

for $n \geq 1$. For $n = 0$ it is simply $2a_2 + 2\nu a_0 = 0$. Rearranging we have

$$a_{n+2} = \frac{2(n-\nu)}{(n+2)(n+1)} a_n \tag{12.29}$$

Such a form 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

$$[\varphi_1 p(x) \frac{d\overline{\varphi_2}}{dx}]_a^b = [e^{-x^2} \varphi_1 \frac{d\overline{\varphi_2}}{dx}]_a^b = 0$$

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 solutions is truncated to a polynomial when $\nu = n$. These polynomials are subsequently known as the

Hermite polynomials. 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

$$a_2 = \frac{2(0-2)}{(2)(1)}a_0 = -2a_0$$
$$a_4 = \frac{2(2-2)}{(4)(3)}a_2 = 0$$

So $y = a_0 - 2a_0x^2 = a_0(1 - 2x^2)$. In the same fashion, when $n = 3$, we have

$$a_3 = \frac{2(1-3)}{(3)(2)}a_1 = -\frac{2}{3}a_1$$
$$a_5 = \frac{2(3-3)}{(5)(4)}a_3 = 0$$

hence $y = a_1x - \frac{2}{3}a_1x^3 = a_1(x - \frac{2}{3}x^3)$. By convention, the Hermite polynomials are scaled in a way such that the leading highest degree term has a coefficient of 2^n. So the first four Hermite polynomials are

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= 2x \\ H_2(x) &= 4x^2 - 2 \\ H_3(x) &= 8x^3 - 12x \end{aligned} \tag{12.30}$$

► Short Exercise: Find the Hermite polynomial of degree 4, $H_4(x)$, corresponding to $\nu = n = 4$.[23] □

12.4.2. Generating Special Polynomials by Gram-Schmidt Orthogonalization

Since a Sturm-Liouville operator is Hermitian given appropriate boundary conditions and Hermiticity 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 polynomials in the previous part, by applying the Gram-Schmidt process on the standard polynomial basis $\{1, x, x^2, x^3, \ldots\}$ to any degree with respect to the inner product in Equation (12.8). The details of the calculation are shown below.

Example 12.4.2. Compute the first four Hermite polynomials as introduced in Example 12.4.1 by Gram-Schmidt Orthogonalization.

Solution. The Hermite polynomials have a weighting of $w(x) = e^{-x^2}$ and are integrated over the entire real axis $(-\infty, \infty)$. It is instructive to first note 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 the footnote below)[24] Now we apply Gram-Schmidt Orthogonalization on the standard polynomials $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}, \vec{u}^{(4)}, \ldots\} = \{1, x, x^2, x^3, \ldots\}$ according

[23] Using the recurrence relation, we have

$$a_2 = \frac{2(0-4)}{(2)(1)} a_0 = -4a_0$$
$$a_4 = \frac{2(2-4)}{(4)(3)} a_2 = -\frac{1}{3} a_2 = \frac{4}{3} a_0$$
$$a_6 = \frac{2(4-4)}{(6)(5)} a_4 = 0$$

hence $y = a_0 - 4a_0x^2 + \frac{4}{3}a_0x^4$ and after scaling it becomes $H_4(x) = 16x^4 - 48x^2 + 12$.

[24] It is a well-known result from multivariable calculus that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

to Definition 12.3.3. The zeroth degree Hermite polynomial is trivially $\vec{v}^{(1)} = H_0(x) = 1$ and $\left\|\vec{v}^{(1)}\right\|^2 = \sqrt{\pi}$. Next, for $n = 1$, we have

$$\begin{aligned}
\vec{v}^{(2)} &= \vec{u}^{(2)} - \frac{\langle \vec{u}^{(2)}, \vec{v}^{(1)} \rangle}{\left\|\vec{v}^{(1)}\right\|^2} \vec{v}^{(1)} \\
&= x - \frac{\int_{-\infty}^{\infty} e^{-x^2}(x)(1)dx}{\left\|\vec{v}^{(1)}\right\|^2} \vec{v}^{(1)} \\
&= x - \frac{\int_{-\infty}^{\infty} xe^{-x^2}dx}{\sqrt{\pi}}(1) = x - (0)(1) = x
\end{aligned}$$

So the Hermite polynomial of degree 1 is just $\vec{v}^{(2)} \leftarrow H_1(x) = 2x$ (scaled by convention, similar for the following). Now

$$\begin{aligned}
\left\|\vec{v}^{(2)}\right\|^2 &= \int_{-\infty}^{\infty} (2x)^2 e^{-x^2} dx \\
&= 4 \int_{-\infty}^{\infty} x^2 e^{-x^2} dx
\end{aligned}$$

---

for the Gaussian integral. By mathematical induction and integration by parts, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} x^{2(m+1)} e^{-x^2} dx &= \int_{-\infty}^{\infty} x^{2m+2} e^{-x^2} dx \\
&= \int_{-\infty}^{\infty} -\frac{1}{2} x^{2m+1} (-2xe^{-x^2}) dx \\
&= \int_{-\infty}^{\infty} -\frac{1}{2} x^{2m+1} d(e^{-x^2}) \\
&= [-\frac{1}{2} x^{2m+1} e^{-x^2}]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-\frac{2m+1}{2} x^{2m}) e^{-x^2} dx \\
&= (0) + \frac{2m+1}{2} \int_{-\infty}^{\infty} x^{2m} e^{-x^2} dx \\
&= \frac{(1)(3)(5)\cdots(2m-1)(2m+1)\sqrt{\pi}}{2^{(m+1)}} \\
&\quad \text{(via the induction hypothesis)}
\end{aligned}$$

so the formula is established.

$$= 4\left(\frac{1}{2}\sqrt{\pi}\right) = 2\sqrt{\pi}$$

For $n = 2$, we have

$$\begin{aligned}
\vec{v}^{(3)} &= \vec{u}^{(3)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(1)} \rangle}{\left\|\vec{v}^{(1)}\right\|^2}\vec{v}^{(1)} - \frac{\langle \vec{u}^{(3)}, \vec{v}^{(2)} \rangle}{\left\|\vec{v}^{(2)}\right\|^2}\vec{v}^{(2)} \\
&= x^2 - \frac{\int_{-\infty}^{\infty}(x^2)(1)e^{-x^2}dx}{\sqrt{\pi}}(1) - \frac{\int_{-\infty}^{\infty}(x^2)(2x)e^{-x^2}dx}{2\sqrt{\pi}}(2x) \\
&= x^2 - \frac{\int_{-\infty}^{\infty}x^2e^{-x^2}dx}{\sqrt{\pi}}(1) - \frac{2\int_{-\infty}^{\infty}x^3e^{-x^2}dx}{2\sqrt{\pi}}(2x) \\
&= x^2 - \frac{(\frac{1}{2}\sqrt{\pi})}{\sqrt{\pi}}(1) - (0)(2x) \\
&= x^2 - \frac{1}{2}
\end{aligned}$$

Hence $\vec{v}^{(3)} \leftarrow H_2(x) = 4x^2 - 2$, and $\left\|\vec{v}^{(3)}\right\|^2 = \int_{-\infty}^{\infty}(4x^2 - 2)^2e^{-x^2}dx = \int_{-\infty}^{\infty}(16x^4 - 16x^2 + 4)e^{-x^2}dx = 16(\frac{3}{4}\sqrt{\pi}) - 16(\frac{1}{2}\sqrt{\pi}) + 4(\sqrt{\pi}) = 8\sqrt{\pi}$. Finally, for $n = 3$

$$\begin{aligned}
\vec{v}^{(4)} &= \vec{u}^{(4)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(1)} \rangle}{\left\|\vec{v}^{(1)}\right\|^2}\vec{v}^{(1)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(2)} \rangle}{\left\|\vec{v}^{(2)}\right\|^2}\vec{v}^{(2)} - \frac{\langle \vec{u}^{(4)}, \vec{v}^{(3)} \rangle}{\left\|\vec{v}^{(3)}\right\|^2}\vec{v}^{(3)} \\
&= x^3 - \frac{\int_{-\infty}^{\infty}(x^3)(1)e^{-x^2}dx}{\sqrt{\pi}}(1) - \frac{\int_{-\infty}^{\infty}(x^3)(2x)e^{-x^2}dx}{2\sqrt{\pi}}(2x) \\
&\quad - \frac{\int_{-\infty}^{\infty}(x^3)(4x^2 - 2)e^{-x^2}dx}{8\sqrt{\pi}}(4x^2 - 2) \\
&= x^3 - (0)(1) - \frac{2\int_{-\infty}^{\infty}x^4e^{-x^2}dx}{2\sqrt{\pi}}(2x) - (0)(4x^2 - 2) \\
&= x^3 - \left(\frac{3\sqrt{\pi}}{4\sqrt{\pi}}\right)(2x) = x^3 - \frac{3}{2}x
\end{aligned}$$

thus $\vec{v}^{(4)} \leftarrow H_3(x) = 8x^3 - 12x$. □

12.5. Earth Science Applications

Example 12.5.1. Consider the *linearized shallow water system* along the Equator:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \beta y v = -\dfrac{\partial \Phi}{\partial x} & (12.31a) \\ \dfrac{\partial v}{\partial t} + \beta y u = -\dfrac{\partial \Phi}{\partial y} & (12.31b) \\ \dfrac{\partial \Phi}{\partial t} = -gH\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}\right) & (12.31c) \end{cases}$$

The first two equations are the momentum equations, and the last one is the continuity equation. u, v, Φ are the zonal, meridional winds, and the geopotential. The constants include β which denotes the beta effect (the change of f over latitudes), and g, H as the gravitational acceleration and equivalent depth. By assuming a traveling wave solution (written in the complex form, i is the imaginary number)

$$u = \hat{u}(y)\exp(i(kx - \omega t)) \tag{12.32a}$$
$$v = \hat{v}(y)\exp(i(kx - \omega t)) \tag{12.32b}$$
$$\Phi = \hat{\Phi}(y)\exp(i(kx - \omega t)) \tag{12.32c}$$

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 change of variables to arrive at the Hermite's equation, and hence derive the dynamical modes of the linearized shallow water system.

Solution. Plugging the ansatz in, the system of equations becomes

$$\begin{aligned} -i\omega\hat{u}(y)e^{i(kx-\omega t)} - \beta y\hat{v}(y)e^{i(kx-\omega t)} &= -ik\hat{\Phi}(y)e^{i(kx-\omega t)} \\ -i\omega\hat{u}(y) - \beta y\hat{v}(y) &= -ik\hat{\Phi}(y) \end{aligned} \tag{12.33}$$

$$-i\omega\hat{v}(y)e^{i(kx-\omega t)} + \beta y\hat{u}(y)e^{i(kx-\omega t)} = -\frac{d\hat{\Phi}(y)}{dy}e^{i(kx-\omega t)}$$
$$-i\omega\hat{v}(y) + \beta y\hat{u}(y) = -\frac{d\hat{\Phi}(y)}{dy} \tag{12.34}$$

and

$$-i\omega\hat{\Phi}(y)e^{i(kx-\omega t)} = -gH\left(ik\hat{u}(y)e^{i(kx-\omega t)} + \frac{d\hat{v}(y)}{dy}e^{i(kx-\omega t)}\right)$$
$$-i\omega\hat{\Phi}(y) = -gH\left(ik\hat{u}(y) + \frac{d\hat{v}(y)}{dy}\right) \tag{12.35}$$

Differentiating (12.35) gives

$$-i\omega\frac{d\hat{\Phi}(y)}{dy} = -gH\left(ik\frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2}\right) \tag{12.36}$$

Substituting (12.34) into (12.36) leads to

$$i\omega(-i\omega\hat{v}(y) + \beta y\hat{u}(y)) = -gH\left(ik\frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2}\right)$$
$$\omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y) = -gH\left(ik\frac{d\hat{u}(y)}{dy} + \frac{d^2\hat{v}(y)}{dy^2}\right) \tag{12.37}$$

Using (12.33) in (12.37) gives

$$\omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y) = -gH\left(-\frac{k}{\omega}\frac{d}{dy}(-ik\hat{\Phi}(y) + \beta y\hat{v}(y)) + \frac{d^2\hat{v}(y)}{dy^2}\right)$$
$$\omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y) = \frac{-igHk^2}{\omega}\frac{d\hat{\Phi}(y)}{dy} + \frac{gHk}{\omega}\left(\beta\hat{v}(y) + \beta y\frac{d\hat{v}(y)}{dy}\right)$$
$$- gH\frac{d^2\hat{v}(y)}{dy^2} \tag{12.38}$$

Applying both (12.34) and (12.35) in (12.38) yields

$$\omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y)$$

$$= \frac{igHk^2}{\omega}(-i\omega\hat{v}(y) + \beta y\hat{u}(y)) + \frac{gHk}{\omega}\left(\beta\hat{v}(y) + \beta y\left(\frac{i\omega}{gH}\hat{\Phi}(y) - ik\hat{u}(y)\right)\right)$$
$$- gH\frac{d^2\hat{v}(y)}{dy^2} \tag{12.39}$$

Rearranging then gives

$$\begin{aligned}
&\omega^2\hat{v}(y) + i\omega\beta y\hat{u}(y) \\
&= gHk^2\hat{v}(y) + \frac{igHk^2}{\omega}\beta y\hat{u}(y) + \frac{gHk}{\omega}\beta\hat{v}(y) + ik\beta y\hat{\Phi}(y) \\
&\quad - \frac{igHk^2}{\omega}\beta y\hat{u}(y) - gH\frac{d^2\hat{v}(y)}{dy^2} \\
&= gHk^2\hat{v}(y) + \frac{gHk}{\omega}\beta\hat{v}(y) + ik\beta y\hat{\Phi}(y) - gH\frac{d^2\hat{v}(y)}{dy^2}
\end{aligned} \tag{12.40}$$

Finally, using (12.33) again, we have

$$\omega^2\hat{v}(y) - \beta^2 y^2\hat{v}(y) = gHk^2\hat{v}(y) + \frac{gHk}{\omega}\beta\hat{v}(y) - gH\frac{d^2\hat{v}(y)}{dy^2} \tag{12.41}$$

Cleaning this up, we obtain

$$\frac{d^2\hat{v}(y)}{dy^2} + \left[\left(\frac{\omega^2}{gH} - k^2 - \beta\frac{k}{\omega}\right) - \frac{\beta^2}{gH}y^2\right]\hat{v}(y) = 0 \tag{12.42}$$

Finally, with a change of variable $\tilde{y} = y/y_0$, $y_0 = (\sqrt{gH}/\beta)^{1/2}$, it becomes

$$\begin{aligned}
\frac{\beta}{\sqrt{gH}}\frac{d^2\hat{v}(\tilde{y})}{d\tilde{y}^2} + \left[\left(\frac{\omega^2}{gH} - k^2 - \beta\frac{k}{\omega}\right) - \frac{\beta^2}{gH}\left(\frac{\sqrt{gH}}{\beta}\tilde{y}^2\right)\right]\hat{v}(\tilde{y}) = 0 \\
\frac{d^2\hat{v}(\tilde{y})}{d\tilde{y}^2} + \left[\frac{\sqrt{gH}}{\beta}\left(\frac{\omega^2}{gH} - k^2 - \beta\frac{k}{\omega}\right) - \tilde{y}^2\right]\hat{v}(\tilde{y}) = 0
\end{aligned} \tag{12.43}$$

Letting

$$\mu = \frac{\sqrt{gH}}{\beta}\left(\frac{\omega^2}{gH} - k^2 - \beta\frac{k}{\omega}\right) \tag{12.44}$$

simplifies the equation to

$$\hat{v}'' + (\mu - \tilde{y}^2)\hat{v} = 0 \tag{12.45}$$

Further, if

$$\hat{v} = \hat{w}\exp\left(-\frac{\tilde{y}^2}{2}\right) \tag{12.46}$$

then

$$\hat{v}' = \hat{w}'\exp\left(-\frac{\tilde{y}^2}{2}\right) - \tilde{y}\hat{w}\exp\left(-\frac{\tilde{y}^2}{2}\right) \tag{12.47}$$

and

$$\begin{aligned}\hat{v}'' &= \hat{w}''\exp\left(-\frac{\tilde{y}^2}{2}\right) - \tilde{y}\hat{w}'\exp\left(-\frac{\tilde{y}^2}{2}\right) - \hat{w}\exp\left(-\frac{\tilde{y}^2}{2}\right)\\ &\quad - \tilde{y}\hat{w}'\exp\left(-\frac{\tilde{y}^2}{2}\right) + \tilde{y}^2\hat{w}\exp\left(-\frac{\tilde{y}^2}{2}\right)\\ &= (\hat{w}'' - 2\tilde{y}\hat{w}' + (\tilde{y}^2 - 1)\hat{w})\exp\left(-\frac{\tilde{y}^2}{2}\right)\end{aligned} \tag{12.48}$$

So

$$\hat{v}'' + (\mu - \tilde{y}^2)\hat{v} = 0 \tag{12.49}$$

becomes

$$\begin{aligned}(\hat{w}'' - 2\tilde{y}\hat{w}' + (\tilde{y}^2 - 1)\hat{w})\exp\left(-\frac{\tilde{y}^2}{2}\right) + (\mu - \tilde{y}^2)\hat{w}\exp\left(-\frac{\tilde{y}^2}{2}\right) &= 0\\ \hat{w}'' - 2\tilde{y}\hat{w}' + (\mu - 1)\hat{w} &= 0\end{aligned} \tag{12.50}$$

This is Hermite's equation in disguise, where we have $\mu = 2\nu + 1$, $\nu = 0, 1, 2, \ldots$ if the solutions have to be bounded. Hence the equatorial wave modes are given in terms of $\hat{w} = H_n(\tilde{y})$ and

$$\hat{v}(\tilde{y}) = \hat{v}_0 H_n(\tilde{y})\exp\left(-\frac{\tilde{y}^2}{2}\right) \tag{12.51}$$

From this, $\hat{u}(\tilde{y})$ and $\hat{\Phi}(\tilde{y})$ can be found via any two of (12.33), (12.34) and (12.35). □

Figure 12.1.: *Horizontal structures of selected zonally propagating wave solutions to the shallow water equations on an equatorial β plane. (Kiladis et al., 2009 [4])*

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 a 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 comes the main part of executing the Gram-Schmidt procedure. The inner loop subtracts the parallel components of each previous vector from the current vector and the outer loop iterates the calculation for every input 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 $\langle A, B\rangle = \mathrm{tr}(AB^*)$ 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

$$\langle \vec{u}, \vec{v}\rangle = \mathbf{u}^T B\mathbf{v}$$

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

$$\langle \vec{u}, \vec{v} \rangle = \mathbf{u}^T B \mathbf{v}$$

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 *Triangle Inequality*

$$\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\| \tag{12.52}$$

as in Exercise 4.5 but now for any inner product.

Exercise 12.5 Show that x^3 and $\cos(2x)$ are orthogonal to each other in the $L^2[-\pi, \pi]$ space with respect to the inner product of Equation (12.4).

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 (12.8)

$$\langle f, g\rangle = \int_a^b w(x)f(x)\overline{g(x)}dx$$

where $w(x) =$ (a) 1, and (b) e^{-x}, with $a > 0$.

Exercise 12.7 Show that given a real inner product induced by $\langle \vec{u}, \vec{v}\rangle = \mathbf{u}^T B\mathbf{v}$ where

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

The linear operator that has a matrix representation of

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

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*

$$(1 - x^2)y'' - 2xy' + \lambda y = 0 \tag{12.53}$$

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 will 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 optimized linear equation in terms of some independent variable(s). For example, in Earth Science, we may regress the frequency of tropical cyclones in a basin against the large-scale vorticity field or wind shear.

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 terms of distance, i.e. the squared error

$$\left\|\vec{h} - \vec{h}_f\right\|^2 = \left\|\vec{h} - A\vec{x}_f\right\|^2 \tag{13.1}$$

is minimized, and $\vec{h}_f = A\vec{x}_f$ (or $\vec{x}_f$) is then referred to as the ***least-square approximation*** to the system. From Properties 10.3.3, we know that the shortest distance will be achieved by the orthogonal projection of $\vec{h}$ 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 $C(A)^{\perp}$[1] of the column space $C(A)$ of A, and we have

$$\langle A\vec{x}, \vec{h} - A\vec{x}_f\rangle = \langle \vec{x}, A^\dagger(\vec{h} - A\vec{x}_f)\rangle = 0 \tag{13.2}$$

Since this holds for any $\vec{x}$, by the last item of Properties 12.1.2 we have

$$\begin{aligned} A^\dagger(\vec{h} - A\vec{x}_f) &= \mathbf{0} \\ A^\dagger A\vec{x}_f &= A^\dagger\vec{h} \end{aligned} \tag{13.3}$$

This is called the ***normal equation*** due to the appearance of the $A^\dagger A$ factor. So any $\vec{x}_f$[2] satisfying this equation will produce the least-square error. We may be tempted to multiply both sides of the equation by the inverse $(A^\dagger A)^{-1}$ to arrive at a formula for the least-square solution $\vec{x}_f = (A^\dagger A)^{-1}A^\dagger\vec{h}$. However, this is only allowed when this inverse indeed exists, and we now discuss under what condition it will happen.

Properties 13.1.1. For an $m \times n$ matrix A, $A^\dagger A$ and A have the same null space.

[1] which may not be equal to $\mathcal{N}(A^T)$ or $\mathcal{N}(A^*)$ but rather $\mathcal{N}(A^\dagger)$ if an inner product other than the standard one is used.

[2] The existence of at least one of such a vector is guaranteed by the uniqueness of orthogonal projection of $\vec{h}$ onto $C(A)$.

Figure 13.1.: *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 3×2 matrix A, that is, $\vec{h}_f = A\vec{x}_f$, are $\mathbb{R}^3$ vectors.*

Proof. It is obvious that if $\vec{x} \in \mathcal{N}(A)$, $A\vec{x} = \mathbf{0}$ then $A^\dagger A\vec{x} = \mathbf{0}$ so $\vec{x} \in \mathcal{N}(A^\dagger A)$ and $\mathcal{N}(A) \subseteq \mathcal{N}(A^\dagger A)$. Now we only need to show that $\mathcal{N}(A^\dagger A) \subseteq \mathcal{N}(A)$. For $\vec{x} \in \mathcal{N}(A^\dagger A)$, we have $A^\dagger A\vec{x} = \mathbf{0}$ and hence $\langle \vec{x}, A^\dagger A\vec{x} \rangle = 0$. 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 $\mathcal{N}(A^\dagger A) \subseteq \mathcal{N}(A)$. Hence $\mathcal{N}(A^\dagger A) = \mathcal{N}(A)$, the null space of $A^\dagger A$ and A coincides. □

Properties 13.1.2. For an $m \times n$ matrix A, $A^\dagger A$ has the same rank as A. As a corollary, if A has linearly independent columns such that $\operatorname{rank}(A) = n$, then $A^\dagger A$ is invertible.

Proof. By the Rank-nullity Theorem 6.3.9, $\operatorname{rank}(A^\dagger A) + \operatorname{nullity}(A^\dagger A) = n = \operatorname{rank}(A) + \operatorname{nullity}(A)$ but by the previous properties $\operatorname{nullity}(A^\dagger A) = \operatorname{nullity}(A)$ so $\operatorname{rank}(A^\dagger A) = \operatorname{rank}(A)$ where $A^\dagger A$ is an $n \times n$ matrix. Therefore if $\operatorname{rank}(A^\dagger A) = \operatorname{rank}(A) = n$, then $A^\dagger A$ is full-rank and by Properties 6.3.10 it is invertible. □

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

$$\vec{x}_f = (A^\dagger A)^{-1} A^\dagger \vec{h} \tag{13.4}$$

such that the squared error $\left\|\vec{h} - \vec{h}_f\right\|^2 = \left\|\vec{h} - A\vec{x}_f\right\|^2$ 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 $A^\dagger = A^T$ and we by chance have the QR decomposition of A, then

$$\begin{aligned}
\vec{x}_f &= (A^T A)^{-1} A^T \vec{h} \\
&= ((QR)^T (QR))^{-1} (QR)^T \vec{h} \\
&= (R^T Q^T Q R)^{-1} (QR)^T \vec{h} \\
&= (R^T R)^{-1} R^T Q^T \vec{h} \\
&= R^{-1} (R^T)^{-1} R^T Q^T \vec{h} \\
&= R^{-1} Q^T \vec{h}
\end{aligned} \tag{13.5}$$

where $Q^T Q = I$ since Q is an 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

$$\begin{aligned}
\vec{x}_f &= (A^T A)^{-1} A^T \vec{h} \\
&= \left(\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix}^T \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix}^T \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\
&= \begin{bmatrix} 11 & 18 \\ 18 & 36 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{11}{72} \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{19}{12} \end{bmatrix}
\end{aligned}$$

and the least-square error is

$$\begin{aligned}
\left\| \vec{h} - A\vec{x}_f \right\|^2 &= \left\| \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{19}{12} \end{bmatrix} \right\|^2 \\
&= \left\| \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} - \begin{bmatrix} \frac{11}{3} \\ \frac{47}{6} \\ \frac{41}{6} \end{bmatrix} \right\|^2
\end{aligned}$$

$$= \left\| \begin{bmatrix} -\frac{2}{3} \\ \frac{1}{6} \\ \frac{1}{6} \end{bmatrix} \right\|^2 = \left(-\frac{2}{3}, \frac{1}{6}, \frac{1}{6}\right)^T \cdot \left(-\frac{2}{3}, \frac{1}{6}, \frac{1}{6}\right)^T = \frac{1}{2}$$

For the inner product in Example 12.3.1, an appropriate adjoint, adapted from Section 12.2.1, to be used in this situation is

$$\begin{aligned} A^\dagger &= C^{-1} A^T B \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \end{aligned}$$

where C can be picked to be any positive-definite matrix[3] and we choose $C = I$ for simplicity. Subsequently, the least-square solution is

$$\begin{aligned} \vec{x}_f &= (A^\dagger A)^{-1} A^\dagger \vec{h} \\ &= \left(\begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\ &= \begin{bmatrix} 34 & 62 \\ 62 & 120 \end{bmatrix}^{-1} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} \\ &= \begin{bmatrix} \frac{30}{59} & -\frac{31}{118} \\ -\frac{31}{118} & \frac{17}{118} \end{bmatrix} \begin{bmatrix} 5 & 8 & 5 \\ 8 & 14 & 12 \end{bmatrix} \begin{bmatrix} 3 \\ 8 \\ 7 \end{bmatrix} = \begin{bmatrix} \frac{10}{59} \\ \frac{103}{59} \end{bmatrix} \end{aligned}$$

[3]The choice of C actually matters if there are multiple least-square solutions and $\vec{x}_f$ is also required to have a minimal norm with respect to some other inner product. This will be explored in Section 16.4.

with the least-square error being

$$\begin{aligned}
\left\|\vec{h} - A\vec{x}_f\right\|^2 &= (\mathbf{h} - A\mathbf{x}_f)^T B(\mathbf{h} - A\mathbf{x}_f) \\
&= \left(\begin{bmatrix}3\\8\\7\end{bmatrix} - \begin{bmatrix}1&2\\3&4\\1&4\end{bmatrix}\begin{bmatrix}\frac{10}{59}\\\frac{103}{59}\end{bmatrix}\right)^T \begin{bmatrix}2&1&0\\1&2&1\\0&1&2\end{bmatrix}\left(\begin{bmatrix}3\\8\\7\end{bmatrix} - \begin{bmatrix}1&2\\3&4\\1&4\end{bmatrix}\begin{bmatrix}\frac{10}{59}\\\frac{103}{59}\end{bmatrix}\right) \\
&= \begin{bmatrix}-\frac{39}{59}\\\frac{30}{59}\\-\frac{9}{59}\end{bmatrix}^T \begin{bmatrix}2&1&0\\1&2&1\\0&1&2\end{bmatrix}\begin{bmatrix}-\frac{39}{59}\\\frac{30}{59}\\-\frac{9}{59}\end{bmatrix} = \frac{36}{59}
\end{aligned}$$

□

We close this section by confirming that the matrix $T = A(A^\dagger A)^{-1}A^\dagger$ as in $\vec{h}_f = A\vec{x}_f = A(A^\dagger A)^{-1}A^\dagger\vec{h}$ 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 $T^2 = T = T^\dagger$. For the first equality, we have

$$\begin{aligned}
T^2 &= (A(A^\dagger A)^{-1}A^\dagger)(A(A^\dagger A)^{-1}A^\dagger) \\
&= A(A^\dagger A)^{-1}(A^\dagger A)(A^\dagger A)^{-1}A^\dagger \\
&= A(A^\dagger A)^{-1}(I)A^\dagger = A(A^\dagger A)^{-1}A^\dagger = T
\end{aligned}$$

and for the second equality, we can use Properties 12.2.3 to get

$$\begin{aligned}
T^\dagger &= (A(A^\dagger A)^{-1}A^\dagger)^\dagger \\
&= (A^\dagger)^\dagger((A^\dagger A)^{-1})^\dagger A^\dagger \\
&= A((A^\dagger A)^\dagger)^{-1}A^\dagger \\
&= A(A^\dagger A)^{-1}A^\dagger = T
\end{aligned}$$

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 on least-square approximation. The simplest type of linear regression is fitting a straight line $y = \alpha + \beta x$, where α and β are its intercept and slope, to m pairs of observation, $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ such that the sum of squared errors $\sum_{k=1}^{m}(y_k - (\alpha + \beta x_k))^2$ is minimized. In this context, x and y are known as the *explanatory* and *response variable* respectively.

Figure 13.2.: *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 squares 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 $\vec{y} = (y_1, y_2, \ldots, y_m)^T$,

while the values predicted by the best-fit straight line will be in the form of $\vec{y}^{\text{pred}} = (\alpha\mathbf{1} + \beta\vec{x})^T = (\alpha + \beta x_1, \alpha + \beta x_2, \ldots, \alpha + \beta x_m)^T$, where $\mathbf{1}$ is a column vector filled with ones, α and β 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 α and β in the matrix system

$$\alpha\mathbf{1} + \beta\vec{x} = \vec{y} \tag{13.6}$$

or alternatively

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{13.7}$$

so that the squared error $\sum_{k=1}^{m}(y_k - y_k^{\text{pred}})^2 = \sum_{k=1}^{m}(y_k - (\alpha + \beta x_k))^2$ 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

$$\sum_{k=1}^{m}(y_k - (\alpha + \beta x_k))^2 = \left\|\vec{y} - [X]\vec{\beta}\right\|^2 \tag{13.8}$$

will be minimized by the best-fit parameters which are computed via

$$\vec{\beta}_f = ([X]^T[X])^{-1}[X]^T\vec{y} \tag{13.9}$$

according to Theorem 13.1.3 where $A = [X]$, and simply $A^\dagger = [X]^T$ 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

$$\begin{bmatrix} \alpha_f \\ \beta_f \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} m & \sum X \\ \sum X & \sum X^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} \\
&= \frac{1}{m(\sum X^2) - (\sum X)^2} \begin{bmatrix} \sum X^2 & -\sum X \\ -\sum X & m \end{bmatrix} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix} \\
&= \frac{1}{m(\sum X^2) - (\sum X)^2} \begin{bmatrix} (\sum Y)(\sum X^2) - (\sum X)(\sum XY) \\ m(\sum XY) - (\sum X)(\sum Y) \end{bmatrix} \qquad (13.10)
\end{aligned}
$$

where

$$\sum X = x_1 + x_2 + \cdots + x_m \qquad (13.11a)$$

$$\sum X^2 = x_1^2 + x_2^2 + \cdots + x_m^2 \qquad (13.11b)$$

$$\sum Y = y_1 + y_2 + \cdots + y_m \qquad (13.11c)$$

$$\sum XY = x_1 y_1 + x_2 y_2 + \cdots + x_m y_m \qquad (13.11d)$$

and we have used the results (2.17) in Example 2.3.5 to calculate the inverse from the second to third line.

Properties 13.2.1. The best-fit parameters for the single predictor variable linear regression are found by

$$\alpha_f = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{m(\sum X^2) - (\sum X)^2} \qquad (13.12a)$$

$$\beta_f = \frac{m(\sum XY) - (\sum X)(\sum Y)}{m(\sum X^2) - (\sum X)^2} \qquad (13.12b)$$

such that the regression line $y = \alpha_f + \beta_f x$ achieves the least-square error, i.e. the value of $\sum_{k=1}^{m}(y_k - (\alpha_f + \beta_f x_k))^2$ is the smallest possible.

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.

$$\begin{aligned}\sum X &= 2+3+4+5+7 = 21\\ \sum(X^2) &= 2^2+3^2+4^2+5^2+7^2 = 103\\ \sum Y &= 4+6+7+9+11 = 37\\ \sum(XY) &= (2)(4)+(3)(6)+(4)(7)+(5)(9)+(7)(11) = 176\end{aligned}$$

By Formula (13.12) in Properties 13.2.1, with $m = 5$, the required parameters are

$$\begin{aligned}\alpha_f &= \frac{\sum Y \sum(X^2) - \sum X \sum(XY)}{m\sum(X^2) - (\sum X)^2} = \frac{(37)(103)-(21)(176)}{(5)(103)-(21)^2} \approx 1.554\\ \beta_f &= \frac{m\sum(XY) - (\sum X)(\sum Y)}{m\sum(X^2) - (\sum X)^2} = \frac{(5)(176)-(21)(37)}{(5)(103)-(21)^2} \approx 1.392\end{aligned}$$

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 $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ as there can be different factors contributing to a phenomenon in any Earth Science scenario. Linear regression for multiple predictor variables will then produce a best-fit equation in the form of $y = \alpha + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_n x^{(n)}$. 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 involve more predictor variables, $\vec{\beta}_f^T = (\alpha_f, \beta_{1f}, \beta_{2f}, \cdots, \beta_{nf})^T$, and the matrix $[X] = [\mathbf{1}|\vec{x}^{(1)}|\vec{x}^{(2)}|\cdots|\vec{x}^{(n)}]$ holds the observed values of those multiple predictor variables column by column.

Example 13.2.2. The university carried out an experiment and tried to quantify the effects of IQ and studying time on the GPA of students. 5 students participated and the data are shown below.

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

Do 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

$$\begin{aligned} \vec{\beta}_f &= ([X]^T[X])^{-1}[X]^T\vec{y} \\ &= \left(\begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix}^T \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 104 & 5.2 \\ 1 & 111 & 6.1 \\ 1 & 107 & 8.3 \\ 1 & 106 & 3.4 \\ 1 & 109 & 9.6 \end{bmatrix}^T \begin{bmatrix} 3.56 \\ 3.71 \\ 3.73 \\ 3.34 \\ 3.88 \end{bmatrix} \end{aligned}$$

$$\approx \begin{bmatrix} 1.4463 \\ 0.0162 \\ 0.0710 \end{bmatrix}$$

So the regression model is 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, although 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 n polynomials in a single variable x to m points, $n < m$, will involve a matrix in the form like the one above:

$$[X] = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix} \tag{13.13}$$

This class of matrices is called ***Vandermonde matrices***.[4]

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 error 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 $A^\dagger(\vec{h} - A\vec{x}_f) = A^\dagger(\vec{h} - \vec{h}_f) = \mathbf{0}$. For linear regression, $A^\dagger = [X]^T$ 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 $\overline{y}$, each actual data and predicted values as y_i and $\hat{y}_i = y_i^{\text{pred}}$ respectively, then

$$\sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \overline{y})^2 \tag{13.14}$$

The term on L.H.S. is called the ***SST/Sum of Squares Total***, while the two terms on R.H.S. are known as the ***SSE/Sum of Squares Error*** and ***SSR/Sum of Squares Regression***. To prove this, we expand the SST, which gives

$$\begin{aligned}
\text{SST} &= \sum_i (y_i - \overline{y})^2 \\
&= \sum_i ((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}))^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \overline{y})^2 + 2\sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \overline{y})) \\
&= \text{SSE} + \text{SSR} + 2\sum_i ((y_i - \hat{y}_i)(\hat{y}_i - \overline{y}))
\end{aligned} \tag{13.15}$$

[4] As long as $n < m$ and the sampled points x_i are distinct, the columns of a Vandermonde matrix will be linearly independent.

The remaining step is to show that the last term equals zero. For simplicity, we work with the case of a single predictor variable, so there are two parameters α and β only, but it can be easily generalized to multiple predictors and β_j. Expanding the product gives

$$\begin{aligned}\sum_i((y_i-\hat{y}_i)(\hat{y}_i-\overline{y})) &= \sum_i(\hat{y}_i(y_i-\hat{y}_i)) - \overline{y}\sum_i(y_i-\hat{y}_i)\\ &= \sum_i((\alpha+\beta x_i)(y_i-\hat{y}_i)) - \overline{y}\sum_i(y_i-\hat{y}_i)\\ &= \beta\sum_i(x_i(y_i-\hat{y}_i)) - (\overline{y}-\alpha)\sum_i(y_i-\hat{y}_i)\\ &= \beta\sum_i(x_ie_i) - (\overline{y}-\alpha)\sum_i e_i \qquad (13.16)\end{aligned}$$

Using the same logic when we are investigating $[X]^T\vec{e} = \mathbf{0}$ before, we know that

$$\mathbf{1}\cdot\vec{e} = \sum_i e_i = 0 \qquad (13.17a)$$

$$\vec{x}\cdot\vec{e} = \sum_i(x_ie_i) = 0 \qquad (13.17b)$$

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 the minimum at the point of best-fit. Substituting this two equations readily implies that the concerned quantity $\sum_i((y_i-\hat{y}_i)(\hat{y}_i-\overline{y})) = 0$ 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 the sum of squares error (SSE) plus the sum of squares regression (SSR), i.e.

$$\sum_i(y_i-\overline{y})^2 = \sum_i(y_i-\hat{y}_i)^2 + \sum_i(\hat{y}_i-\overline{y})^2 \qquad (13.18a)$$

$$\text{SST} = \text{SSE} + \text{SSR} \tag{13.18b}$$

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*[5], 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.

$$\begin{aligned}
\overline{y} &= \frac{1}{5}(3.56 + 3.71 + 3.73 + 3.34 + 3.88) \\
&= 3.644 \\
\text{SST} &= \sum_i (y_i - \overline{y})^2 \\
&= (3.56 - 3.644)^2 + (3.71 - 3.644)^2 + (3.73 - 3.644)^2 \\
&\quad + (3.34 - 3.644)^2 + (3.88 - 3.644)^2 \\
&\approx 0.167
\end{aligned}$$

[5]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.

$$
\begin{aligned}
\text{SSR} &= \sum_i (\hat{y}_i - \overline{y})^2 \\
&= ((1.4463 + 0.0162(104) + 0.0710(5.2)) - 3.644)^2 \\
&\quad + ((1.4463 + 0.0162(111) + 0.0710(6.1)) - 3.644)^2 \\
&\quad + ((1.4463 + 0.0162(107) + 0.0710(8.3)) - 3.644)^2 \\
&\quad + ((1.4463 + 0.0162(106) + 0.0710(3.4)) - 3.644)^2 \\
&\quad + ((1.4463 + 0.0162(109) + 0.0710(9.6)) - 3.644)^2 \\
&= (3.500 - 3.644)^2 + (3.678 - 3.644)^2 + (3.769 - 3.644)^2 \\
&\quad + (3.405 - 3.644)^2 + (3.894 - 3.644)^2 \\
&\approx 0.157
\end{aligned}
$$

Hence $R^2 \approx 0.157/0.167 = 0.94$ which indicates a good fit. □

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 3.3.4, we have found that the corresponding 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

$$A = \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \qquad \text{and} \qquad \vec{h} = \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix}$$

So

$$\vec{x}_f = \left(\begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix}^T \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix} \right)^{-1} \begin{bmatrix} 10 & 20 \\ 25 & 15 \\ -10 & 5 \end{bmatrix}^T \begin{bmatrix} 0.2 \\ 0.3 \\ -0.2 \end{bmatrix}$$
$$\approx \begin{bmatrix} 0.01357 \\ 0.00058 \end{bmatrix}$$

So the approximated solution of the temperature gradients that minimizes the squared errors is $\partial T/\partial x = 1.36 \times 10^{-2}$ °C/km and $\partial T/\partial y = 0.58 \times 10^{-3}$ °C/km.

□

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 more round-off errors.

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

$$\vec{x}_f = (A^\dagger A)^{-1} A^\dagger \vec{h}$$

will be reduced to simply

$$\vec{x}_f = A^{-1}\vec{h}$$

the exact solution as derived in Section 3.2.2.

```
                            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
==============================================================================
```

Figure 13.3.: *The output of the Python program that calculates the OLS for Example 13.2.2.*

Exercise 13.2 Find the least-square solution to the overdetermined linear system below.

$$\begin{cases} x + 2y - 3z = 4 \\ x - y + 3z \quad = 5 \\ 2x + y - z \quad = 1 \\ x + y + z \qquad = 2 \end{cases}$$

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 the global carbon dioxide level. Also, calculate the root-mean-square error for each fit.

Figure 13.4.: *The trend of atmospheric carbon dioxide level measured at Mauna Loa Observatory for Example 13.4.*

Years passed since 1960	0	5	10	15	20	25
CO_2 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_2 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))

Exercise 13.5 Radioactive decay is modeled 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 expenses, 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 point of view. In the last chapter, we have seen how to fit a polynomial curve to some given data. It is then natural to further ask, whether there are other suitable types of functions that can be used for data fitting as well. Recall in Chapter 12, we have derived the Fourier series into which any reasonable function can be expanded using sines and cosines. Coincidentally, in the area of Earth Science, many phenomena can be described by the notion of *waves*, which are often assumed to be *sinusoidal* during their derivation, e.g. atmospheric gravity waves, and seismic waves. However, in real life, our data sampling 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 (DFT)*.

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 reasonable function $f(x)$ in the interval $[-\pi, \pi]$ can be written as

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{n=1}^{\infty} b_n \sin(nx) \tag{14.1}$$

where the Fourier coefficients a_m and b_n are given by Formulae (12.17a) and (12.17b). By a change of variable

$$x = \frac{2\pi}{N} t - \pi \tag{14.2}$$

the interval is scaled to $t \in [0, N]$ where the cosines and sines are now in the form of

$$\cos\left(m\left(\frac{2\pi}{N} t - \pi\right)\right) = \cos\left(\frac{2m\pi}{N} t\right) \tag{14.3a}$$

$$\sin\left(n\left(\frac{2\pi}{N} t - \pi\right)\right) = \sin\left(\frac{2n\pi}{N} t\right) \tag{14.3b}$$

The negative sign arising from the $m\pi$ or $n\pi$ term inside is absorbed whenever needed. Since it involves a linear variable transformation only, these cosines and sines

$$\begin{aligned}\Bigg\{&\sqrt{\frac{2}{N}} \sin\left(\frac{2\pi}{N} t\right), \sqrt{\frac{2}{N}} \sin\left(\frac{2\pi(2)}{N} t\right), \sqrt{\frac{2}{N}} \sin\left(\frac{2\pi(3)}{N} t\right), \ldots, \frac{1}{\sqrt{N}}, \\ &\sqrt{\frac{2}{N}} \cos\left(\frac{2\pi}{N} t\right), \sqrt{\frac{2}{N}} \cos\left(\frac{2\pi(2)}{N} t\right), \sqrt{\frac{2}{N}} \cos\left(\frac{2\pi(3)}{N} t\right), \ldots\Bigg\}\end{aligned} \tag{14.4}$$

[1] are still an orthonormal basis to the new $L^2[0, N]$ space, mapped from the initial $L^2[-\pi, \pi]$[2]. The Fourier coefficients are now computed by

$$a_m = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos\left(\frac{2m\pi}{N}t\right)d(\frac{2\pi}{N}t) = \frac{2}{N}\int_0^N f(t)\cos\left(\frac{2m\pi}{N}t\right)dt \tag{14.5a}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin\left(\frac{2n\pi}{N}t\right)d(\frac{2\pi}{N}t) = \frac{2}{N}\int_0^N f(t)\sin\left(\frac{2n\pi}{N}t\right)dt \tag{14.5b}$$

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

$$S_p[f](t) = \frac{a_0}{2} + \sum_{m=1}^{p} a_m\cos\left(\frac{2m\pi}{N}t\right) + \sum_{n=1}^{p} b_n\sin\left(\frac{2n\pi}{N}t\right) \tag{14.6a}$$

$$= \frac{a_0}{2} + a_1\cos\left(\frac{2\pi}{N}t\right) + a_2\cos\left(\frac{2\pi(2)}{N}t\right) + \cdots + a_p\cos\left(\frac{2\pi p}{N}t\right)$$
$$+ b_1\sin\left(\frac{2\pi}{N}t\right) + b_2\sin\left(\frac{2\pi(2)}{N}t\right) + \cdots + b_p\sin\left(\frac{2\pi p}{N}t\right) \tag{14.6b}$$

---

[1] The normalization factor of $\sqrt{2/N}$ can be deduced by calculating the squared norm of just the cosines themselves

$$\int_0^N \left|\cos\left(\frac{2m\pi}{N}t\right)\right|^2 dt = \int_0^N \frac{1}{2}\left(1 + \cos\left(\frac{4m\pi}{N}t\right)\right)dt$$
$$= \frac{1}{2}(N) + (0) = \frac{N}{2}$$

and the same goes for the sines.

[2] 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 for $L^2[-\pi, \pi]$ as given, $c_1\varphi_j(x) + c_2\varphi_j(x) + \cdots = 0(x) \equiv 0$ where $0(x)$ is the zero function over $x \in [-\pi, \pi]$, has the trivial solution $c_j = 0$ as its only solution. Assume the form of the linear variable change is $x = \alpha t + \beta = X(t)$. Replacing x by $X(t)$ (possible since $\alpha \neq 0$) then immediately produces the equality $c_1\varphi_j(X(t)) + c_2\varphi_j(X(t)) + \cdots = 0(X(t)) \equiv 0$ which automatically inherits the desired property of possessing only the trivial solution $c_j = 0$ for the zero function $0(X(t)) = 0(t)$ as well and shows that the new basis functions are now $\hat{\varphi}_j(t) = \varphi_j(X(t))$.

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 in Equation (12.4), a.k.a. the best-fit trigonometric (Fourier) polynomial of degree p. This is known as the ***Best Approximation Theorem*** which directly follows from Properties 10.3.3 and the related discussion about orthogonal projections onto the Fourier basis in Section 12.3.3: Such a Fourier partial sum $S_p[f](t)$ is essentially the orthogonal projection of $f(t)$ onto the subspace spanned by the corresponding orthonormal trigonometric basis up to order p. Hence, the higher the degree of the partial sum of the Fourier expansion, the more closer the approximation to the given function. We will use Example 12.3.2 as an illustration: Note that 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 (Figure 14.1) then shows the original function and the partial sums of its Fourier expansion up to degrees $3, 10, 20$ respectively. From this, we can see that as the degree p goes up, $S_p[f](t)$ becomes a better approximation to the straight line $y = \frac{2\pi}{N}t - \pi$.[3] Eventually, as $p \longrightarrow \infty$ and the expression tends to the full Fourier series, $S_p[f]$ will converge to f in the L^2 sense[4], with respect to the inner product in Equation (12.4). To

[3]Notice that at the two 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 behaves as the periodic extension of the original function (see Footnote 20 of Chapter 12) and will converge to the average value of the left and right limits $(f_-(t) + f_+(t))/2$ at point of discontinuity. Therefore, as every end-point is essentially a discontinuity where one of the side limits is π while the other is $-\pi$, it converges to $((-\pi) + \pi)/2 = 0$.

[4]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.

Figure 14.1.: *The Fourier expansion of the linear function up to degrees* 3, 10, *and* 20.

show this heuristically, we need the ***Bessel's Inequality*** (and *Parseval's Identity*) for partial sums regarding a complete orthonormal basis in any L^2 space, which includes the case of the Fourier series:

$$\lim_{p\to\infty} \left\| f - S_p[f] \right\|^2 = 0 \tag{14.7}$$

The detailed derivation is placed in Appendix A.3.

Returning to practices in Earth Science, and other fields like Engineering, we are often 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 apply the idea of Fourier 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 $t = 0, 1, 2, \ldots, N-1$. Further assume that for now, we only use a

few sines and cosines for the approximation, such that the degree p is much 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 (14.6a) where N is the period of the time series. Since we only have finite data points, we cannot carry out the integrations needed to compute (14.5a) and (14.5b). Nevertheless, we can borrow the idea in the last chapter and do the approximation in a way such that the Fourier partial sum will achieve the least-square error when summed over the sampling points. Subsequently, the system to be optimized will be

$$\begin{cases} C_0 + A_1 \cos\left(\frac{2\pi}{N}(0)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(0)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(0)\right) \\ \qquad + B_1 \sin\left(\frac{2\pi}{N}(0)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(0)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(0)\right) = f(0) \\ C_0 + A_1 \cos\left(\frac{2\pi}{N}(1)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(1)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(1)\right) \\ \qquad + B_1 \sin\left(\frac{2\pi}{N}(1)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(1)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(1)\right) = f(1) \\ C_0 + A_1 \cos\left(\frac{2\pi}{N}(2)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(2)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(2)\right) \\ \qquad + B_1 \sin\left(\frac{2\pi}{N}(2)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(2)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(2)\right) = f(2) \\ \qquad\qquad \vdots \\ C_0 + A_1 \cos\left(\frac{2\pi}{N}(N-1)\right) + A_2 \cos\left(\frac{2\pi(2)}{N}(N-1)\right) + \cdots + A_p \cos\left(\frac{2\pi p}{N}(N-1)\right) \\ \qquad + B_1 \sin\left(\frac{2\pi}{N}(N-1)\right) + B_2 \sin\left(\frac{2\pi(2)}{N}(N-1)\right) + \cdots + B_p \sin\left(\frac{2\pi p}{N}(N-1)\right) = f(N-1) \end{cases} \tag{14.8}$$

where we have replaced the small letters a_k, b_k by capital letters A_k, B_k and a_0 by C_0. It can be expressed in the form of a matrix system

$$G\vec{\beta} = \vec{d} \tag{14.9}$$

where

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & \cdots & 1 & 0 \\ 1 & \cos\left(\frac{2\pi}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \cos\left(\frac{2\pi(2)}{N}\right) & \cdots & \cos\left(\frac{2\pi p}{N}\right) & \sin\left(\frac{2\pi p}{N}\right) \\ 1 & \cos\left(\frac{2\pi(2)}{N}\right) & \sin\left(\frac{2\pi(2)}{N}\right) & \cos\left(\frac{2\pi(2)(2)}{N}\right) & \cdots & \cos\left(\frac{2\pi p(2)}{N}\right) & \sin\left(\frac{2\pi p(2)}{N}\right) \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \cos\left(\frac{2\pi(N-1)}{N}\right) & \sin\left(\frac{2\pi(N-1)}{N}\right) & \cos\left(\frac{2\pi(2)(N-1)}{N}\right) & \cdots & \cos\left(\frac{2\pi p(N-1)}{N}\right) & \sin\left(\frac{2\pi p(N-1)}{N}\right) \end{bmatrix} \tag{14.10}$$

is a $N \times (2p+1)$ matrix and

$$\vec{\beta} = \begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \\ A_p \\ B_p \end{bmatrix} \qquad \vec{d} = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \\ f(N-1) \end{bmatrix} \tag{14.11}$$

are vectors with $2p+1$ and N components respectively. The least-square method then sets out to find the best-fit parameters for $\vec{\beta} = (C_0, A_1, B_1, \cdots, A_p, B_p)^T$ to this system of equations. From Theorem 13.1.3, we know that the best parameters are found by

$$\vec{\beta}_f = (G^T G)^{-1} G^T \vec{d} \tag{14.12}$$

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 in a series as a complex exponential using Euler's formula (8.6), then the two column vectors of a sine-cosine pair in G with a particular frequency of $2\pi k/N$, where both k and N are integers, can be expressed by the real and imaginary parts of

$$\left\{\exp\left(i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} = \left\{\cos\left(\frac{2\pi k}{N}t\right) + i\sin\left(\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \tag{14.13a}$$

or alternatively,

$$\left\{\exp\left(-i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} = \left\{\cos\left(\frac{2\pi k}{N}t\right) - i\sin\left(\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \tag{14.13b}$$

We now first prove that the two column vectors of a sine-cosine pair that have the same frequency are orthogonal. Taking the sum of squares over (14.13a)

gives

$$\sum_{t=0}^{N-1}\left(\exp\left(i\frac{2\pi k}{N}t\right)\right)^2 = \sum_{t=0}^{N-1}\left(\cos\left(\frac{2\pi k}{N}t\right) + i\sin\left(\frac{2\pi k}{N}t\right)\right)^2$$
$$\sum_{t=0}^{N-1}\exp\left(i\frac{4\pi k}{N}t\right) = \sum_{t=0}^{N-1}\left(\cos^2(\frac{2\pi k}{N}t) + 2i\sin\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi k}{N}t\right) - \sin^2(\frac{2\pi k}{N}t)\right) \tag{14.14}$$

Notice that the L.H.S. is a geometric sequence with a common ratio $r = \exp(i4\pi k/N)$[5], the sum of which is seen to be

$$\frac{1-r^N}{1-r} = \frac{1-\exp(i4\pi k)}{1-\exp(i4\pi k/N)} = 0 \tag{14.15}$$

as $\exp(i4\pi k)$ is just 1. By comparing the real and imaginary parts, we know that

$$\sum_{t=0}^{N-1}\cos^2(\frac{2\pi k}{N}t) = \sum_{t=0}^{N-1}\sin^2(\frac{2\pi k}{N}t) \tag{14.16a}$$

$$\sum_{t=0}^{N-1}\sin\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi k}{N}t\right) = 0 \tag{14.16b}$$

The second equation (14.16b) shows that the two column vectors representing sine and cosine waves of the same frequency have a dot product of zero and hence are orthogonal.

Utilizing the complex formulation, 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 with a frequency of $2\pi k/N$, the second series is also a sine, of a frequency of $2\pi l/N$, and $k \neq l$ are distinct integers. We start by considering the sum of products between

$$\left\{\exp\left(i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} = \left\{\cos\left(\frac{2\pi k}{N}t\right) + i\sin\left(\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \tag{14.17a}$$

[5] which will not be equal to 1 as we have demanded that $p, k \leq \frac{N-1}{2}$.

and

$$\left\{\exp\left(i\frac{2\pi l}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} = \left\{\cos\left(\frac{2\pi l}{N}t\right) + i\sin\left(\frac{2\pi l}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \tag{14.17b}$$

The analysis is similar to the one above. Particularly, if we take the sum of products over these two equations, the L.H.S. will be zero, and by considering the real part of the expression on R.H.S., we have

$$\sum_{t=0}^{N-1}\cos\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi l}{N}t\right) - \sum_{t=0}^{N-1}\sin\left(\frac{2\pi k}{N}t\right)\sin\left(\frac{2\pi l}{N}t\right) = 0 \tag{14.18}$$

as long as $k + l$ is not the integer multiples of N. (why?)[6] We can also consider another sum of products between

$$\left\{\exp\left(i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} = \left\{\cos\left(\frac{2\pi k}{N}t\right) + i\sin\left(\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \tag{14.19a}$$

and

$$\left\{\exp\left(-i\frac{2\pi l}{N}t\right)\right\}_{t=0,1,2,\cdots,N-1} = \left\{\cos\left(\frac{2\pi l}{N}t\right) - i\sin\left(\frac{2\pi l}{N}t\right)\right\}_{t=0,1,2,\cdots,N-1} \tag{14.19b}$$

Again by looking at the real part, this yields another relation as[7]

$$\sum_{t=0}^{N-1}\cos\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi l}{N}t\right) + \sum_{t=0}^{N-1}\sin\left(\frac{2\pi k}{N}t\right)\sin\left(\frac{2\pi l}{N}t\right) = 0 \tag{14.20}$$

---

[6] If $k + l = qN$, the sum of complex exponentials on L.H.S. will then become

$$\begin{aligned}\sum_{t=0}^{N-1}\exp\left(i\frac{2\pi k}{N}t\right)\exp\left(i\frac{2\pi l}{N}t\right) &= \sum_{t=0}^{N-1}\exp\left(i\frac{2\pi k}{N}t\right)\exp\left(i\frac{2\pi(qN-k)}{N}t\right)\\ &= \sum_{t=0}^{N-1}\exp(i2\pi qt) = \sum_{t=0}^{N-1}1 = N \neq 0\end{aligned}$$

and the argument fails.

[7] Similarly we have the constraint that k and l are not differed by an integer multiple of N.

From the two equations (14.18) and (14.20) just derived, we can hence conclude that

$$\sum_{t=0}^{N-1} \cos\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi l}{N}t\right) = 0 \tag{14.21a}$$

$$\sum_{t=0}^{N-1} \sin\left(\frac{2\pi k}{N}t\right)\sin\left(\frac{2\pi l}{N}t\right) = 0 \tag{14.21b}$$

The orthogonality relations can be proven between sine and cosine of different frequencies as well in a very similar essence and we leave it to the readers. We will now establish the last result about 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

$$\left\{\exp\left(i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1} \qquad \text{and} \qquad \left\{\exp\left(-i\frac{2\pi k}{N}t\right)\right\}_{t=0,1,2,\ldots,N-1}$$

This time, the L.H.S. is not a geometric series, but rather N terms of 1. The relation is then

$$\begin{aligned} N = \sum_{t=0}^{N-1}(1) &= \sum_{t=0}^{N-1} \exp\left(i\frac{2\pi k}{N}t\right)\exp\left(-i\frac{2\pi k}{N}t\right) \\ &= \sum_{t=0}^{N-1}\left(\cos\left(\frac{2\pi k}{N}t\right) + i\sin\left(\frac{2\pi k}{N}t\right)\right)\left(\cos\left(\frac{2\pi k}{N}t\right) - i\sin\left(\frac{2\pi k}{N}t\right)\right) \\ &= \sum_{t=0}^{N-1}\cos^2\left(\frac{2\pi k}{N}t\right) + \sum_{t=0}^{N-1}\sin^2\left(\frac{2\pi k}{N}t\right) \end{aligned} \tag{14.22}$$

Actually, it can also be inferred from the fact that the sum of a sine-cosine square pair is 1. Solving the two equations (14.16a) and (14.22) then yields

$$\sum_{t=0}^{N-1}\cos^2\left(\frac{2\pi k}{N}t\right) = \sum_{t=0}^{N-1}\sin^2\left(\frac{2\pi k}{n}t\right) = \frac{N}{2} \tag{14.23}$$

Hence the product $G^T G$, where each entry will be the dot product between the corresponding series of sines and cosines, will be

$$G^T G = \begin{bmatrix} N & 0 & 0 & \cdots \\ 0 & \frac{N}{2} & 0 & \\ 0 & 0 & \frac{N}{2} & \\ \vdots & & & \ddots \end{bmatrix} \tag{14.24}$$

and

$$(G^T G)^{-1} = \frac{1}{N} \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 2 & 0 & \\ 0 & 0 & 2 & \\ \vdots & & & \ddots \end{bmatrix} \tag{14.25}$$

So the best-fit parameters are

$$\vec{\beta}_f = (G^T G)^{-1} G^T \vec{d}$$

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 2 & 0 & \\ 0 & 0 & 2 & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & \cos\left(\frac{2\pi}{N}\right) & \cos\left(\frac{2\pi(2)}{N}\right) & \\ 0 & \sin\left(\frac{2\pi}{N}\right) & \sin\left(\frac{2\pi(2)}{N}\right) & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \\ \vdots \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 2 & 2\cos\left(\frac{2\pi}{N}\right) & 2\cos\left(\frac{2\pi(2)}{N}\right) & \\ 0 & 2\sin\left(\frac{2\pi}{N}\right) & 2\sin\left(\frac{2\pi(2)}{N}\right) & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} \tag{14.26}$$

Detailed expressions are

$$C_0 = \frac{1}{N}(f(0) + f(1) + f(2) + \cdots + f(N-1))$$

$$= \frac{1}{N}\sum_{t=0}^{n-1} f(t) \tag{14.27a}$$

$$A_k = \frac{2}{N}\left(f(0) + f(1)\cos\left(\frac{2\pi k}{N}\right) + \cdots + f(N-1)\cos\left(\frac{2\pi k(N-1)}{N}\right)\right)$$

$$= \frac{2}{N}\sum_{t=0}^{N-1} f(t)\cos\left(\frac{2\pi kt}{N}\right) \tag{14.27b}$$

$$B_k = \frac{2}{N}\left(f(1)\sin\left(\frac{2\pi k}{N}\right) + \cdots + f(N-1)\sin\left(\frac{2\pi k(N-1)}{N}\right)\right)$$

$$= \frac{2}{N}\sum_{t=0}^{N-1} f(t)\sin\left(\frac{2\pi kt}{N}\right) \tag{14.27c}$$

That makes sense, at least they seem to be. But in the worked short exercise that follows the example below, you will immediately notice that there is a big caveat to this prototype.

Example 14.1.1. Fit the following time series with the Fourier trigonometric 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 it means that there are only three components, which are the constant term, and a sine-cosine pair with a frequency of $2\pi k/N$ where $k = 1$, $N = 5$. The best-fit parameters will be

$$\begin{bmatrix} C_0 \\ A_1 \\ B_1 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2\cos\left(\frac{2\pi}{5}\right) & 2\cos\left(\frac{2\pi(2)}{5}\right) & 2\cos\left(\frac{2\pi(3)}{5}\right) & 2\cos\left(\frac{2\pi(4)}{5}\right) \\ 0 & 2\sin\left(\frac{2\pi}{5}\right) & 2\sin\left(\frac{2\pi(2)}{5}\right) & 2\sin\left(\frac{2\pi(3)}{5}\right) & 2\sin\left(\frac{2\pi(4)}{5}\right) \end{bmatrix}\begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \\ f(4) \end{bmatrix}$$

$$C_0 = \frac{1}{5}(4 + 1 + 2 + 3 + 1) = \frac{11}{5}$$

$$
\begin{aligned}
A_1 &= \frac{2}{5}\left(4 + 1\cos\left(\frac{2\pi}{5}\right) + 2\cos\left(\frac{2\pi(2)}{5}\right) + 3\cos\left(\frac{2\pi(3)}{5}\right) + 1\cos\left(\frac{2\pi(4)}{5}\right)\right) \\
&\approx 0.229 \\
B_1 &= \frac{2}{5}\left(0 + 1\sin\left(\frac{2\pi}{5}\right) + 2\sin\left(\frac{2\pi(2)}{5}\right) + 3\sin\left(\frac{2\pi(3)}{5}\right) + 1\sin\left(\frac{2\pi(4)}{5}\right)\right) \\
&\approx -0.235
\end{aligned}
$$

So the best trigonometric fit of order $p = 1$ for the time series concerned is

$$
f(t) \approx \frac{11}{5} + 0.229\cos\left(\frac{2\pi}{5}t\right) - 0.235\sin\left(\frac{2\pi}{5}t\right)
$$

□

Example 14.1.2. As a follow-up short exercise for the last example, find an improved approximation with order $p = 2$. What happens if $p = 3$?

Solution. 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

$$
\begin{aligned}
A_2 &= \frac{2}{5}\left(4 + 1\cos\left(\frac{2\pi(2)}{5}\right) + 2\cos\left(\frac{2\pi(2)(2)}{5}\right)\right. \\
&\qquad \left. + 3\cos\left(\frac{2\pi(2)(3)}{5}\right) + 1\cos\left(\frac{2\pi(2)(4)}{5}\right)\right) \approx 1.571 \\
B_2 &= \frac{2}{5}\left(0 + 1\sin\left(\frac{2\pi(2)}{5}\right) + 2\sin\left(\frac{2\pi(2)(2)}{5}\right)\right.
\end{aligned}
$$

$$+ 3\sin\left(\frac{2\pi(2)(3)}{5}\right) + 1\sin\left(\frac{2\pi(2)(4)}{5}\right)\Bigg) \approx 0.380$$

Hence the new approximation will be

$$\begin{aligned} f(t) \approx \frac{11}{5} &+ 0.229\cos\left(\frac{2\pi}{5}t\right) - 0.235\sin\left(\frac{2\pi}{5}t\right) \\ &+ 1.571\cos\left(\frac{2\pi(2)}{5}t\right) + 0.380\sin\left(\frac{2\pi(2)}{5}t\right) \end{aligned}$$

When p is increased to 3, the matrix G involved in the best-fit formula becomes

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & \cos\left(\frac{2\pi}{5}\right) & \sin\left(\frac{2\pi}{5}\right) & \cos\left(\frac{2\pi(2)}{5}\right) & \sin\left(\frac{2\pi(2)}{5}\right) & \cos\left(\frac{2\pi(3)}{5}\right) & \sin\left(\frac{2\pi(3)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(2)}{5}\right) & \sin\left(\frac{2\pi(2)}{5}\right) & \cos\left(\frac{2\pi(2)(2)}{5}\right) & \sin\left(\frac{2\pi(2)(2)}{5}\right) & \cos\left(\frac{2\pi(3)(2)}{5}\right) & \sin\left(\frac{2\pi(3)(2)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(3)}{5}\right) & \sin\left(\frac{2\pi(3)}{5}\right) & \cos\left(\frac{2\pi(2)(3)}{5}\right) & \sin\left(\frac{2\pi(2)(3)}{5}\right) & \cos\left(\frac{2\pi(3)(3)}{5}\right) & \sin\left(\frac{2\pi(3)(3)}{5}\right) \\ 1 & \cos\left(\frac{2\pi(4)}{5}\right) & \sin\left(\frac{2\pi(4)}{5}\right) & \cos\left(\frac{2\pi(2)(4)}{5}\right) & \sin\left(\frac{2\pi(2)(4)}{5}\right) & \cos\left(\frac{2\pi(3)(4)}{5}\right) & \sin\left(\frac{2\pi(3)(4)}{5}\right) \end{bmatrix}$$

and a routine computation will show that

$$G^T G = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{5}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{5}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{5}{2} & 0 & \frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{2} & 0 & -\frac{5}{2} \\ 0 & 0 & 0 & \frac{5}{2} & 0 & \frac{5}{2} & 0 \\ 0 & 0 & 0 & 0 & -\frac{5}{2} & 0 & \frac{5}{2} \end{bmatrix}$$

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 (14.27b) and (14.27c) to compute A_3 and B_3 we will get

$$\begin{aligned} A_3 = \frac{2}{5}\Bigg(4 &+ 1\cos\left(\frac{2\pi(3)}{5}\right) + 2\cos\left(\frac{2\pi(3)(2)}{5}\right) \\ &+ 3\cos\left(\frac{2\pi(3)(3)}{5}\right) + 1\cos\left(\frac{2\pi(3)(4)}{5}\right)\Bigg) = A_2 \end{aligned}$$

$$B_3 = \frac{2}{5}\left(0 + 1\sin\left(\frac{2\pi(3)}{5}\right) + 2\sin\left(\frac{2\pi(3)(2)}{5}\right)\right.$$
$$\left. + 3\sin\left(\frac{2\pi(3)(3)}{5}\right) + 1\sin\left(\frac{2\pi(3)(4)}{5}\right)\right) = -B_2$$

So A_2 $[B_2]$ and A_3 $[B_3]$ carry the same information and one of them will be redundant. □

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 short exercise of 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 $t = 0, 1, \ldots, N - 1$

$$\begin{aligned}\cos\left(\frac{2\pi l}{N}t\right) &= \cos\left(\frac{2\pi(N-k)}{N}t\right)\\ &= \cos\left(2\pi t - \frac{2\pi k}{N}t\right)\\ &= \cos\left(\frac{2\pi k}{N}t\right)\end{aligned} \tag{14.28}$$

so the two cosine waves, despite having different frequencies, coincide at every data point and will be indistinguishable within the time series. The same problem arises similarly for the sine signals. The condition $k + l = N$ above offers a hint that the maximum value of p should be $N/2$. This corresponds to an angular frequency of

$$\omega_{\mathrm{Ny}} = \frac{2\pi}{N}\left(\frac{N}{2}\right) = \pi \tag{14.29}$$

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 order $\lfloor N/2 \rfloor$[8]. We have just shown a part of the theorem that if $k + l = N$, then the sinusoidal series of a higher order $l = \lfloor N/2 \rfloor + 1, \ldots, N$ will be redundant in the derivation (14.28) above. We will complete the theorem by verifying that the sine/cosine terms of an order $l = N + 1, N + 2, \ldots$ beyond will also lead to duplicated modes: let $l = k + qN$ where q is a positive integer, then

$$\begin{aligned}
\cos\left(\frac{2\pi l}{N}t\right) &= \cos\left(\frac{2\pi(k+qN)}{N}t\right) \\
&= \cos\left(\frac{2\pi k}{N}t + 2\pi q t\right) \\
&= \cos\left(\frac{2\pi k}{N}t\right)
\end{aligned} \tag{14.30}$$

for all time steps $t = 0, 1, \ldots, N - 1$, and the same goes 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 Fourier coefficient of a_l in the "ground truth"),

[8]The floor symbol $\lfloor\ \rfloor$ returns the largest integer smaller than or equal to the argument.

then (14.27b) will yield

$$\begin{aligned}
A_{k\leftarrow l} &= \frac{2}{N}\sum_{t=0}^{N-1}\left(a_l\cos\left(\frac{2\pi lt}{N}\right)\right)\cos\left(\frac{2\pi kt}{N}\right)\\
&= \frac{2}{N}\sum_{t=0}^{N-1}a_l\cos\left(\frac{2\pi(N-k)t}{N}\right)\cos\left(\frac{2\pi kt}{N}\right)\\
&= \frac{2}{N}\sum_{t=0}^{N-1}a_l\cos\left(2\pi t-\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi kt}{N}\right)\\
&= \frac{2}{N}\sum_{t=0}^{N-1}a_l\cos\left(\frac{2\pi k}{N}t\right)\cos\left(\frac{2\pi kt}{N}\right)\\
&= \frac{2}{N}\sum_{t=0}^{N-1}a_l\cos^2\left(\frac{2\pi kt}{N}\right)\\
&= \frac{2}{N}\left(a_l\left(\frac{N}{2}\right)\right) = a_l \qquad \text{(by (14.23))} \qquad (14.31)
\end{aligned}$$

Hence any signal with a frequency higher than the Nyquist frequency will contribute to the respective DFT frequency and contaminate it.[9]

Now we can formally derive the ***(real) Discrete Fourier Transform*** for a time series which is given by (14.27a), (14.27b), and (14.27c). 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, we will have frequencies from zero to $k = \frac{N}{2}$. Notice that the constant term of zero frequency contributes a single parameter, every other frequency contributes two coefficients via a pair of sine and cosine, and the maximum frequency $k_{\text{Ny}} = \frac{N}{2}$ only gives rise to one coefficient from the cosine series which takes the form of alternating $(1, -1, 1, -1, \ldots, 1, -1)$ (the corresponding diagonal entry in the $(G^TG)^{-1}$

[9] On the other hand, if the true function only contains the lower frequency mode, then by the same steps we can show that the prototype DFT coefficient A_k will coincide exactly with the usual Fourier coefficients a_k.

term of the least-square formula will hence be $1/N$ instead of $2/N$), as the sine function of the Nyquist frequency will be always zero at the sampling points. In both cases, there can be at most N sinusoidal curves for fitting N data points and G will be square (and hence invertible). Since the amount of parameters is the same as the number of data, it becomes a unique 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 14.1.3. 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

$$\begin{aligned} &(C_0, A_1, B_1, A_2, B_2, A_3, B_3, A_4)^T \\ &= (6, -2.586, 2.243, 2, 10, -5.414, 6.243, 8)^T \end{aligned}$$

□

14.1.3. Complex DFT

The real DFT approach above has the drawback of using sine-cosine pairs for computation. In Section 8.1, we have learned the power of complex exponentials

to simultaneously represent sines and cosines through Euler's formula, which has also been exploited to derive the relations between the various sinusoidal functions when we are developing the DFT prototype. Therefore, there is an 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 subsection, we propose an interpolation scheme that 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

$$-\frac{N}{2}, -\left(\frac{N}{2}-1\right), \ldots, -1, 0, 1, \ldots, \frac{N}{2}-1$$

for even N and

$$-\frac{N-1}{2}, -\left(\frac{N-1}{2}-1\right), \ldots, -1, 0, 1, \ldots, \frac{N-1}{2}$$

for odd N. Both ranges ensure that the total number of complex exponentials used in the fitting process is exactly N. Each pair of complex exponentials with the wavenumber 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 (8.7), 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

$$G = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \exp\left(i\frac{2\pi}{N}\right) & \exp\left(i\frac{2\pi(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})}{N}\right) & \cdots & \exp\left(-i\frac{2\pi}{N}\right) \\ 1 & \exp\left(i\frac{2\pi(2)}{N}\right) & \exp\left(i\frac{2\pi(2)(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)(2)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})(2)}{N}\right) & \cdots & \exp\left(-i\frac{2\pi(2)}{N}\right) \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \exp\left(i\frac{2\pi(N-1)}{N}\right) & \exp\left(i\frac{2\pi(2)(N-1)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)(N-1)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})(N-1)}{N}\right) & \cdots & \exp\left(-i\frac{2\pi(N-1)}{N}\right) \end{bmatrix} \tag{14.32}$$

for even N, where each column represents frequencies corresponding to

$$k = 0, 1, \ldots, \frac{N}{2} - 1, -\frac{N}{2}, \ldots, -1$$

This is a common convention where we start from $k = 0$, and it increases to the largest positive $k = \frac{N}{2} - 1$, then we flip the sign and resume from the most negative $k = -\frac{N}{2}$, finally go 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 is from $k = \frac{N-1}{2}$ to $k = -\frac{N-1}{2}$.

The entries of G^*G in Formula (13.4) from Theorem 13.1.3 are then the complex dot products between the sequences 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.[10] Thus, the expression G^*G is just N times the identity I, and $(G^*G)^{-1} = \frac{1}{N}I$.

Now we denote the new, complex best-fit parameters, or coefficients by C_k. Subsequently,

$$\vec{\beta}_f = (G^*G)^{-1}G^*\vec{d} = \left(\frac{1}{N}I\right)G^*\vec{d}$$

$$\begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_{\frac{N}{2}-1} \\ C_{-\frac{N}{2}} \\ \vdots \\ C_{-1} \end{bmatrix} = \frac{1}{N}\begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & \exp\left(-i\frac{2\pi}{N}\right) & \exp\left(-i\frac{2\pi(2)}{N}\right) & \cdots \\ \vdots & \vdots & \vdots & \\ 1 & \exp\left(-i\frac{2\pi(\frac{N}{2}-1)}{N}\right) & \exp\left(-i\frac{2\pi(\frac{N}{2}-1)(2)}{N}\right) & \cdots \\ 1 & \exp\left(-i\frac{2\pi(-\frac{N}{2})}{N}\right) & \exp\left(-i\frac{2\pi(-\frac{N}{2})(2)}{N}\right) & \cdots \\ \vdots & \vdots & \vdots & \\ 1 & \exp\left(-i\frac{2\pi(-1)}{N}\right) & \exp\left(-i\frac{2\pi(-1)(2)}{N}\right) & \cdots \end{bmatrix}\begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} \tag{14.33}$$

Again this is for even N, and we ought to replace the indices for coefficients and complex exponentials appropriately for odd N. The $\frac{1}{N}$ factor will also be ignored. We now conclude the method of ***(complex) Discrete Fourier Transform*** in a compact way as follows.

Definition 14.1.2 (Discrete Fourier Transform (DFT))**.** The coefficients, or amplitudes, of the DFT in complex form are computed by

$$C_k = \sum_{t=0}^{N-1} f(t) \exp\left(-i\frac{2\pi k}{N}t\right) \tag{14.34}$$

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 Δt different from 1 ($t = 0, \Delta t, 2\Delta t, \ldots, (N-1)\Delta t$), the appropriate expression is

$$C_k = \sum_{s=0}^{N-1} f(s\Delta t) \exp\left(-i\frac{2\pi k}{N}s\right) \tag{14.35}$$

[10]When $k \neq l$, we have

$$\begin{aligned}
\sum_{t=0}^{N-1} \exp(i(2\pi k/N)t)\overline{\exp(i(2\pi l/N)t)} &= \sum_{t=0}^{N-1} \exp(i(2\pi k/N)t)\exp(-i(2\pi l/N)t) \\
&= \sum_{t=0}^{N-1} \exp(i(2\pi(k-l)/N)t) \\
&= \frac{1-\exp(i(2\pi(k-l)/N))^N}{1-\exp(i(2\pi(k-l)/N))} \\
&= \frac{1-\exp(i(2\pi(k-l)))}{1-\exp(i(2\pi(k-l)/N))} \\
&= \frac{1-1}{1-\exp(i(2\pi(k-l)/N))} = 0
\end{aligned}$$

The $\exp(i(2\pi(k-l)/N))$ term in the denominator will not equal to 1 and the denominator will not be zero as long as $k \neq l$ and the indices are in the specified range. 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.

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 (8.7) of Properties 8.1.7 and comparing the expression for the real case[11], that is

$$A_k = \frac{C_k + C_{-k}}{N} \tag{14.36a}$$

$$B_k = -\frac{(C_k - C_{-k})}{iN} \tag{14.36b}$$

for $k \neq 0$. If N is even, then we define $C_{N/2} = 0$ for convenience. Moreover, since sine is an odd function, and cosine is an even function, $\mathrm{Re}(C_k) = \mathrm{Re}(C_{-k})$, and $\mathrm{Im}(C_k) = -\mathrm{Im}(C_{-k})$ 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

Properties 14.1.3. Given a real time series, the amplitudes A_k, B_k in real DFT (14.27b) and (14.27c), and C_k in complex DFT (14.34) satisfy the relations

$$A_k = 2\,\mathrm{Re}(C_k)/N \tag{14.37a}$$

$$B_k = -2\,\mathrm{Im}(C_k)/N \tag{14.37b}$$

11

$$\begin{aligned}
\frac{C_k + C_{-k}}{N} &= \frac{1}{N}\left(\sum_{t=0}^{N-1} f(t)\exp\left(-i\frac{2\pi k}{N}t\right) + \sum_{t=0}^{N-1} f(t)\exp\left(-i\frac{2\pi(-k)}{N}t\right)\right) \\
&= \frac{1}{N}\left(\sum_{t=0}^{N-1} f(t)\left(\exp\left(-i\frac{2\pi k}{N}t\right) + \exp\left(i\frac{2\pi k}{N}t\right)\right)\right) \\
&= \frac{2}{N}\sum_{t=0}^{N-1} f(t)\cos\left(\frac{2\pi k}{N}t\right) \qquad \text{(by (8.7a))}
\end{aligned}$$

which is just equal to A_k in Equation (14.27b). The derivation is similar for B_k.

for any $0 \neq |k| < \frac{N}{2}$. Meanwhile, C_0 is the same in both types of DFT and when N is even, $A_{N/2} = \frac{1}{N}C_{-N/2}$. The $\frac{1}{N}$ factor is optional and depends on if any convention is used.

Example 14.1.4. Find the complex DFT for the time series in example 14.1.1.

Solution. Using Formula (14.34) in Definition 14.1.2, we have

$$\begin{aligned}
C_0 &= 4 + 1 + 2 + 3 + 1 = 11 \\
C_1 &= 4 + 1\exp\left(-i\frac{2\pi}{5}\right) + 2\exp\left(-i\frac{2\pi(2)}{5}\right) \\
&\quad + 3\exp\left(-i\frac{2\pi(3)}{5}\right) + 1\exp\left(-i\frac{2\pi(4)}{5}\right) = 0.573 + 0.588i \\
C_2 &= 4 + 1\exp\left(-i\frac{2\pi(2)}{5}\right) + 2\exp\left(-i\frac{2\pi(2)(2)}{5}\right) \\
&\quad + 3\exp\left(-i\frac{2\pi(2)(3)}{5}\right) + 1\exp\left(-i\frac{2\pi(2)(4)}{5}\right) = 3.927 - 0.951i
\end{aligned}$$

Either by the property just mentioned or a direct computation, $C_{-1} = \overline{C_1}$ and $C_{-2} = \overline{C_2}$ are seen to be the complex conjugates of C_1 and C_2 respectively. □

► Short Exercise: Check if Properties 14.1.3 is true in this example.[12]

14.1.4. Inverse DFT

During the derivation of complex DFT, we have found that $G^*G = I$ (omitting the N factor), 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 $\hat{\mathbf{f}} = G^*\vec{d}$, we can undo the DFT and

recover the original time series by multiplying to the left with the inverse $(G^*)^{-1}$ so that $\vec{d} = (G^*)^{-1}\hat{\mathbf{f}} = G\hat{\mathbf{f}}$. This means that

$$\begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ \vdots \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \exp\left(i\frac{2\pi}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})}{N}\right) & \cdots & \exp\left(i\frac{2\pi(-1)}{N}\right) \\ 1 & \exp\left(i\frac{2\pi(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(\frac{N}{2}-1)(2)}{N}\right) & \exp\left(i\frac{2\pi(-\frac{N}{2})(2)}{N}\right) & \cdots & \exp\left(i\frac{2\pi(-1)(2)}{N}\right) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \hat{f}_0 \\ \hat{f}_1 \\ \vdots \\ \hat{f}_{\frac{N}{2}-1} \\ \hat{f}_{-\frac{N}{2}} \\ \vdots \\ \hat{f}_{-1} \end{bmatrix} \tag{14.38}$$

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 follows.

Definition 14.1.4 (Inverse DFT)**.** The inverse Discrete Fourier Transform is computed by

$$f(t) = \frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k) \exp\left(i\frac{2\pi k}{N}t\right) \tag{14.39}$$

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 Equation (14.34) from Definition 14.1.2 (with a dummy variable t') into R.H.S., we have

$$\begin{aligned} &\frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} \left(\sum_{t'=0}^{N-1} f(t') \exp\left(-i\frac{2\pi k}{N}t'\right) \right) \exp\left(i\frac{2\pi k}{N}t\right) \\ &= \frac{1}{N} \sum_{t'=0}^{N-1} f(t') \left(\sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} \exp\left(i\frac{2\pi k}{N}(t-t')\right) \right) \end{aligned}$$

[12] We will only check A_1 and B_1 and leave A_2 and B_2 to the readers. $2\,\mathrm{Re}(C_1)/N = 2(0.573)/5 \approx 0.229 = A_1$, $-2\,\mathrm{Im}(C_1)/N = -2(0.588)/5 \approx -0.235 = B_1$.

where

$$\sum_{k=-\lfloor N/2\rfloor}^{\lfloor (N-1)/2\rfloor} \exp\left(i\frac{2\pi k}{N}(t-t')\right) = \begin{cases} N & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \tag{14.40}$$

The first case should be obvious, while the second case is derived in the footnote below.[13] Hence

$$\begin{aligned} &\frac{1}{N}\sum_{t'=0}^{N-1} f(t')\left(\sum_{k=-\lfloor N/2\rfloor}^{\lfloor (N-1)/2\rfloor} \exp\left(i\frac{2\pi k}{N}(t-t')\right)\right) \\ &= \frac{1}{N}((0)+\cdots+f(t)(N)+\cdots+(0)) \\ &= f(t) \end{aligned}$$

so the R.H.S. indeed reproduces the original function on L.H.S.

Example 14.1.5. Apply inverse DFT on the complex DFT coefficients $C_{-3} = 1, C_{-2} = 1+i, C_{-1} = 2-i, C_0 = 3, C_1 = 2+i, C_2 = 1-i$ to retrieve the physical time series.

[13]For an integer $t - t' = \Delta t \neq 0$, we have

$$\begin{aligned} \sum_{k=-\lfloor N/2\rfloor}^{\lfloor (N-1)/2\rfloor} \exp\left(i\frac{2\pi k}{N}\Delta t\right) &= \exp\left(i\frac{2\pi(-\lfloor \frac{N}{2}\rfloor)}{N}(\Delta t)\right)\frac{1-\exp\left(i\frac{2\pi\Delta t}{N}\right)^N}{1-\exp\left(i\frac{2\pi\Delta t}{N}\right)} \\ &= \exp\left(i\frac{-2\pi\lfloor \frac{N}{2}\rfloor}{N}(\Delta t)\right)\frac{1-\exp(i2\pi\Delta t)}{1-\exp\left(i\frac{2\pi\Delta t}{N}\right)} \\ &= \exp\left(i\frac{-2\pi\lfloor \frac{N}{2}\rfloor}{N}(\Delta t)\right)\frac{1-1}{1-\exp\left(i\frac{2\pi\Delta t}{N}\right)} = 0 \end{aligned}$$

as a geometric sum that evaluates to zero. Δt cannot be an integer multiple of N as well, otherwise, the denominator will be problematic, but the range of summation prevents this.

Solution. The time series $f(t)$ will have a period of 6 where $t = 0, 1, 2, 3, 4, 5$. By Formula (14.39) in Definition 14.1.4, we have

$$
\begin{aligned}
f(0) &= \frac{1}{N}\sum_{k=-3}^{2} F(k)\exp\left(i\frac{2\pi k}{N}(0)\right) \\
&= \frac{1}{6}[C_{-3}\exp\left(i\frac{2\pi(-3)}{6}(0)\right) + C_{-2}\exp\left(i\frac{2\pi(-2)}{6}(0)\right) \\
&\quad + C_{-1}\exp\left(i\frac{2\pi(-1)}{6}(0)\right) + C_0\exp\left(i\frac{2\pi(0)}{6}(0)\right) \\
&\quad + C_1\exp\left(i\frac{2\pi(1)}{6}(0)\right) + C_2\exp\left(i\frac{2\pi(2)}{6}(0)\right)] \\
&= \frac{1}{6}[(1) + (1+i) + (2-i) + 3 + (2+i) + (1-i)] = \frac{10}{6} \\
f(1) &= \frac{1}{N}\sum_{k=-3}^{2} F(k)\exp\left(i\frac{2\pi k}{N}(1)\right) \\
&= \frac{1}{6}[C_{-3}\exp\left(i\frac{2\pi(-3)}{6}(1)\right) + C_{-2}\exp\left(i\frac{2\pi(-2)}{6}(1)\right) \\
&\quad + C_{-1}\exp\left(i\frac{2\pi(-1)}{6}(1)\right) + C_0\exp\left(i\frac{2\pi(0)}{6}(1)\right) \\
&\quad + C_1\exp\left(i\frac{2\pi(1)}{6}(1)\right) + C_2\exp\left(i\frac{2\pi(2)}{6}(1)\right)] \\
&= \frac{1}{6}[(1)(-1) + (1+i)e^{-i\frac{2\pi}{3}} + (2-i)e^{-i\frac{\pi}{3}} + 3 + (2+i)e^{i\frac{\pi}{3}} + (1-i)e^{i\frac{2\pi}{3}}] \\
&= \frac{1}{2} \\
f(2) &= \frac{1}{N}\sum_{k=-3}^{2} F(k)\exp\left(i\frac{2\pi k}{N}(2)\right) \\
&= \frac{1}{6}[C_{-3}\exp\left(i\frac{2\pi(-3)}{6}(2)\right) + C_{-2}\exp\left(i\frac{2\pi(-2)}{6}(2)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ C_{-1}\exp\left(i\frac{2\pi(-1)}{6}(2)\right) + C_0\exp\left(i\frac{2\pi(0)}{6}(2)\right) \\
&+ C_1\exp\left(i\frac{2\pi(1)}{6}(2)\right) + C_2\exp\left(i\frac{2\pi(2)}{6}(2)\right)] \\
&= \frac{1}{6}[(1)(1) + (1+i)e^{-i\frac{4\pi}{3}} + (2-i)e^{-i\frac{2\pi}{3}} + 3 + (2+i)e^{i\frac{2\pi}{3}} + (1-i)e^{i\frac{4\pi}{3}}] \\
&\approx -0.4107
\end{aligned}
$$

We leave the calculations for the remaining three data points to the readers. They are $f(3) = 0$, $f(4) \approx 0.7440$, and $f(5) = \frac{1}{2}$. □

14.2. Properties of DFT

14.2.1. Power Spectrum and Parseval's Theorem

The complex DFT coefficients actually store the relevant information about the sinusoidal signals in the corresponding frequencies, namely the *phase* and *amplitude*. To see this, go back to Properties 14.1.3, where

$$A_k = 2\,\mathrm{Re}(C_k)/N \tag{14.41a}$$

$$B_k = -2\,\mathrm{Im}(C_k)/N \tag{14.41b}$$

but also, the wave signal of that particular frequency takes the form of

$$A_k\cos\left(\frac{2\pi k}{N}t\right) + B_k\sin\left(\frac{2\pi k}{N}t\right) \tag{14.42}$$

Plugging (14.41) into (14.42) above gives

$$\frac{2}{N}\left(\mathrm{Re}(C_k)\cos\left(\frac{2\pi k}{N}t\right) - \mathrm{Im}(C_k)\sin\left(\frac{2\pi k}{N}t\right)\right)$$

$$= \frac{2}{N}\left(\hat{A}_k \cos\phi_k \cos\left(\frac{2\pi k}{N}t\right) - \hat{A}_k \sin\phi_k \sin\left(\frac{2\pi k}{N}t\right)\right)$$
$$= \frac{2}{N}\hat{A}_k \cos\left(\frac{2\pi k}{N}t + \phi_k\right) \tag{14.43}$$

where we let $-\pi \leq \phi_k < \pi$ and $\hat{A}_k \geq 0$ as real-valued quantities, in a way such that

$$\mathrm{Re}(C_k) = \hat{A}_k \cos\phi_k \tag{14.44a}$$
$$\text{and} \quad \mathrm{Im}(C_k) = \hat{A}_k \sin\phi_k \tag{14.44b}$$

and then apply the trigonometric identity $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$. ϕ_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 ϕ_k and $\hat{A}_k$ are then derived according to

$$\frac{\mathrm{Im}(C_k)}{\mathrm{Re}(C_k)} = \frac{\sin\phi_k}{\cos\phi_k} = \tan\phi_k \tag{14.45a}$$
$$\mathrm{Re}(C_k)^2 + \mathrm{Im}(C_k)^2 = \hat{A}_k^2\cos^2\phi_k + \hat{A}_k^2\sin^2\phi_k = \hat{A}_k^2 \tag{14.45b}$$

Thus

$$\phi_k = \arctan\left(\frac{\mathrm{Im}(C_k)}{\mathrm{Re}(C_k)}\right) \tag{14.46a}$$
$$\text{and} \quad \hat{A}_k = \sqrt{\mathrm{Re}(C_k)^2 + \mathrm{Im}(C_k)^2} \tag{14.46b}$$

Recalling (8.4) and (8.5), 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

$$|C_k|^2 = C_k\overline{C_k} = C_k C_{-k} \tag{14.47}$$

due to Equation (8.8) 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 14.1.4.

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

$$\phi_1 = \arctan\left(\frac{\mathrm{Im}(C_1)}{\mathrm{Re}(C_1)}\right) = \arctan\left(\frac{0.588}{0.573}\right) = 0.80\,\mathrm{rad}$$

and the power is

$$\begin{aligned}|C_1|^2 = |C_{-1}|^2 = C_1 C_{-1} &= (0.573 + 0.588i)(0.573 - 0.588i)\\ &= 0.67\end{aligned}$$

Similarly, for the second frequency, the phase is

$$\phi_2 = \arctan\left(\frac{\mathrm{Im}(C_2)}{\mathrm{Re}(C_2)}\right) = \arctan\left(\frac{-0.951}{3.927}\right) = -0.24\,\mathrm{rad}$$

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 is related to the initial time series according to the ***Parseval's Theorem***.

Theorem 14.2.1 (Parseval's Theorem)**.** The sum of powers in each DFT coefficient, divided by the length N, is equal to the sum of all data values in the (real) time series squared, i.e.

$$\sum_{t=0}^{N-1} f(t)^2 = \frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} |F(k)|^2 \tag{14.48}$$

Proof. By substituting the form of inverse DFT given as (14.39) in Definition 14.1.4, the L.H.S. becomes

$$\begin{aligned}
&\sum_{t=0}^{N-1} f(t)^2 \\
&= \sum_{t=0}^{N-1} \left[\frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k) \exp\left(i \frac{2\pi k}{N} t \right) \overline{\left(\frac{1}{N} \sum_{l=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(l) \exp\left(i \frac{2\pi l}{N} t \right) \right)} \right] \\
&= \frac{1}{N^2} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k) \left(\sum_{t=0}^{N-1} \exp\left(i \frac{2\pi k}{N} t \right) \left(\sum_{l=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} \overline{F(l)} \exp\left(-i \frac{2\pi l}{N} t \right) \right) \right) \\
&= \frac{1}{N^2} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k) \left(\sum_{l=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} \sum_{t=0}^{N-1} \overline{F(l)} \exp\left(i \frac{2\pi (k-l)}{N} t \right) \right)
\end{aligned}$$

where we have used two different dummy indices k and l initially. Similar to (14.40), we have

$$\sum_{t=0}^{N-1} \exp\left(i \frac{2\pi (k-l)}{N} t \right) = \begin{cases} N & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \tag{14.49}$$

Thus

$$\frac{1}{N^2} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k) \left(\sum_{l=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} \sum_{t=0}^{N-1} \overline{F(l)} \exp\left(i \frac{2\pi (k-l)}{N} t \right) \right)$$

$$
\begin{aligned}
&= \frac{1}{N^2} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} F(k)[(0) + \cdots + (\overline{F(k)})(N) + \cdots + (0)] \\
&= \frac{1}{N^2} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} NF(k)\overline{F(k)} \\
&= \frac{1}{N} \sum_{k=-\lfloor N/2 \rfloor}^{\lfloor (N-1)/2 \rfloor} |F(k)|^2
\end{aligned}
$$

is equal to the R.H.S. □

Example 14.2.2. Verify the Parseval's Theorem for Examples 14.1.1 and 14.1.4.

Solution. The L.H.S. of Equation (14.48) 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 14.1.4, the R.H.S. is

$$
\begin{aligned}
\frac{1}{5} \sum_{k=-2}^{2} |F(k)|^2 &= \frac{1}{5}(|C_{-2}|^2 + |C_{-1}|^2 + |C_0|^2 + |C_1|^2 + |C_2|^2) \\
&= \frac{1}{5}(16.33 + 0.67 + 121 + 0.67 + 16.33) \\
&= \frac{1}{5}(155) = 31
\end{aligned}
$$

which are indeed equal. □

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

$$h(t) = f(t) * g(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(t-\tau) \tag{14.50}$$

if τ 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 14.2. 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

$$\begin{aligned}
h(0) &= f(0)g(0) \\
h(1) &= f(0)g(1-0) + f(1)g(1-1) = f(0)g(1) + f(1)g(0) \\
h(4) &= f(0)g(4-0) + f(1)g(4-1) + f(2)g(4-2) \\
&\quad + f(3)g(4-3) + f(4)g(4-4) \\
&= f(0)g(4) + f(1)g(3) + f(2)g(2) + f(3)g(1) + f(4)g(0)
\end{aligned}$$

and

$$h(6) = f(0)g(6-0) + f(1)g(6-1) + f(2)g(6-2)$$

$$
\begin{aligned}
&+ f(3)g(6-3) + f(4)g(6-4) \\
&= f(0)g(6) + f(1)g(5) + f(2)g(4) + f(3)g(3) + f(4)g(2) \\
h(9) &= f(3)g(6-0) + f(4)g(6-1) = f(3)g(6) + f(4)g(5) \\
h(10) &= f(4)g(6)
\end{aligned}
$$

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 some years t from now on, assuming she will only get pregnant after marriage.

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

$$
p(t = 5) = P(\text{Birth 5 yrs later}|\text{Married now})P(\text{Married now})
$$

Figure 14.2.: *An example schematic of convolution.*

$$
\begin{aligned}
&+ P(\text{Birth 4 yrs later}|\text{Married 1 yr later})P(\text{Married 1 yr later}) \\
&+ \cdots \\
&+ P(\text{Birth 1 yr later}|\text{Married 4 yrs later})P(\text{Married 4 yrs later}) \\
&+ P(\text{Birth now}|\text{Married 5 yrs later})P(\text{Married 5 yrs later}) \\
&= q(0)r(5) + q(1)r(4) + q(2)r(3) \\
&\quad + q(3)r(2) + q(4)r(1) + q(5)r(0) \\
&= (0.2)(0.08) + (0.2)(0.15) + (0.2)(0.15) \\
&\quad + (0.1)(0.15) + (0.1)(0.15) + (0.1)(0) \\
&= 0.106
\end{aligned}
$$

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.[14]

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 14.3).

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

$$h_c(t) = f(t) \circledast g(t) = \sum_{\tau=0}^{N-1} f(\tau)g((t-\tau) \bmod N) \tag{14.51}$$

where $m \bmod N$ is the *modulo* operation that returns an integer $0 \leq n < N$

[14] It is $q(0)r(8) + q(1)r(7) + q(2)r(6) + q(3)r(5) + q(4)r(4) + q(5)r(3) + q(6)r(2) = (0.2)(0.08) + (0.2)(0.08) + (0.2)(0.08) + (0.1)(0.08) + (0.1)(0.15) + (0.1)(0.15) + (0.1)(0.15) = 0.101$.

between 0 and $N-1$ such that $m + qN = n$ for some integer q. In this context

$$(t-\tau) \bmod N = \begin{cases} t-\tau & \text{if } N > t \geq \tau \\ t+N-\tau & \text{if } 0 \leq t < \tau \end{cases} \tag{14.52}$$

Alternatively, one can manipulate the periodic extension of $f(t)$ and $g(t)$ by repeating them indefinitely to produce $f_{\text{px}}(t)$ and $g_{\text{px}}(t)$ such that

$$h_c(t) = f(t) \circledast g(t) = \sum_{\tau=-\infty}^{\infty} f_{\text{px}}(\tau) g_{\text{px}}(t-\tau) \tag{14.53}$$

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

$$F[f \circledast g](k) = \hat{f}(k)\hat{g}(k) \tag{14.54}$$

This means that convolution in the physical/time domain is equivalent to component-wise multiplication in the frequency domain.

Proof. We will 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 14.3.: *An illustration for circular convolution.*

$$= \sum_{\tau=0}^{N-1} \sum_{t'=-\tau}^{N-1-\tau} f(\tau) g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N} t'\right) \exp\left(-i\frac{2\pi k}{N}\tau\right)$$

$$= \sum_{\tau=0}^{N-1} f(\tau) \left(\sum_{t'=-\tau}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N} t'\right)\right) \exp\left(-ih\frac{2\pi k}{N}\tau\right)$$

where we have made a change of variable from t to $t' = t - \tau$ with τ 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 $(t' - N) \bmod N = t'$, $\exp\left(-i\frac{2\pi k}{N}(t'-N)\right) = \exp\left(-i(\frac{2\pi k}{N}t' - 2\pi k)\right) = \exp\left(-i\frac{2\pi k}{N}t'\right)$. In the last line, we simply invoke Definition 14.1.2. Subsequently,

$$\begin{aligned}
&\sum_{\tau=0}^{N-1} f(\tau) \left(\sum_{t'=-\tau}^{N-1-\tau} g(t' \bmod N) \exp\left(-i\frac{2\pi k}{N}t'\right)\right) \exp\left(-i\frac{2\pi k}{N}\tau\right) \\
&= \sum_{\tau=0}^{N-1} f(\tau)\hat{g}(k) \exp\left(-i\frac{2\pi k}{N}\tau\right) \\
&= \left(\sum_{\tau=0}^{N-1} f(\tau) \exp\left(-i\frac{2\pi k}{N}\tau\right)\right) \hat{g}(k) = \hat{f}(k)\hat{g}(k)
\end{aligned}$$

where we use Definition 14.1.2 again and the desired equality is established. □

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 14.1.4 as $(11, 0.573 + 0.588i, 3.927 - 0.951i, 3.927 + 0.951i, 0.573 - 0.588i)$. 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 $(f \circledast g)(3) = 18$ and $(f \circledast g)(4) = 11$. Thus $(f \circledast g)(t) = (11, 21, 5, 18, 11)$ and by Definition 14.1.2 its DFT is

$$(\widehat{f \circledast g})(0) = 11 + 21 + 5 + 18 + 11 = 66$$

$$
\begin{aligned}
(\widehat{f \circledast g})(1) &= 11 + 21\exp\left(-i\frac{2\pi}{5}\right) + 5\exp\left(-i\frac{2\pi(2)}{5}\right) \\
&\quad + 18\exp\left(-i\frac{2\pi(3)}{5}\right) + 11\exp\left(-i\frac{2\pi(4)}{5}\right) = 2.281 - 1.869i \\
(\widehat{f \circledast g})(2) &= 11 + 21\exp\left(-i\frac{2\pi(2)}{5}\right) + 5\exp\left(-i\frac{2\pi(2)(2)}{5}\right) \\
&\quad + 18\exp\left(-i\frac{2\pi(2)(3)}{5}\right) + 11\exp\left(-i\frac{2\pi(2)(4)}{5}\right) \\
&= -7.781 - 18.242i
\end{aligned}
$$

Therefore, we can verify that

$$
\begin{aligned}
&F[f \circledast g](k) \\
&= (66, 2.281 - 1.869i, -7.781 - 18.242i, -7.781 + 18.242i, 2.281 + 1.869i) \\
&= (11, 0.573 + 0.588i, 3.927 - 0.951i, 3.927 + 0.951i, 0.573 - 0.588i) \\
&\quad \odot (6, 0.309 - 3.580i, -0.809 - 4.841i, -0.809 + 4.841i, 0.309 + 3.580i) \\
&= \hat{f}(k)\hat{g}(k)
\end{aligned}
$$

Theorem 14.2.4 holds, where $\odot$ represents component-wise multiplication. □

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 $O(N^2)$. This 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 (FFT)***. 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

$$G^* = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_N & \omega_N^2 & \omega_N^3 & \cdots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \omega_N^6 & \cdots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \omega_N^{3(N-1)} & \cdots & \omega_N^{(N-1)^2} \end{bmatrix} \tag{14.55}$$

where $\omega_N = \exp\left(-i\frac{2\pi}{N}\right)$ is the *base/fundamental frequency* (see the coming Footnote 15). The algorithm works most effectively when N is the power of 2. For an easier explanation, we consider a smaller value of $N = 2^3 = 8$. Then we denote

$$G^* = F_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8 & \omega_8^2 & \omega_8^3 & \omega_8^4 & \omega_8^5 & \omega_8^6 & \omega_8^7 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8^8 & \omega_8^{10} & \omega_8^{12} & \omega_8^{14} \\ 1 & \omega_8^3 & \omega_8^6 & \omega_8^9 & \omega_8^{12} & \omega_8^{15} & \omega_8^{18} & \omega_8^{21} \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^{16} & \omega_8^{20} & \omega_8^{24} & \omega_8^{28} \\ 1 & \omega_8^5 & \omega_8^{10} & \omega_8^{15} & \omega_8^{20} & \omega_8^{25} & \omega_8^{30} & \omega_8^{35} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^{24} & \omega_8^{30} & \omega_8^{36} & \omega_8^{42} \\ 1 & \omega_8^7 & \omega_8^{14} & \omega_8^{21} & \omega_8^{28} & \omega_8^{35} & \omega_8^{42} & \omega_8^{49} \end{bmatrix} \tag{14.56}$$

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

$$\hat{\mathbf{f}} = \begin{bmatrix} I_4 & D_{4\to 8} \\ I_4 & -D_{4\to 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} \begin{bmatrix} \mathbf{f}_{\text{even}} \\ \mathbf{f}_{\text{odd}} \end{bmatrix} \tag{14.57}$$

where

$$D_{4\to 8} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \omega_8 & 0 & 0 \\ 0 & 0 & \omega_8^2 & 0 \\ 0 & 0 & 0 & \omega_8^3 \end{bmatrix} \tag{14.58}$$

and

$$F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_4 & \omega_4^2 & \omega_4^3 \\ 1 & \omega_4^2 & \omega_4^4 & \omega_4^6 \\ 1 & \omega_4^3 & \omega_4^6 & \omega_4^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} \end{bmatrix} \tag{14.59}$$

where $\omega_4 = \exp\left(-i\frac{2\pi}{4}\right) = \exp\left(-i\frac{2\pi}{8}(2)\right) = \omega_8^2$. Subsequently,

$$\begin{aligned} \begin{bmatrix} I_4 & D_{4\to 8} \\ I_4 & -D_{4\to 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} &= \begin{bmatrix} I_4F_4 & D_{4\to 8}F_4 \\ I_4F_4 & -D_{4\to 8}F_4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8 & \omega_8^3 & \omega_8^5 & \omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^2 & \omega_8^6 & \omega_8^{10} & \omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^3 & \omega_8^9 & \omega_8^{15} & \omega_8^{21} \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & -\omega_8 & -\omega_8^3 & -\omega_8^5 & -\omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & -\omega_8^2 & -\omega_8^6 & -\omega_8^{10} & -\omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & -\omega_8^3 & -\omega_8^9 & -\omega_8^{15} & -\omega_8^{21} \end{bmatrix} \end{aligned} \tag{14.60}$$

Note that $\omega_N^{k+qN} = \omega_N^k$ [15] for any integer q and $-1 = \omega_N^{N/2}$, hence the matrix can be further rewritten into

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8 & \omega_8^3 & \omega_8^5 & \omega_8^7 \\ 1 & \omega_8^4 & \omega_8^8 & \omega_8^{12} & \omega_8^2 & \omega_8^6 & \omega_8^{10} & \omega_8^{14} \\ 1 & \omega_8^6 & \omega_8^{12} & \omega_8^{18} & \omega_8^3 & \omega_8^9 & \omega_8^{15} & \omega_8^{21} \\ 1 & \omega_8^8 & \omega_8^{16} & \omega_8^{24} & \omega_8^4 & \omega_8^{12} & \omega_8^{20} & \omega_8^{28} \\ 1 & \omega_8^{10} & \omega_8^{20} & \omega_8^{30} & \omega_8^5 & \omega_8^{15} & \omega_8^{25} & \omega_8^{35} \\ 1 & \omega_8^{12} & \omega_8^{24} & \omega_8^{36} & \omega_8^6 & \omega_8^{18} & \omega_8^{30} & \omega_8^{42} \\ 1 & \omega_8^{14} & \omega_8^{28} & \omega_8^{42} & \omega_8^7 & \omega_8^{21} & \omega_8^{35} & \omega_8^{49} \end{bmatrix} \tag{14.61}$$

which is essentially the same as F_8 except the columns have been rearranged into even and odd just like what we have done to $\mathbf{f}$. Hence the FFT formulation (14.57)

$$\hat{\mathbf{f}} = \begin{bmatrix} I_4 & D_{4\to 8} \\ I_4 & -D_{4\to 8} \end{bmatrix} \begin{bmatrix} F_4 & 0 \\ 0 & F_4 \end{bmatrix} \begin{bmatrix} \mathbf{f}_{\text{even}} \\ \mathbf{f}_{\text{odd}} \end{bmatrix}$$

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 14.4 below. For a bigger N, e.g. $N = 2^{10} = 1024$, the method proceeds in the same principle such that $F_{1024} \to F_{512} \to F_{256} \to \cdots$. The FFT time complexity is *log-linear* $O(n \log n)$, where the $\log n$ factor has replaced n thanks to its bisection 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.

[15] $\omega_N^{k+qN} = \exp\left(-i\frac{2\pi}{N}(k+qN)\right) = \exp\left(-i(\frac{2\pi}{N}k + 2\pi q)\right) = \exp\left(-i\frac{2\pi}{N}k\right) = \omega_N^k$.

Figure 14.4.: *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 the 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-year 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 FFT-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. Alternatively, write np.abs(
    Nino34_fft)**2.
```

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
```

Finally, we can plot the power spectrum as follows (Figure 14.5).

```
import matplotlib.pyplot as plt

def reciprocal(x): # Function to transform between frequency
    and period for the secondary axis.
     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", dpi=300, bbox_inches="tight")
```

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 applying 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.

Figure 14.5.: *Power Spectrum of the 120-year Niño 3.4 SST.*

```
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("Nino 3.4 Index")
plt.legend()
plt.title("ENSO-filtered Nino 3.4 time series")
plt.savefig("NinoENSOFilter", dpi=300, bbox_inches="tight")
```

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 for a heuristic purpose.

For more information, search about the *"Gibbs Phenomenon"* and also read the comprehensive discussion in this DSP StackExchange post (6220).

Figure 14.6.: *Raw and ENSO-filtered Niño 3.4 signals from 1981 to 2020 data.*

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 $\omega = \frac{2\pi(3)}{10}$.

Exercise 14.2 Download an ERA5 Temperature Dataset over any time period of 20 years. Select any location as you like, and extract the temperature time series there. Apply DFT to the time series and identify any dominant frequency or period with a large power. 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 11.8 and deduce the characteristic 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 it with the one in the `scipy.fft` library by testing them on any time series.

Chapter 15

Markov Chains

In Earth Science, we often look at the evolution of variables in time, like temperature, rainfall, etc. Sometimes, the variables can be modeled as *random variables*. A simple daily example will be tossing a coin (no matter if it is fair or not), where the outcomes (head/tail) are probabilistic. Some Earth System examples are the chances of extreme weather and slow geological processes. Such processes involving random variables are known as *Stochastic Processes*, and we will visit the related Statistical concepts. Particularly, we will investigate the so-called *Markov Chains*, which assumes that the present state of a stochastic process only depends on the past. Modeling a stochastic process with a Markov Chain can give more satisfactory results when it is known that the process inherently correlates strongly with previous states, or is said to have *memory*.

15.1. Statistical Prerequisites for Markov Chains

15.1.1. Lagged Auto-correlation

First, let's talk about how to determine if a ***time series*** possesses some sort of *memory* as mentioned in the introduction. Memory causes the past of a time series to influence the present, and hence will leave a mark when we try to

correlate the past and present segments of the time series. This leads to the idea of ***lagged auto-correlation***, which is the correlation between a time series and itself, but one of them is lagged and shifted by a certain amount of days, let's say k (the direction of the shifting does not matter as it is an auto-correlation, and we only care about the overlapping part). Then, it is more specifically known as the *Lag-k Auto-correlation*.

Definition 15.1.1 (Lagged Auto-correlation)**.** The lag-k auto-correlation r_k of a time series $\{x_i\}_{i=1}^n$ is defined as

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x_-})(x_{i+k} - \overline{x_+})}{\sqrt{(\sum_{i=1}^{n-k}(x_i - \overline{x_-})^2)(\sum_{j=k+1}^{n}(x_j - \overline{x_+})^2)}} \tag{15.1}$$

the correlation (Definition 11.2.4) between $\{x_-\} = \{x_i\}_{i=1}^{n-k}$ and $\{x_+\} = \{x_i\}_{i=k+1}^{n}$ where they are the two sub-sequences extracted from the earliest and latest $n-k$ data of the original time series, and the overline denotes an average. Alternatively, the formula can be written as

$$r_k = \frac{\mathrm{Cov}(\{x_-\}, \{x_+\})}{\sqrt{\mathrm{Var}(\{x_-\})\mathrm{Var}(\{x_+\})}} \tag{15.2a}$$

$$= \frac{\mathrm{Cov}(\{x_-\}, \{x_+\})}{\sqrt{\mathrm{Cov}(\{x_-\}, \{x_-\})\mathrm{Cov}(\{x_+\}, \{x_+\})}} \tag{15.2b}$$

where the definitions of variance and covariance are based on Definitions 11.2.1 and 11.2.3.

If the lag-k auto-correlation of a time series is close to 1 or -1, it means that the state at a certain time will likely lead to a similar/opposite state k time steps later. This may be used to infer causality but the implication is not always definite. There may be situations when the generated time series is regarded to be infinitely long, and in such cases, we simply apply the shifting and compute the lagged summation over the entire time axis.

Example 15.1.1. For the traffic flow data over some seven days of a highway below, find its lag-1 auto-correlation.

Day	1	2	3	4	5	6	7
Vehicles per Hour	680	820	760	790	840	1030	1080

We first locate the first and last $7 - 1 = 6$ data and construct the two smaller time series, which are $\{x_-\} = \{680, 820, 760, 790, 840, 1030\}$ and $\{x_+\} = \{820, 760, 790, 840, 1030, 1080\}$. The means of the two time series are easily found to be $\overline{x_-} = 820$ and $\overline{x_+} = 2660/3$ (the readers can verify these numbers), and the sample variances of the two time series are thus

$$\begin{aligned}
\mathrm{Var}(\{x_-\}) &= \frac{1}{6-1}[(680-820)^2 + (820-820)^2 + (760-820)^2 \\
&\quad + (790-820)^2 + (840-820)^2 + (1030-820)^2] \\
&= 13720 \text{ (Vehicles per Hour)}^2 \\
\mathrm{Var}(\{x_+\}) &= \frac{1}{6-1}\left[\left(820 - \frac{2660}{3}\right)^2 + \left(760 - \frac{2660}{3}\right)^2 + \left(790 - \frac{2660}{3}\right)^2\right. \\
&\quad \left. + \left(840 - \frac{2660}{3}\right)^2 + \left(1030 - \frac{2660}{3}\right)^2 + \left(1080 - \frac{2660}{3}\right)^2\right] \\
&= 17987 \text{ (Vehicles per Hour)}^2
\end{aligned}$$

and their sample covariance is

$$\begin{aligned}
&\mathrm{Cov}(\{x_-\}, \{x_+\}) \\
&= \frac{1}{6-1}\left[(680-820)\left(820 - \frac{2660}{3}\right) + (820-820)\left(760 - \frac{2660}{3}\right)\right. \\
&\quad + (760-820)\left(790 - \frac{2660}{3}\right) + (790-820)\left(840 - \frac{2660}{3}\right) \\
&\quad \left. +(840-820)\left(1030 - \frac{2660}{3}\right) + (1030-820)\left(1080 - \frac{2660}{3}\right)\right]
\end{aligned}$$

$$= 12000 \text{ (Vehicles per Hour)}^2$$

Hence by Formula (15.2) in Definition 15.1.1 above, the lag-1 auto-correlation is

$$r_1 = \frac{12000}{\sqrt{(13720)(17987)}} = 0.7639$$

It means that heavy (quiet) traffic will likely be followed by more or less heavy (quiet) traffic the next day, which is quite reasonable.

► Short Exercise: Compute the lag-2 auto-correlation for the same dataset.[1]

15.1.2. Conditional Probabilities, Stochastic Matrices

To understand the idea of Markov Chains, we also need to know what *conditional probabilities* are. The ***conditional probability*** $P(A|B)$ is the probability of event A occurring, given event B has occurred. For an evolving system that has a finite number of states A_i, $i = 1, 2, \ldots, n$, and can only possess one state at a time, e.g. a binary on-and-off, the conditional probability $P(A_i^{[k+1]}|A_j^{[k]})$ represents the probability of state A_i occurring at time step $k+1$ if the state is A_j at time step k. A simple Atmospheric Sciences example is the daily weather report, which in a simplistic sense, can take the state of either sunny, rainy, or windy. Usually, if it is rainy today, then there is a relatively high chance it is rainy tomorrow, i.e. $P(\text{rainy}^{[k+1]}|\text{rainy}^{[k]})$ is high.

For N finite, distinct states/events in a well-defined, *closed* system so that it takes one and only one of the events as its state, we have the following observation.

Properties 15.1.2. The sum of conditional probabilities for a changing system with N *mutually exclusive* (only one state at a time) and *exhaustive* (the states

cover all possibilities) events $A_1, A_2, \cdots, A_N$ is

$$\sum_{i=1}^{N} P(A_i^{[k+1]}|A_j^{[k]}) = P(A_1^{[k+1]}|A_j^{[k]}) + P(A_2^{[k+1]}|A_j^{[k]}) + \cdots + P(A_N^{[k+1]}|A_j^{[k]})$$
$$P(\Omega^{[k+1]}|A_j^{[k]}) = 1 \tag{15.3}$$

where $\Omega = \bigcup_i A_i$ covers all the states and k denotes a particular time step. The time increment is 1 here but can be replaced by any other positive integer. It means that any given event A_j must consequently lead to one of the possible states (including A_j itself) at the next (few) time step(s).

This is essentially a rephrasing of the *Law of Total Probability* from elementary Statistics. As a result, we can express all the conditional probabilities $P(A_i^{[k+1]}|A_j^{[k]})$ for a particular previous state A_j in a system using a column vector with components that sum up to 1. Using the daily weather as an example, assume there are only three states (sunny/windy/rainy), if a sunny day has a chance of 0.8 to be followed by another sunny day, and 0.15/0.05 for another windy/rainy day. Then we can write

$k+1 \setminus k$	Sunny
Sunny	0.8
Windy	0.15
Rainy	0.05

as a column vector

$$\begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix}$$

[1]The new $\{x_-\}$ and $\{x_+\}$ will be $\{680, 820, 760, 790, 840\}$ and $\{760, 790, 840, 1030, 1080\}$. We provide the relevant numbers for the readers to check. Their sample (co)variances are $\mathrm{Var}(\{x_-\}) = 3920$, $\mathrm{Var}(\{x_+\}) = 21150$ and $\mathrm{Cov}(\{x_-\}, \{x_+\}) = 5725$. So the lag-2 auto-correlation will be $r_2 = \frac{5725}{\sqrt{(3920)(21150)}} = 0.6287$.

We can do the same for the other two states. If a windy day has a probability of 0.2/0.5/0.3 leading to a sunny/windy/rainy day, and a rainy day has a probability of 0.1/0.2/0.7 leading to a sunny/windy/rainy day, then

$k+1 \setminus k$	Sunny	Windy	Rainy
Sunny	0.8	0.2	0.1
Windy	0.15	0.5	0.2
Rainy	0.05	0.3	0.7

This can be summarized by the so-called ***stochastic (transition) matrix*** as

$$P = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix}$$

Definition 15.1.3 (Stochastic Matrix)**.** The stochastic matrix for a closed system with mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_N$, is

$$P = \begin{bmatrix} P(A_1^{[k+1]}|A_1^{[k]}) & P(A_1^{[k+1]}|A_2^{[k]}) & \cdots & P(A_1^{[k+1]}|A_N^{[k]}) \\ P(A_2^{[k+1]}|A_1^{[k]}) & P(A_2^{[k+1]}|A_2^{[k]}) & & P(A_2^{[k+1]}|A_N^{[k]}) \\ \vdots & & \ddots & \vdots \\ P(A_N^{[k+1]}|A_1^{[k]}) & P(A_N^{[k+1]}|A_2^{[k]}) & \cdots & P(A_N^{[k+1]}|A_N^{[k]}) \end{bmatrix} \tag{15.4}$$

where $P_{ij} = P(A_i^{[k+1]}|A_j^{[k]})$ is the conditional probability of the system moving to state i at the next time step from the present state j.

Again, notice that the entries along any column add up to 1. The j-th column holds the conditional probabilities of state j leading to different states at the next time step.

15.2. Construction of and Prediction by Markov Chains

15.2.1. State Vector, Steady State

Processes that can be represented by such stochastic matrices proposed above are known as ***Markov Chains***. It is assumed that the conditional probabilities outlined in the stochastic matrices P do not change in time (*stationary*). Given a probability vector (or a ***state vector***) $\vec{x}^{[k]}$ consisting of the probabilities of having different states at a certain time step k, we can calculate the probability vector $\vec{x}^{[k+1]}$ at the next time step $k+1$ as $P\vec{x}^{[k]}$.

Properties 15.2.1. Given a Markov Chain, with a stochastic matrix P described by (15.4), then the state vector $\vec{x}^{[k+1]}$ at time step $k+1$ is decided by

$$\vec{x}^{[k+1]} = P\vec{x}^{[k]} \tag{15.5}$$

Proof. If we look at the i-th entry on both sides, we have

$$\vec{x_i}^{[k+1]} = P_{i1}\vec{x_1}^{[k]} + P_{i2}\vec{x_2}^{[k]} + P_{i3}\vec{x_3}^{[k]} \cdots$$

When explicitly written in terms of (conditional) probabilities, it is

$$\begin{aligned} P(A_i^{[k+1]}) &= P(A_i^{[k+1]}|A_1^{[k]})P(A_1^{[k]}) + P(A_i^{[k+1]}|A_2^{[k]})P(A_2^{[k]}) + \cdots \\ &= \sum_{j=1}^{N} P(A_i^{[k+1]}|A_j^{[k]})P(A_j^{[k]}) \end{aligned}$$

which is exactly the manifestation of the *Law of Total Probability* again, given the events are mutually exclusive and exhaustive. □

Similarly, at the $k+2$-th time step, the probability vector is $\vec{x}^{[k+2]} = P^2\vec{x}^{[k]}$. In general, we have $\vec{x}^{[1]} = P\vec{x}^{[0]}, \vec{x}^{[2]} = P\vec{x}^{[1]} = P(P\vec{x}^{[0]}) = P^2\vec{x}^{[0]}$ and

$$\vec{x}^{[n]} = P^n\vec{x}^{[0]} \qquad (\vec{x}^{[k+n]} = P^n\vec{x}^{[k]}) \tag{15.6}$$

Example 15.2.1. Using the previous example of daily weather, the stochastic matrix is

$$P = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix}$$

Find the probabilities of each type of weather occurring on Day 2 and Day 3, if we know that the chances of being sunny/windy/rainy on Day 1 are 0.3/0.4/0.3.

Solution. By Formula (15.6) we have just derived, the required state vector is

$$\begin{aligned} \vec{x}^{[2]} &= P\vec{x}^{[1]} \\ &= \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix} \\ &= 0.3 \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix} + 0.4 \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix} + 0.3 \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix} \\ &= \begin{bmatrix} 0.35 \\ 0.305 \\ 0.345 \end{bmatrix} \end{aligned}$$

So on Day 2, the chances of sunny/windy/rainy are 0.35/0.305/0.345. It is emphasized that in the end, the required state vector on the next day is just the linear combination of the column vectors that contain the conditional probabilities, with the weightings specified by the current state vector. Again, it

is the Law of Total Probability working. Similarly, for Day 3, the state vector will be

$$\begin{aligned}
\vec{x}^{[3]} &= P\vec{x}^{[2]} = P^2\vec{x}^{[1]} \\
&= \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix} \begin{bmatrix} 0.35 \\ 0.305 \\ 0.345 \end{bmatrix} \\
&= 0.35 \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix} + 0.305 \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix} + 0.345 \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix} \\
&= \begin{bmatrix} 0.3755 \\ 0.274 \\ 0.3505 \end{bmatrix}
\end{aligned}$$

□

► Short Exercise: Find the state vector on the next day if today is windy.[2]

For a Markov Chain which has been ongoing for a long period of time, and the initial state becomes effectively forgotten (*memoryless*), it is desirable to obtain the ***steady-state vector*** which represents the "average" probability of every state given no prior knowledge of the initial state. The steady-state vector $\vec{q}$ remains the same for any time step and the probabilities are stationary. Hence, we have

$$\vec{q} = P\vec{q} = \cdots = P^n\vec{q} \tag{15.7}$$

Rearrangement of the relation $\vec{q} = P\vec{q}$ gives

$$(I - P)\vec{q} = \mathbf{0} \tag{15.8}$$

which can be recognized as an eigenvalue-eigenvector problem (see Section 9.1): The steady-state vector $\vec{q}$ corresponds to the eigenvector of P with an eigenvalue of $\lambda = 1$. This eigenvector $\vec{q}$ is then found by following the usual procedure

to solve the linear system $(I - P)\vec{q} = \mathbf{0}$. Some may suspect if this steady-state vector always exists. To address this question, we will first show that one of the eigenvalues in any Markov Chain must be 1.

Properties 15.2.2. The stochastic matrix P of a Markov chain always has an eigenvalue of $\lambda = 1$. The steady state of a Markov Chain is then simply the eigenvector of P that corresponds to the eigenvalue of $\lambda = 1$ as in $(P - \lambda I)\vec{q} = \mathbf{0}$, and normalized by the sum of entries so that they add up to 1.

Proof. We will show that the transpose P^T has an eigenvalue of $\lambda = 1$. Then by Properties 9.1.5, P must also share this same eigenvalue and we are done. Particularly, we can observe that $\vec{z} = (1, 1, 1, \ldots, 1)^T$ is an eigenvector of P^T that corresponds to the desired eigenvalue of 1, because

$$P^T\vec{z} = \begin{bmatrix} P_{11} & P_{21} & \cdots & P_{n1} \\ P_{12} & P_{22} & \cdots & P_{n2} \\ \vdots & & \ddots & \vdots \\ P_{1n} & P_{2n} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} + P_{21} + \cdots + P_{n1} \\ P_{12} + P_{22} + \cdots + P_{n2} \\ \vdots \\ P_{1n} + P_{2n} + \cdots + P_{nn} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \vec{z}$$

where by Properties 15.1.2 the entries in each column of P sum to one: $P_{1j} + P_{2j} + \cdots + P_{nj} = 1$. □

Example 15.2.2. Find the steady-state vector in the last example about daily

[2]It is simply $P\vec{x}$ where $\vec{x} = (0, 1, 0)^T$ and thus

$$\begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}$$

so the probabilities are simply given by the components in the second column of the stochastic matrix.

weather where

$$P = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.5 & 0.2 \\ 0.05 & 0.3 & 0.7 \end{bmatrix}$$

Solution. We seek to solve the linear system

$$I - P = \begin{bmatrix} 0.2 & -0.2 & -0.1 \\ -0.15 & 0.5 & -0.2 \\ -0.05 & -0.3 & 0.3 \end{bmatrix}$$

and a simple calculation by Gaussian Elimination reveals that the desired eigenvector of $\lambda = 1$ is $\vec{q} = (\frac{9}{7}, \frac{11}{14}, 1)^T$ (You may check by substituting this into $(I - P)\vec{q} = \mathbf{0}$). Since the probabilities have to add up to 1, we divide each component by their sum and this yields the steady-state vector $\vec{q} = (\frac{18}{43}, \frac{11}{43}, \frac{14}{43})^T \approx (0.419, 0.256, 0.326)^T$, meaning that there is 41.9%/25.6%/32.6% chance of having a sunny/windy/rainy day on average. □

Nevertheless, there is a subtlety. We have only shown that there is always an eigenvalue of 1 (or in general it is the modulus $|\lambda| = 1$ equals to one, as we will see soon) for a Markov Chain, but we don't know if there is only one unique corresponding eigenvector. Consider an extreme example where there exist three states, A, B, and C. State A always leads to state B, which in turn always leads to state C, and state C will always move back to state A. The stochastic matrix is then simply

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

It can be found that there are three eigenvalues $\lambda = 1, -\frac{1}{2} + \frac{\sqrt{3}}{2}i, -\frac{1}{2} - \frac{\sqrt{3}}{2}i$ where the two complex eigenvalues have a modulus of 1 as well. This system can be

easily seen to be unstable. Any state vector other than $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$ (the eigenvector for $\lambda = 1$) will keep oscillating where the components are "cycled". Therefore, the system will not converge to the steady state despite the existence of a "steady-state" vector of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$. The imaginary part of the two complex eigenvalues represents a rotation of the two complex eigenvectors and the modulus of $|\lambda| = 1$ means that the magnitude of this rotation will not decay.[3] So there should be some restriction if the system has to converge to a unique steady state. It turns out that a sufficient condition is whether the Markov Chain is *regular*. To proceed, we need to introduce some terminologies.

Definition 15.2.3 (Regular Markov Chain)**.** A stochastic matrix P, or the corresponding Markov Chain, is said to be ***regular*** if some power of the stochastic matrix P^k, where k is any positive integer, has all positive entries, i.e. P^k is a ***positive stochastic matrix*** (or simply positive).

Properties 15.2.4. For a positive stochastic matrix A, the only eigenvalue with a modulus $|\lambda| = 1$ is $\lambda = 1$. Moreover, the geometric multiplicity of this eigenvalue $\lambda = 1$ belonging to the matrix A is strictly 1, i.e. there is only one eigenvector corresponding to $|\lambda| = 1$.

The proof is provided in Appendix A.4. For a regular stochastic matrix P, $A = P^k$ is positive for some k by Definition 15.2.3 and hence the only eigenvalue of P^k that has a modulus of 1 is $\lambda = 1$ and its geometric multiplicity is 1 by Properties 15.2.4 above. It is not hard to show that, therefore, P will also have $\lambda = 1$ as the only eigenvalue that has a modulus of one, $|\lambda| = 1$, and the geometric multiplicity is also 1.[4] Hence by combining the two results, we have

[3] For the complex eigenvalues λ_c and complex eigenvectors $\vec{x}_c$, $\|P\vec{x}_c\|^2 = \|\lambda\vec{x}_c\|^2 = |\lambda|^2\|\vec{x}_c\|^2 = \|\vec{x}_c\|^2$ as $|\lambda| = 1$. However, the entries of a state vector are always real-valued and it will be a linear combination of these conjugate complex eigenvectors where the imaginary part is exactly canceled out underneath.

Properties 15.2.5. A regular Markov chain always has a unique steady-state vector with an eigenvalue of 1.

The stochastic matrix in Example 15.2.2 is obviously regular, hence we have been able to derive the steady state for it. However, note that Properties 15.2.5 provides a sufficient condition only, and a non-regular Markov chain may still converge to a unique steady state. Take

$$P = \begin{bmatrix} 1 & 0.5 \\ 0 & 0.5 \end{bmatrix}$$

as an example. It is not hard to see that the bottom-left entry in P^k where k is any positive integer, will always be 0.[5] However, it is also easy to see that given any starting probability vector, it will approach the steady-state vector $(1, 0)^T$ as it goes on. Another problem is that in Definition 15.2.3 we haven't specified how large the k we have to keep checking and when we can stop. For the small example above and by the accompanying footnote, knowing P^2 is not positive is adequate to decide that it is not regular. For completeness, we note

[4]Assume the contrary that there are two different eigenvectors of P, $\vec{q}^{(1)}$ and $\vec{q}^{(2)}$, the eigenvalues of which have a modulus $|\lambda_1| = |\lambda_2| = 1$ of one. Then $P^k\vec{q}^{(1)} = P^{k-1}(P\vec{q}^{(1)}) = \lambda_1 P^{k-1}\vec{q}^{(1)} = \lambda_1 P^{k-2}(P\vec{q}^{(1)}) = \lambda_1^2 P^{k-2}\vec{q}^{(1)} = \cdots = \lambda_1^k\vec{q}^{(1)}$, and similarly we have $P^k\vec{q}^{(2)} = \lambda_2^k\vec{q}^{(2)}$. These show that $\vec{q}^{(1)}$ and $\vec{q}^{(2)}$ will be also two eigenvectors to $A = P^k$ with an eigenvalue of λ_1^k and λ_2^k where both of their modulus $\left|\lambda_j^k\right| = \left|\lambda_j\right|^k = 1$, $j = 1, 2$, equal to one, contradicting Properties 15.2.4.

[5]In general, for any two 2×2 matrices both in the form of

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

their product

$$\begin{bmatrix} * & * \\ 0 & * \end{bmatrix}\begin{bmatrix} * & * \\ 0 & * \end{bmatrix} = \begin{bmatrix} * & * \\ (0)(*) + (*)(0) & * \end{bmatrix} = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}$$

will take such a form too.

that in general, given an $n \times n$ stochastic matrix P, the required k to test is up to $n^2 - 2n + 2$ (so for $n = 2$, we will check $k = 1, 2$; and $n = 3$, $k = 1, 2, 3, 4, 5$).[6]

15.2.2. Expected Value Problem

For the special case in which one of the states in Markov Chain is *absorbing*, i.e. the probability of staying in this node/state is 1 and the probabilities of moving to other states are all 0, all pathways will eventually lead to accumulation in this particular *sink* no matter what the initial state is. This is referred to as an ***absorbing Markov Chain*** and it is possible to predict the expected time needed for any state to arrive at the absorbing state. Note that the previous example of

$$P = \begin{bmatrix} 1 & 0.5 \\ 0 & 0.5 \end{bmatrix}$$

is exactly an absorbing Markov Chain where the sink is in the first node.

We can attack this problem by finding the total expected time spent in other non-absorbing nodes which is equivalent to the expected time to arrive at the sink. At the zeroth time step, or $t = 0$, the time spent in the starting node and other nodes is one and zero unit time respectively. At $t = k$, the average times spent during the k-th time step in different nodes are exactly the probabilities of landing on these nodes after k time steps.

Hence, the idea is to add up the times staying in the non-absorbing nodes over every time step. Therefore, we consider the variant of the stochastic matrix in which the row and column of the absorbing state are deleted. Assume that there are three states in the absorbing Markov Chain, and its stochastic matrix is

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.7 & 0 & 0 \\ 0.1 & 0.2 & 1 \end{bmatrix}$$

[6]This "magic number" is due to a result by Wielandt but the proof is a bit too advanced to be included here.

Along the third column, the third component is equal to 1 and the others are 0, indicating that the third node is absorbing. After deleting the row and column related to the absorbing state, and only looking at those non-absorbing states, we have

$$P_0 = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0 \end{bmatrix}$$

Assume we start at the first node, i.e. the probability vector representing the initial state is $(1, 0)^T$ in this case, the formula to calculate the total staying time in the non-absorbing nodes over all time steps would be

$$I \begin{bmatrix} 1 \\ 0 \end{bmatrix} + P \begin{bmatrix} 1 \\ 0 \end{bmatrix} + P^2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \cdots = (I + P_0 + P_0^2 + \cdots) \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

where the first term on L.H.S. is the time staying in other nodes at the zeroth time step, the second term is that at the first time step, and so on. t_1 and t_2 will be the total expected time spent at nodes 1 and 2 throughout the process. For the case where the starting condition is probabilistic, e.g. 50% chance in the first node and 50% chance in the second node, we simply change $(1, 0)^T$ to $(0.5, 0.5)^T$. We can utilize the formula of *geometric sum for a matrix*, and generalize it to accommodate more states.

Properties 15.2.6. The expected time for an initial state in a Markov Chain to be absorbed into the sink (assumed that there is only one sink) can be inferred from the relation

$$\begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \end{bmatrix} = (I + P_0 + P_0^2 + \cdots)\vec{x}^{[0]} \tag{15.9a}$$

$$= (I - P_0)^{-1}\vec{x}^{[0]} \tag{15.9b}$$

where P is the associated stochastic matrix and P_0 is P with the row and column

of the absorbing state removed. $\vec{x}^{[0]}$ is the initial state vector, also with the entry corresponding to the absorbing node removed. The required absorption time to the sink is the sum of t_i over all valid i.

The only caveat is that the formula of geometric sum for a matrix may not hold in the second equality, just like the original formula for a number will fail if the common ratio has a magnitude larger than or equal to 1. However, it is guaranteed that, for a stochastic matrix without other sinks or closed loops which make traveling from any state to the sink impossible

$$(I - P_0)^{-1} = I + P_0 + P_0^2 + \cdots \tag{15.10}$$

indeed holds.[7]

Example 15.2.3. For the absorbing Markov Chain that we have been discussing, where

$$P = \begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.7 & 0 & 0 \\ 0.1 & 0.2 & 1 \end{bmatrix} \qquad P_0 = \begin{bmatrix} 0.2 & 0.8 \\ 0.7 & 0 \end{bmatrix}$$

Find the expected time required to reach the third node if we start from the first node.

Solution. We have

$$I - P_0 = \begin{bmatrix} 0.8 & -0.8 \\ -0.7 & 1 \end{bmatrix}$$

It is not hard to obtain (for example, just do it like in Example 2.3.5)

$$(I - P_0)^{-1} = \begin{bmatrix} 4.166 & 3.333 \\ 2.917 & 3.333 \end{bmatrix}$$

[7] An equivalent condition is that the original stochastic matrix P only has one eigenvalue with the value of $\lambda = 1$ that represents the sink and there are no other eigenvalues that also have a modulus of $|\lambda| = 1$.

$$(I - P_0)^{-1}\vec{x}^{[0]} = \begin{bmatrix} 4.166 & 3.333 \\ 2.917 & 3.333 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4.166 \\ 2.917 \end{bmatrix}$$

Hence the expected time required to reach the third node starting from the first node will be just the sum of the first column in P_0, $4.166 + 2.917 = 7.083$. □

► Short Exercise: Find the expected time required to reach the third node if it has 50%/50% chances to start at the first/second node instead.[8]

15.2.3. Lagged Auto-correlation Predicted by Markov Chains

The last topic we will touch on Markov Chains is their inherent auto-correlation (see the beginning of this chapter). Consider the simplest case, where a Markov Chain only has two possible states, with a 2×2 stochastic matrix, in the form of

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ 1 - P_{11} & 1 - P_{12} \end{bmatrix} \tag{15.11}$$

Assume the first/second state represents a value of 1/0 without the loss of generality. The theoretical lag-1 auto-correlation between the binary states, predicted by the Markov Chain, can be inferred from (15.2) in Definition 15.1.1, that is

$$\begin{aligned} r_1 &= \frac{\mathrm{Cov}(\{x_-\}, \{x_+\})}{\sqrt{\mathrm{Var}(\{x_-\})\mathrm{Var}(\{x_+\})}} \\ &= \frac{E[X_- X_+] - E[X_-]E[X_+]}{\sqrt{(E[X_-^2] - (E[X_-])^2)(E[X_+^2] - (E[X_+])^2)}} \end{aligned}$$

where we have used the short-cut formulae for variance and covariance listed in Section 11.2.1. If the Markov Chain has operated for a long enough time, then

the means of the two time series, despite lagged by a day, will be equal due to the loss of memory, implying that

$$E[X_-] = E[X_+] = q_1 \tag{15.12}$$

where $\vec{q}$ is the steady-state vector, with q_1 being the average probability of being in state 1. Similarly, we have

$$E[X_-^2] = E[X_+^2] = q_1 \tag{15.13}$$

since squaring a time series composed purely of 1 and 0 will return itself. The remaining job is to determine $E[X_-X_+]$, which is

$$\begin{aligned} E[X_-X_+] &= P(X_+ = 1 \text{ and } X_- = 1) \\ &= P(X_+ = 1|X_- = 1)P(X_- = 1) \\ &= P_{11}q_1 \end{aligned} \tag{15.14}$$

as per the definition of conditional probability. So the lag-1 auto-correlation is

$$\begin{aligned} r_1 &= \frac{(P_{11}q_1 - q_1^2)}{\sqrt{(q_1 - q_1^2)(q_1 - q_1^2)}} \\ &= \frac{q_1(P_{11} - q_1)}{(q_1 - q_1^2)} \\ &= \frac{P_{11} - q_1}{1 - q_1} \end{aligned} \tag{15.15}$$

[8] It is simply a matter of computing

$$(I - P_0)^{-1}\vec{x}^{[0]} = \begin{bmatrix} 4.166 & 3.333 \\ 2.917 & 3.333 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 3.75 \\ 3.125 \end{bmatrix}$$

where $\vec{x}^{[0]}$ is now $(0.5, 0.5)^T$ and the required time is then $3.75 + 3.125 = 6.875$.

But, by the definition of the stationary eigenvector $P\vec{q} = \vec{q}$ in (15.7), considering its first component we have

$$\begin{aligned} P_{11}q_1 + P_{12}q_2 &= q_1 \\ P_{11}q_1 + P_{12}(1 - q_1) &= q_1 \\ (1 - P_{11} + P_{12})q_1 &= P_{12} \end{aligned} \tag{15.16}$$

Substituting this into the formula of lag-1 auto-correlation (15.15) just derived, we arrive at

$$\begin{aligned} r_1 = \frac{P_{11} - q_1}{1 - q_1} &= \frac{(1 - P_{11} + P_{12})(P_{11} - q_1)}{(1 - P_{11} + P_{12})(1 - q_1)} \\ &= \frac{P_{11}(1 - P_{11} + P_{12}) - (1 - P_{11} + P_{12})q_1}{1 - P_{11} + P_{12} - (1 - P_{11} + P_{12})q_1} \\ &= \frac{P_{11}(1 - P_{11} + P_{12}) - P_{12}}{1 - P_{11} + P_{12} - P_{12}} \\ &= \frac{(1 - P_{11})(P_{11} - P_{12})}{1 - P_{11}} \\ &= P_{11} - P_{12} \end{aligned} \tag{15.17}$$

Therefore, the magnitude of lag-1 auto-correlation predicted by a binary Markov Chain is simply the difference between the two conditional probabilities on the same row in the stochastic matrix. By the same essence, the lag-k auto-correlation can be inferred from $r_k = Q_{11} - Q_{12}$, where $Q = P^k$. Fortunately, due to the structure of stochastic matrices, we can derive a straightforward formula for lag-k auto-correlation, which is just the k-th power of the lag-1 auto-correlation, $r_k = r_1^k$.

Properties 15.2.7. The lag-k auto-correlation of a binary Markov Chain, associated with a stochastic matrix P, is

$$r_k = (P^k)_{11} - (P^k)_{12} = (P_{11} - P_{12})^k = r_1^k \tag{15.18}$$

where $r_1 = P_{11} - P_{12}$.

Proof. The formula holds for $k = 1$ as we just see, and now we will prove the equality $(P^k)_{11} - (P^k)_{12} = (P_{11} - P_{12})^k$ for any general k. First, it is instructive to find the eigenvalues and eigenvectors of P and diagonalize the matrix. We set out to solve the characteristic equation

$$\begin{aligned}
\det(P - \lambda I) = \begin{vmatrix} P_{11} - \lambda & P_{12} \\ 1 - P_{11} & 1 - P_{12} - \lambda \end{vmatrix} &= 0 \\
(P_{11} - \lambda)(1 - P_{12} - \lambda) - (1 - P_{11})P_{12} &= 0 \\
P_{11} - P_{11}P_{12} - \lambda P_{11} - \lambda + \lambda P_{12} + \lambda^2 - P_{12} + P_{11}P_{12} &= 0 \\
(P_{11} - P_{12}) - \lambda(1 + (P_{11} - P_{12})) + \lambda^2 &= 0 \\
(1 - \lambda)((P_{11} - P_{12}) - \lambda) &= 0 \qquad (15.19)
\end{aligned}$$

So the eigenvalues are $\lambda_1 = 1$ (as expected) and $\lambda_2 = P_{11} - P_{12}$. It is not hard to find the steady-state vector (not yet normalized) $\vec{q}$ that corresponds to $\lambda = 1$, which is

$$\vec{q} = \begin{bmatrix} P_{12} \\ 1 - P_{11} \end{bmatrix} \qquad (15.20)$$

and we leave it to the readers to check. For $\lambda_2 = P_{11} - P_{12}$, the eigenvector can also be easily found from

$$\begin{aligned}
&\left[\begin{array}{cc|c} P_{11} - (P_{11} - P_{12}) & P_{12} & 0 \\ 1 - P_{11} & 1 - P_{12} - (P_{11} - P_{12}) & 0 \end{array}\right] \\
&= \left[\begin{array}{cc|c} P_{12} & P_{12} & 0 \\ 1 - P_{11} & 1 - P_{11} & 0 \end{array}\right] \\
&\to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 - P_{11} & 1 - P_{11} & 0 \end{array}\right] && \frac{1}{P_{12}} R_1 \to R_1 \\
&\to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] && R_2 - (1 - P_{11})R_1 \to R_2
\end{aligned}$$

hence it has a very simple form of $(-1, 1)^T$. Using the same technique as in Example 9.2.3, we diagonalize P according to $S^{-1}PS = D$ where now we use S

to denote the matrix formed by the eigenvectors of P, and then

$$\begin{aligned}
P^k &= (SDS^{-1})^k \\
&= (SDS^{-1})(SDS^{-1})\cdots(SDS^{-1})(SDS^{-1}) \\
&= SD(S^{-1}S)D\cdots D(S^{-1}S)DS^{-1} = SD^kS^{-1} \\
&= \begin{bmatrix} P_{12} & -1 \\ 1-P_{11} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (P_{11}-P_{12})^k \end{bmatrix} \begin{bmatrix} P_{12} & -1 \\ 1-P_{11} & 1 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} P_{12} & -(P_{11}-P_{12})^k \\ 1-P_{11} & (P_{11}-P_{12})^k \end{bmatrix} \left(\frac{1}{P_{12}+(1-P_{11})} \begin{bmatrix} 1 & 1 \\ -(1-P_{11}) & P_{12} \end{bmatrix} \right) \\
&= \frac{1}{P_{12}+(1-P_{11})} \begin{bmatrix} P_{12}+(1-P_{11})(P_{11}-P_{12})^k & P_{12}-P_{12}(P_{11}-P_{12})^k \\ (1-P_{11})-(1-P_{11})(P_{11}-P_{12})^k & (1-P_{11})+P_{12}(P_{11}-P_{12})^k \end{bmatrix}
\end{aligned} \tag{15.21}$$

So

$$\begin{aligned}
&(P^k)_{11} - (P^k)_{12} \\
&= \frac{(P_{12}+(1-P_{11})(P_{11}-P_{12})^k) - (P_{12}-P_{12}(P_{11}-P_{12})^k)}{P_{12}+(1-P_{11})} \\
&= \frac{((1-P_{11})+P_{12})(P_{11}-P_{12})^k}{P_{12}+(1-P_{11})} = (P_{11}-P_{12})^k
\end{aligned}$$

the desired formula is established. □

For instance, given a two-state Markov Chain with the stochastic matrix

$$P = \begin{bmatrix} 0.35 & 0.75 \\ 0.65 & 0.25 \end{bmatrix}$$

Its lag-1 auto-correlation will be $0.35 - 0.75 = 0.25 - 0.65 = -0.4$, and the lag-4 auto-correlation will be $(-0.4)^4 = 0.0256$. As just shown, the lag-k auto-correlation of any (first-order) binary Markov Chain always decays exponentially.

15.3. Python Programming and Earth System Applications

Let's use the 2024 Hualien Earthquake sequence, the mainshock of which occurred on April 3rd as an illustration. First, download the .csv records from https://scweb.cwa.gov.tw/en-us/earthquake/data for April and May. We will first read the data in April and extract the largest earthquake magnitude each day.

```
import numpy as np
import pandas as pd

equake_Apr = pd.read_csv("Seismic activity_April.csv", usecols
    =np.arange(7), parse_dates=["Orgin date"])
equake_Apr_daily_mag_max = equake_Apr.groupby(equake_Apr["
    Orgin date"].dt.day)["Magnitude"].max()
```

Write a custom function to separate the cases into three levels of magnitude: 6.0–6.9 → 0, 5.0–5.9 → 1, 4.0–4.9 → 2.

```
# Three magnitude classes: 4.0-4.9, 5.0-5.9, 6.0-6.9
def classify_mag(mag):
    if mag >= 6.0:
        return(0)
    elif (5.0 <= mag <= 5.9):
        return(1)
    elif (mag <= 4.9):
        return(2)
```

Now comes the main part. Shift the categorized time series by one day and count the incidence of transitions between different states using `np.unique` and the function above, to construct the stochastic matrix.

```
equake_Apr_prev_states = (equake_Apr_daily_mag_max.loc[3:29].
    apply(classify_mag)).values
equake_Apr_next_states = (equake_Apr_daily_mag_max.loc[4:30].
    apply(classify_mag)).values
```

```
transition_matrix = np.zeros([3,3])
trans_states, Apr_counts = np.unique(np.c_[
    equake_Apr_next_states, equake_Apr_prev_states],
    return_counts=True, axis=0)

transition_matrix[trans_states[:,0], trans_states[:,1]] =
    Apr_counts
transition_matrix = transition_matrix/np.sum(transition_matrix
    , axis=0)
```

We have normalized the columns by their respective sums. Finally, we compute the steady-state eigenvector for this Markov Chain. (Note that the eigenvalue of 1 will be the one with the largest modulus.)

```
import scipy.linalg as linalg

eigvals, eigvecs = linalg.eig(transition_matrix)
q = eigvecs[:, np.argmax(np.abs(eigvals))] / np.sum(eigvecs[:,
     np.argmax(np.abs(eigvals))])
print(q)
```

This gives `[0.0767 0.4063 0.5170]`, which means that on average, during the active period on any day, the magnitude of the most intense earthquake will have 7.7%/40.6%/51.7% to be ≤ 4.9, within 5.0–5.9, and ≥ 6.0 respectively. This is obtained over the training dataset. Now let's read the file for May and see if the actual probabilities match those expected by our Markov Chain formulated from data in April.

```
equake_May = pd.read_csv("Seismic activity_May.csv", usecols=
    np.arange(7), parse_dates=["Orgin date"])
equake_May_daily_mag_max = equake_May.groupby(equake_May["
    Orgin date"].dt.day)["Magnitude"].max()
equake_May_states = (equake_May_daily_mag_max.apply(
    classify_mag)).values
May_counts = np.bincount(equake_May_states)
May_expected_counts_from_Markov = q*31
print(May_counts, May_expected_counts_from_Markov)
```

which returns `[2 5 24]` and `[2.4 12.6 16.0]` correspondingly. There are differences clearly in the last two bins and we can confirm it using the *Chi-square Test*:

```
from scipy.stats import chisquare
print(chisquare(f_obs=May_counts, f_exp=
    May_expected_counts_from_Markov))
```

that yields a p-value of `0.0135` which exceeds the 5% significance level. Therefore, we can conclude that the actual distribution of Earthquake events over the testing set deviates from that predicted by the Markov Chain with 95% confidence. This makes both logical and physical sense. Earthquakes are dynamic processes, which means that the transition probabilities will not be stationary, violating the most fundamental assumption of a Markov Chain. Moreover, we implicitly use a first-order Markov Chain in which the current state only depends on the previous day. However, the more earlier earthquakes will also consume the accumulated stress in the tectonic plates, influencing and retarding much later earthquake events, so the Markov process should be a higher-order one. Particularly, the seismic activity should decay and its distribution will gradually shift to the dormant side.

15.4. Exercise

Exercise 15.1 For the island Gu-Nei-Ng-Dou, if it is a sunny day, then the next day has 75% chance of being sunny, 10% chance of being windy, and 15% of being rainy. Meanwhile, a windy day has 60%/25%/15% chance leading to a sunny/windy/rainy day, and a rainy day has 30%/20%/50% chance leading to a sunny/windy/rainy day. Find:

(a) the probability of getting a rainy day three days after a rainy day,

(b) the steady-state vector for all three types of weather,

Figure 15.1.: *Transport between reservoirs in Exercise 15.3.*

(c) the probability of getting at least one sunny day today and tomorrow if it was sunny yesterday.

Exercise 15.2 Go to https://ggweather.com/enso/oni.htm for the historical yearly ENSO status and construct a Markov Chain for the El Niño/La Niña/Neutral conditions.

Exercise 15.3 For the carbon cycle shown below (Figure 15.1), calculate the rate constant $k = \frac{\text{flux}}{\text{stock mass}}$ for each transport process, and verify that on average it takes about 213000 years for carbon in the ocean reservoir to be buried in the sediment by (a) absorbing Markov Chain, (b) considering the biosphere-atmosphere-ocean as a single reservoir and calculating the lifetime $\tau = \frac{1}{k} = \frac{\text{total stock mass}}{\text{flux}}$. Also, if the pathway to sediment is neglected, find the steady-state vector of the biosphere-atmosphere-ocean system and confirm they are already in a steady state as one would expect.

Exercise 15.4 Daniel is drunk in a bar. He wants to go to the train station so that he can go home. However, since he is drunk, he loses his sense of direction, and will move randomly along the green edge (shown in the map as Figure 15.2 below) once at a time. Assume that at any location, the chances of moving to all

Figure 15.2.: *The map for Exercise 15.4.*

other neighboring locations are equal. Nevertheless, once he arrives at the train station, he will stop wandering and take the train. How many moves does it take on average for Daniel to reach the train station?

Exercise 15.5 Attempt Project Euler Problem 84, preferably with programming.

Chapter 16

Matrix Factorization Methods

In this chapter, we are going to discuss some matrix factorization methods. We have introduced two of them previously: CR Factorization and QR Decomposition, in Chapters 6 and 7 respectively. We will first discuss two factorization methods for square matrices, *Cholesky* and *LU/LDU*, and subsequently move to the very crucial *Singular Value Decomposition* for non-square matrices. This eventually leads to the notions of *psuedoinverses* and *minimal solutions*. These decomposition methods have been widely applied in many fields of Earth Science to extract important patterns from data. Also, as Machine Learning gains popularity, Research in Earth Science starts to involve more matrix factorization techniques, which have been a main instrument in Machine Learning. Apart from this, a less visible usage of matrix factorization is embedded in the implementation of linear algebra packages in programming languages (e.g. LAPACK, for Fortran). Those matrix factorization methods enable faster and more stable computations of linear algebra problems, such as finding inverses or solving linear systems. Other potential applications include image processing and more.

16.1. Square Matrix Factorization

16.1.1. Cholesky Factorization

Cholesky Decomposition is for the special class of real symmetric and positive-definite matrices. We have talked about how a real symmetric matrix is positive-definite in Section 11.1.2. For such a matrix A, Cholesky Decomposition factorizes it into $U^T U$, where U is an upper-triangular real matrix, and U^T is hence a lower-triangular matrix (remember that upper/lower-triangular implies that non-zero entries are only present along or above/below the main diagonal). An example of Cholesky Factorization would be

$$\begin{aligned} A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 6 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \\ &= U^T U \end{aligned}$$

We can compute the Cholesky Decomposition of any positive-definite symmetric matrix step by step, first rewriting the $n \times n$ matrix A into the presumed factorized form of

$$A = U^T U = \begin{bmatrix} u_{11} & \vec{r}_1^T \\ \mathbf{0} & U_b \end{bmatrix}^T \begin{bmatrix} u_{11} & \vec{r}_1^T \\ \mathbf{0} & U_b \end{bmatrix} = \begin{bmatrix} u_{11} & \mathbf{0}^T \\ \vec{r}_1 & U_b^T \end{bmatrix} \begin{bmatrix} u_{11} & \vec{r}_1^T \\ \mathbf{0} & U_b \end{bmatrix} \tag{16.1}$$

as a 2×2 block matrix, where u_{11} will be the first diagonal entry of U, $\vec{r}_1$ is a column vector of length $n-1$ and U_b is a submatrix with size $(n-1) \times (n-1)$. The block matrix product (refer to Section 8.3.1) on R.H.S. gives

$$A = \begin{bmatrix} \alpha_{11} & \vec{a}_1^T \\ \vec{a}_1 & \tilde{A}_b \end{bmatrix} = \begin{bmatrix} u_{11}^2 & u_{11}\vec{r}_1^T \\ u_{11}\vec{r}_1 & \vec{r}_1\vec{r}_1^T + U_b^T U_b \end{bmatrix} = U^T U \tag{16.2}$$

where α_{11} is the first diagonal element of A, $\vec{a}_1$ and $\tilde{A}_b$ is also a column vector and a submatrix having the same shapes as $\vec{r}_1$ and U_b. Comparing the both sides, we have

$$\alpha_{11} = u_{11}^2 \tag{16.3a}$$
$$\vec{a}_1 = u_{11}\vec{r}_1 \tag{16.3b}$$
$$\tilde{A}_b = \vec{r}_1\vec{r}_1^T + U_b^T U_b \tag{16.3c}$$

and thus

$$u_{11} = \sqrt{\alpha_{11}} \tag{16.4a}$$
$$\vec{r}_1 = \frac{\vec{a}_1}{\sqrt{\alpha_{11}}} \tag{16.4b}$$
$$U_b^T U_b = \tilde{A}_b - \vec{r}_1\vec{r}_1^T \tag{16.4c}$$

By the relations above, we can determine u_{11}, and hence $\vec{r}_1$, retrieving the first row and column of U. Subsequently, the remaining block U_b is found by applying the same procedure on the $(n-1)\times(n-1)$ submatrix $U_{b=2}^T U_{b=2}$ which has been produced by the last relation (16.4c), and then invoking the method recursively to reduce the resulting block until the final entry is processed.

Definition 16.1.1 (Cholesky Factorization)**.** The Cholesky Factorization $U^T U$ of a real symmetric, positive-definite matrix A, is constructed by the recursive relations

$$u_{mm} = \sqrt{\alpha_{mm}} \tag{16.5a}$$
$$\vec{r}_m = \frac{\vec{a}_m}{\sqrt{\alpha_{mm}}} \tag{16.5b}$$
$$U_{b=m+1}^T U_{b=m+1} = \tilde{A}_{b=m+1} - \vec{r}_m\vec{r}_m^T \tag{16.5c}$$

where the subscript m implies we are at the m-th step, and

$$U_m^T U_m = \begin{bmatrix} \alpha_{mm} & \vec{a}_m^T \\ \vec{a}_m & \tilde{A}_{m+1} \end{bmatrix} = \begin{bmatrix} u_{mm}^2 & u_{mm}\vec{r}_m^T \\ u_{mm}\vec{r}_m & \vec{r}_m\vec{r}_m^T + U_{m+1}^T U_{m+1} \end{bmatrix} \tag{16.6}$$

The formulae are iterated over $U_{b=m+1}^T U_{b=m+1}$ acquired at the end of each step.

We require that at every step, α_{mm} and u_{mm} to be positive so that (16.5a) and (16.5b) make sense. It has been shown that matrices in the form of $U^T U$ will always be symmetric, and under this restriction, $U^T U$ will also be positive-definite according to the result in Exercise 11.5.[1]

Example 16.1.1. Perform Cholesky Factorization on the symmetric, positive-definite matrix

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

Solution. The first step results in

$$\begin{aligned} u_{11} &= \sqrt{\alpha_{11}} = \sqrt{4} = 2 \\ \vec{r}_1 &= \frac{1}{\sqrt{\alpha_{11}}}\vec{a}_1 = \frac{1}{\sqrt{4}}\begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ U_2^T U_2 = A_2 - \vec{r}_1 \vec{r}_1^T &= \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} \end{aligned}$$

So we know that

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

[1] This is because the diagonal elements u_{mm} cannot be zero and thus the upper-triangular U is invertible (why?) so that Exercise 11.5 can apply.

Similarly, the next iteration on $U_2^T U$ gives

$$\begin{aligned} u_{22} &= \sqrt{1} = 1 \\ \vec{r}_2 &= \frac{1}{\sqrt{1}} \begin{bmatrix} 2 \end{bmatrix} = \begin{bmatrix} 2 \end{bmatrix} \\ U_3^T U_3 = A_3 - \vec{r}_2 \vec{r}_2^T &= 5 - \begin{bmatrix} 2 \end{bmatrix} \begin{bmatrix} 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 \end{bmatrix} \end{aligned}$$

We still need to deal with the 1×1 matrix that remains. At the third step, we simply take

$$u_{33} = \sqrt{\alpha_{33}} = \sqrt{1} = 1$$

So the final expression of U is given by

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

□

► Short Exercise: Check if $A = U^T U$.[2]

As a side note, Cholesky Decomposition actually has a very interesting and useful side-effect of determining if a given symmetric matrix is positive-definite. If at any step, α_{mm} is not positive and Formula (16.5a) fails, then it means that the matrix cannot be positive-definite.[3]

16.1.2. LU/LDU Factorization

Another method, ***LU Factorization***, is quite similar to Cholesky Factorization. It decomposes any square matrix into one lower and one upper-triangular matrix,

L and U, but the original matrix does not need to be symmetric or positive-definite. The key to LU factorization is the method of Gaussian Elimination discussed in Sections 2.2.2 and 2.2.3. Recall that by Gaussian Elimination, we can always reduce a given matrix to an upper-triangular matrix (in row echelon form) during the forward phase, through a sequence of elementary row operations, or equivalently multiplying by the corresponding elementary matrices to the left. Denoting such operations by $E'_1, E'_2, \cdots, E'_m$ in a spirit similar to the proof of Theorem 2.2.10, we have $E'_m \cdots E'_2 E'_1 A = U$, where U is an upper-triangular matrix produced from the forward phase of Gaussian Elimination. Rearrangement gives

$$A = (E_1'^{-1} E_2'^{-1} \cdots E_m'^{-1}) U \tag{16.7}$$

If none of the E'_i involves row interchange, then they (as well as the inverses $E_i'^{-1}$) would be all lower-triangular matrices (particularly, the multiplicative constant for adding an upper row to a lower row will appear below the main

[2]It is simply computing

$$U^T U = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}^T \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 5 \end{bmatrix} = A$$

[3]Heuristically, towards the m-th step, we have

$$U_m^T U_m = \begin{bmatrix} \alpha_{mm} & \cdots \\ \vdots & \ddots \end{bmatrix}$$

and if $\alpha_{mm} \leq 0$ and we set $\vec{w}_m = (1, 0, \ldots)^T$ where only the first component is 1 and the others are 0, we have

$$\vec{w}_m^T (U_m^T U_m) \vec{w}_m = \begin{bmatrix} 1 & 0 & \cdots \end{bmatrix} \begin{bmatrix} \alpha_{mm} & \cdots \\ \vdots & \ddots \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \alpha_{mm} \leq 0$$

which shows that $U_m^T U_m$ is not positive-definite and it is impossible.

diagonal in that elementary matrix) as required by the forward stage of Gaussian Elimination (meanwhile, if there is row interchange, then E'_i will not be a lower triangular matrix)[4]. As a result, their product $E'^{-1}_1 E'^{-1}_2 \cdots E'^{-1}_m$ will also be a lower triangular matrix L (see Exercise 16.1):

$$L = E'^{-1}_1 E'^{-1}_2 \cdots E'^{-1}_m \tag{16.8}$$

and the LU Factorization $A = LU$ is consequentially derived.

Theorem 16.1.2 (LU Factorization)**.** The LU Factorization of a square matrix A is possible if it can be reduced to a row echelon form U, which will be an upper-triangular matrix, by forward Gaussian Elimination without any row interchange. The steps and elementary matrices used to produce U can be grouped together, and inverted to give a lower triangular matrix L correspondingly. The LU Factorization will then simply be

$$A = LU \tag{16.9}$$

Example 16.1.2. Compute a LU Factorization for

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

Solution. By Gaussian Elimination, we have

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 4 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 4 \end{bmatrix} \qquad E'_1 = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\to \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix} \qquad E'_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}$$

[4]If rows p and q are interchanged, and let's say $p < q$, then above the main diagonal $E_{pq} = 1$.

$$\rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad E_3' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}$$

Therefore, we can take

$$\begin{aligned} U &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \\ L &= E_1'^{-1} E_2'^{-1} E_3'^{-1} \\ &= \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix} \end{aligned}$$

This sequence amounts to the series of elementary row operations in the forward reduction phase. Doing them from the right to the left, it is easy to arrive at

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

□

Note that LU Factorization is not always unique. However, one can derive ***LDU Factorization*** that is unique whenever an LU decomposition of the matrix exists. Here, D is an extra diagonal matrix while all the diagonal entries L and U are now 1. Using the above example, we observe that

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

where we factor out a diagonal matrix from L along each column to its right, to force all the entries along the main diagonal to be 1. If needed, we can also do the same procedure on U which produces a diagonal matrix to the left instead. So the corresponding LDU Factorization in the previous example is

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

There is another variant of LU Factorization which is called ***PLU Factorization***, applicable to any square matrix A. Since the only obstacle to obtaining a LU Decomposition is the presence of row interchanging operations (sometimes necessary, sometimes due to pivoting for numerical stability), the idea is to complete them right from the start so that LU Factorization becomes possible. After obtaining the LU Factorization, we collect all the elementary row matrices involved in the initial row interchanges as a matrix P appended to the left. In practice, we will carry out the forward step of Gaussian Elimination as usual and mark all the row interchanging operations down. The matrix P will simply be the product of all the row interchanging elementary matrices, as swapping rows does not really modify the actual entries and only affects the row indices so that we can move all of them to the left in the end. Whenever we have to move a row interchanging elementary matrix across the other two types of elementary matrices, we can appropriately relocate the addition/multiplication constant c in one of the two rows to another. For example, if we have $R_3 + cR_1 \to R_3$ and $R_2 \leftrightarrow R_3$, then

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

and if $cR_2 \to R_2$ and $R_2 \leftrightarrow R_3$ this time, we have

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

The same "bubble" logic is similarly applicable to any other pair of indices.

Example 16.1.3. Find a PLU Factorization for

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

Solution. We will apply Gaussian Elimination as usual and do the bookmarking for any row interchange.

$$\begin{aligned}
\begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 3 \\ 1 & 3 & 0 \end{bmatrix} &\to \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & -1 \\ 1 & 3 & 0 \end{bmatrix} & E_1' &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&\to \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & -1 \\ 0 & 2 & -2 \end{bmatrix} & E_2' &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \\
&\to \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & -2 \\ 0 & 0 & -1 \end{bmatrix} & E_3' &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \\
&\to \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix} & E_4' &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&\to \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} & E_5' &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}
\end{aligned}$$

So

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

and we move E_3' all the way to the left and modify $E_1'^{-1}$ and $E_2'^{-1}$ accordingly, to acquire

$$
\begin{aligned}
PL &= E_1'^{-1}E_2'^{-1}E_3'^{-1}E_4'^{-1}E_5'^{-1} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 0 & -1 \end{bmatrix}
\end{aligned}
$$

It is not hard to check that

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

Notice that PLU Decomposition is not unique as there are many possible ways of doing row exchanges. □

► Short Exercise: What will the PLU Decomposition in the above example look like if we incorporate the part of "D" as in the LDU variant?[5]

16.1.3. Solving Linear Systems with LU Decomposition

As mentioned at the start of this chapter, matrix factorization greatly helps when it comes to electronic calculation. For instance, LU factorization is widely

applied to solving linear systems in computers, preferred over direct Gaussian Elimination or using inverses (we will discuss the reasons at the end of the chapter). Basically, it is a two-step procedure. For a linear system $A\vec{x} = \vec{h}$, if we have the LU Decomposition of the matrix A, then it can be rewritten as

$$LU\vec{x} = \vec{h} \tag{16.10}$$

Subsequently, we let $U\vec{x} = \vec{y}$, and solve

$$L\vec{y} = \vec{h} \tag{16.11}$$

in the first stage, and then come back to solve

$$U\vec{x} = \vec{y} \tag{16.12}$$

in the second stage. Due to the triangular nature of L and U, we can do a forward and backward substitution respectively to solve the two systems with relative ease.

Example 16.1.4. Solve the linear system $A\vec{x} = \vec{h}$, where

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

via LU Factorization.

Solution. We note that one possible LU factorization of A is

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

[5] It will look like

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

which is left to the readers to verify. We can solve the equivalent problem $LU\vec{x} = \vec{h}$, where we set $U\vec{x} = \vec{y}$ and thus $L\vec{y} = \vec{h}$. The first stage involve solving $L\vec{y} = \vec{h}$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$$

By forward substitution from the top to bottom, we have $y_1 = 3$, $2y_2 = 2 \rightarrow y_2 = 1$, and $y_1 + y_2 + 3y_3 = (3) + (1) + 3y_3 = 4 \rightarrow y_3 = 0$, so $\vec{y} = (3, 1, 0)^T$. Similarly, we do a backward substitution over $U\vec{x} = \vec{y}$ in the second stage:

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}$$

which gives $x_3 = 0$, $x_2 + x_3 = x_2 + (0) = 1 \rightarrow x_2 = 1$ and $x_1 + x_2 = x_1 + (1) = 3 \rightarrow x_1 = 2$, and thus the final answer is $\vec{x} = (2, 1, 0)^T$. □

Remarks

To solve a family of linear systems $A\vec{x} = \vec{h}^{(j)}$, where $\vec{h}^{(j)}$ are many different column vectors, can be very time-consuming if we use Gaussian Elimination every time. LU Decomposition extracts and encapsulates the information from Gaussian Elimination, and saves the effort needed to compute Gaussian Elimination for different $\vec{h}^{(j)}$ every time since we will only have to do the forward and backward substitution (more details in the answer to this Math Stackexchange post (266355)). While it is also possible to achieve similar effects by utilizing the inverse A^{-1}, calculating the inverse for large systems can be numerically unstable and expensive (which has been discussed in Section 3.4, and more thoroughly at https://gregorygundersen.com/blog/2020/12/09/matrix-inversion).

16.2. Singular Value Decomposition (SVD)

16.2.1. Mathematical Ideas of SVD and its Computation

The previous LU Factorization can also be applied to a non-square matrix. However, there is another very important factorization method called the ***Singular Value Decomposition (SVD)*** that is more useful for signal processing, data compression, and feature extraction. SVD proposes that any matrix A, no matter square or non-square, can be rewritten into the product of three matrices:

$$A = U\Sigma V^* \tag{16.13}$$

where A is a given $m \times n$ matrix, U and V are two unitary square matrices of the shape $m \times m$ and $n \times n$ respectively, and Σ is an $m \times n$ diagonal[6] matrix. The principle of SVD essentially revolves around a generalized orthogonal/unitary change of coordinates for a matrix (or linear transformation). To see this, compare (16.13) to (7.12) in Properties 7.2.1 about coordinate transformation for a matrix in general, and set $[T]_{\beta'}^{\gamma'} = \Sigma$, $Q_{\gamma'}^{\gamma} = U$, $[T]_{\beta}^{\gamma} = A$ and $P_{\beta'}^{\beta} = V$. Subsequently, Equation (7.12) becomes

$$\Sigma = U^{-1}AV \tag{16.14}$$

Since U and V are both unitary (Definition 10.4.1, $U^{-1} = U^*$, $UU^* = I = VV^*$), it can be expressed in two ways.

$$\begin{aligned} \Sigma &= U^*AV \\ \text{as well as} \quad U\Sigma V^* &= UU^*AVV^* \\ &= (UU^*)A(VV^*) = (I)A(I) = A \end{aligned}$$

[6] In this chapter, we generalize and use the word "diagonal" for non-square matrices too, in the sense that non-zero entries only occurs when the row and column indices are equal just like for square matrices.

So we arrive at (16.13) again. This means that the central idea of SVD is to find two orthonormal (unitary) bases for the input/output vector spaces $\mathcal{V}$ and $\mathcal{W}$ so that after changing the respective coordinate systems to these two bases, the linear transformation will then have a diagonal representation. As a result, the first step is to show for any linear transformation, we can always find some suitable orthonormal bases for $\mathcal{V}$ and $\mathcal{W}$ so that any basis vector $\vec{v}^{(j)} \in \mathcal{V}$ from the starting vector space is mapped to a constant multiple of one and only one corresponding basis vector $\vec{u}^{(j)} \in \mathcal{W}$ in the other vector space.

Theorem 16.2.1 (Singular Value Theorem for Linear Transformations)**.** For two inner product spaces $\mathcal{V}$ and $\mathcal{W}$ of finite dimensions n and m respectively, as well as given a linear transformation $T : \mathcal{V} \to \mathcal{W}$ between them, there exist orthonormal bases $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}\}$ for $\mathcal{V}$ and $\gamma = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(m)}\}$ for $\mathcal{W}$ so that

$$T(\vec{v}^{(j)}) = \sigma_j \vec{u}^{(j)} \tag{16.15}$$

for $1 \leq j \leq \min(m, n)$ and $T(\vec{v}^{(j)}) = \mathbf{0}$ when $j > \min(m, n)$. The factors σ_j are known as the ***singular values*** of T and are real. The required basis vectors $\vec{v}^{(j)}$ are the eigenvectors of $T^\dagger T$, the eigenvalues of which are equal to σ_j^2. We will take the values for σ_j up to $j = \min(m, n)$ and arrange them in descending order as $\sigma_1 \geq \sigma_2 \geq \sigma_{\min(m,n)} \geq 0$. The singular values σ_j are uniquely determined by a given T.

Proof. First, note that $T^\dagger T$ is positive-semidefinite with respect to any inner product.[7] $T^\dagger T$ is also clearly self-adjoint and automatically Hermitian when the

[7] Consider the Hermitian form $\langle \vec{v}, T^\dagger T \vec{v} \rangle$ appropriate for the inner product. Then by Definition 12.2.1 and Properties 12.1.3, we have

$$\begin{aligned} \langle \vec{v}, T^\dagger T(\vec{v}) \rangle &= \langle T(\vec{v}), T(\vec{v}) \rangle \\ &= \|T(\vec{v})\|^2 \geq 0 \end{aligned}$$

which shows that $T^\dagger T$ is positive-semidefinite.

inner product spaces are finite-dimensional so by Properties 12.2.6 and the Spectral Theorem 12.3.4 there exists an orthonormal basis $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(n)}\}$ for $\mathcal{V}$, which are the eigenvectors of $T^\dagger T$ with corresponding eigenvalues of $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\min(m,n)} \geq 0$, and if $n > m$ then $\lambda_j = 0$ for $j > m$.[8] By Properties 12.2.5, these λ_j will be real. Now we set $\vec{u}^{(j)} = \frac{1}{\sigma_j} T(\vec{v}^{(j)})$ where the singular values $\sigma_j = \sqrt{\lambda_j}$ for $1 \leq j \leq \min(m,n)$ are then real too, and the remaining task is to show that $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(\min(m,n))}\}$ is also an orthonormal subset of $\mathcal{W}$. For any pair of indices $1 \leq i, j \leq \min(m,n)$, we have

$$\begin{aligned}
\langle \vec{u}^{(i)}, \vec{u}^{(j)} \rangle &= \langle \frac{1}{\sigma_i} T(\vec{v}^{(i)}), \frac{1}{\sigma_j} T(\vec{v}^{(j)}) \rangle \\
&= \frac{1}{\sigma_i \sigma_j} \langle \vec{v}^{(i)}, T^\dagger T(\vec{v}^{(j)}) \rangle && \text{(Definition 12.2.1)} \\
&= \frac{1}{\sigma_i \sigma_j} \langle \vec{v}^{(i)}, \lambda_j \vec{v}^{(j)} \rangle \\
&= \frac{\sigma_j^2}{\sigma_i \sigma_j} \langle \vec{v}^{(i)}, \vec{v}^{(j)} \rangle \\
&= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} && (\vec{v}^{(j)} \text{ are set to be orthonormal})
\end{aligned} \tag{16.16}$$

which shows that $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(\min(m,n))}\}$ are orthonormal. By part (c) of Properties 6.2.7 and the Gram-Schmidt process (Definition 12.3.3), we can complete an orthonormal basis $\gamma = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(m)}\}$ for $\mathcal{W}$ if $m > n$. As we have set $\vec{u}^{(j)} = \frac{1}{\sigma_j} T(\vec{v}^{(j)})$, we immediately have $T(\vec{v}^{(j)}) = \sigma_j \vec{u}^{(j)}$ for $1 \leq j \leq \min(m,n)$. And if $j > \min(m,n)$, we have $T(\vec{v}^{(j)}) = \mathbf{0}$ because $T^\dagger T(\vec{v}^{(j)}) = \lambda_j \vec{v}^{(j)} = \mathbf{0}$ as $\lambda_j = 0$ whenever $j > \min(m,n)$.[9] The last part is to show the uniqueness of the singular values σ_j. If the premise holds such that

[8] By Properties 13.1.2 we know that if $n > m$ then the maximum rank of $T^\dagger T$ will be capped by m, so the nullity of $T^\dagger T$ and thus the number of zero eigenvalues will be at least $n - m$. (Rank-Nullity Theorem 7.1.9)

$T(\vec{v}^{(j)}) = \sigma_j \vec{u}^{(j)}$ for $1 \leq j \leq \min(m,n)$ and $T(\vec{v}^{(j)}) = \mathbf{0}$ when $j > \min(m,n)$, then

$$\begin{aligned}\langle T^\dagger(\vec{u}^{(i)}), \vec{v}^{(j)}\rangle &= \langle \vec{u}^{(i)}, T(\vec{v}^{(j)})\rangle \\ &= \langle \vec{u}^{(i)}, \sigma_j \vec{u}^{(j)}\rangle \\ &= \begin{cases} \sigma_i & \text{if } 1 \leq i = j \leq \min(m,n) \\ 0 & \text{otherwise} \end{cases}\end{aligned} \tag{16.17}$$

but since $\vec{v}^{(j)}$ forms an orthonormal basis, by part (d) of Spectral Theorem 12.3.4 we can rewrite $T^\dagger(\vec{u}^{(i)})$, decomposed as the sum of orthogonal projections into $\vec{v}^{(j)}$:

$$\begin{aligned}T^\dagger(\vec{u}^{(i)}) &= \sum_{j=1}^{n} \langle T^\dagger(\vec{u}^{(i)}), \vec{v}^{(j)}\rangle \vec{v}^{(j)} \\ &= \begin{cases} \sigma_i \vec{v}^{(i)} & \text{if } 1 \leq i = j \leq \min(m,n) \\ 0 & \text{otherwise} \end{cases}\end{aligned} \tag{16.18}$$

Therefore, for $1 \leq j \leq \min(m,n)$,

$$T^\dagger T(\vec{v}^{(j)}) = T^\dagger(\sigma_j \vec{u}^{(j)}) = \sigma_j T^\dagger(\vec{u}^{(j)}) = \sigma_j^2 \vec{v}^{(j)} \tag{16.19}$$

so $\vec{v}^{(j)}$ must be the eigenvectors of $T^\dagger T$ whose eigenvalues are σ_j^2. We leave the case of $j > \min(m,n)$, where the singular values will be zero, to the readers. □

With the above theorem for a general linear transformation established, we can now show the equivalent SVD statement for matrices. We will discuss the

[9] Again, as in Footnote 7,

$$\begin{aligned}\langle \vec{v}^{(j)}, T^\dagger T(\vec{v}^{(j)})\rangle &= \langle T(\vec{v}^{(j)}), T(\vec{v}^{(j)})\rangle \\ &= \left\| T(\vec{v}^{(j)}) \right\|^2\end{aligned}$$

but $\langle \vec{v}^{(j)}, T^\dagger T(\vec{v}^{(j)})\rangle = \langle (\vec{v}^{(j)}), \mathbf{0}\rangle = 0$ so $\left\| T(\vec{v}^{(j)}) \right\|^2 = 0$ and by positivity in Definition 12.1.1, $T(\vec{v}^{(j)}) = \mathbf{0}$.

simplest case where the inner product is just the standard (complex) dot product and thus $A^\dagger = A^*$ and its singular values will be the eigenvalues of A^*A.

Theorem 16.2.2 (Singular Value Decomposition)**.** For any $m \times n$ matrix A, its Singular Value Decomposition is

$$A = U\Sigma V^* \tag{16.20}$$

where U and V are $m \times m$ and $n \times n$ unitary matrices respectively, Σ is an $m \times n$ diagonal matrix. The matrix V is constructed by

$$V = \left[\hat{v}^{(1)}|\hat{v}^{(2)}|\cdots|\hat{v}^{(n)}\right] \tag{16.21}$$

where the unit column vectors $\hat{v}^{(1)}, \hat{v}^{(2)}, \ldots, \hat{v}^{(n)}$ are the orthonormal eigenvectors of the Hermitian matrix product A^*A (Properties 12.2.6) called ***right singular vectors***, ordered by decreasing, real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Σ then consists of the first $k = \min(m, n)$ singular values σ_j along the main diagonal, which are the square roots of the eigenvalues of A^*A (also in decreasing order) of the corresponding column vectors $\hat{v}^{(j)}$ in the unitary matrix V, and has a value of zero elsewhere. Finally, U is made up of the orthonormal column vectors that satisfy

$$\hat{u}^{(j)} = \frac{1}{\sigma_j} A\hat{v}^{(j)} \tag{16.22}$$

for any j that is associated to a non-zero singular value $\sigma_j \neq 0$. These vectors $\hat{u}^{(j)}$ are called ***left singular vectors***. When there are not enough column vectors to construct U by this relation, we can use Gram-Schmidt Orthogonalization (or other feasible methods) to create orthonormal vectors that extend the basis of $\hat{u}^{(j)}$ up to $j = m$.

This is just Theorem 16.2.1 applied on Properties 7.2.1 in another form, where now we update the basis subscripts/superscripts to match the notation appropriately and have $[T]_\beta^\gamma = \Sigma$, $Q_\gamma^S = U$, $[T]_S = A$ and $P_\beta^S = V$ (see the flowchart

Figure 16.1.: Same as Figure 7.2 but adapted to the context of SVD.

of Figure 16.1 below) where S denotes the standard bases for both $\mathbb{R}^n/\mathbb{C}^n$ and $\mathbb{R}^m/\mathbb{C}^m$ (but can be replaced by any other arbitrary basis when needed), so

$$\begin{aligned} A = [T]_S &= Q_\gamma^S [T]_\beta^\gamma (P_\beta^S)^{-1} \\ &= U\Sigma V^{-1} = U\Sigma V^* \qquad\qquad ((U \text{ and}) \ V \text{ is unitary}) \end{aligned}$$

We emphasize that we allow complex eigenvectors and eigenvalues but we still use the word "orthonormal" when referring to unitary matrices for convenience in this chapter. And if we work over reals, V^* is simply reduced to V^T.

Example 16.2.1. Find the SVD of the following 2×3 real matrix.

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

Solution. First, we compute its singular values, by considering the normal matrix:

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

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

which can be checked to have the eigenvalues of $\lambda = 6, 1, 0$, and hence A has a decreasing sequence of singular values of $\sigma = \sqrt{6}, 1, 0$. The last singular value of zero is expected by Theorem 16.2.1 and will be discarded. We note that the corresponding orthonormal eigenvectors for $A^T A$ are $(\frac{\sqrt{5}}{\sqrt{6}}, \frac{\sqrt{2}}{\sqrt{15}}, \frac{1}{\sqrt{30}})^T$, $(0, -\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})^T$, and $(-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}})^T$ and the readers can verify them. So we have

$$V = \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{6}} \\ \frac{\sqrt{2}}{\sqrt{15}} & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The two column vectors for U are found by

$$\begin{aligned} \hat{u}^{(1)} &= \frac{1}{\sigma_1} A \hat{v}^{(1)} \\ &= \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} \\ \frac{\sqrt{2}}{\sqrt{15}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \hat{u}^{(2)} &= \frac{1}{\sigma_2} A \hat{v}^{(2)} \\ &= \frac{1}{1} \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{bmatrix} \end{aligned}$$

Therefore, we conclude that

$$
\begin{aligned}
U &= \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix} \\
A &= U\Sigma V^T \\
&= \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{15}} & \frac{1}{\sqrt{30}} \\ 0 & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} 5 & 2 & 1 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}
\end{aligned}
$$

□

Example 16.2.2. Carry out SVD for the 3×2 matrix below.

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

Solution. We leave it to the readers to check that the two orthonormal eigenvectors of

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

are

$$\hat{v}^{(1)} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad \hat{v}^{(2)} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

for $\lambda_1 = 3$, $\lambda_2 = 1$, so $\sigma_1 = \sqrt{3}$, $\sigma_2 = 1$, and

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$\hat{u}^{(1)}$, $\hat{u}^{(2)}$ are then given by

$$\hat{u}^{(1)} = \frac{1}{\sigma_1} A\hat{v}^{(1)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

$$\hat{u}^{(2)} = \frac{1}{\sigma_2} A\hat{v}^{(2)} = \frac{1}{1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

We need another unit vector $\hat{u}^{(3)}$ that is orthogonal to both $\hat{u}^{(1)}$ and $\hat{u}^{(2)}$. A quick way for this special case of three-dimensional vectors is to find the cross product $\hat{u}^{(1)} \times \hat{u}^{(2)}$, that is

$$\hat{u}^{(3)} = \hat{u}^{(1)} \times \hat{u}^{(2)} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{vmatrix} = -\frac{1}{\sqrt{3}}\hat{\imath} + \frac{1}{\sqrt{3}}\hat{\jmath} + \frac{1}{\sqrt{3}}\hat{k} = \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$$

As a result,

$$\begin{aligned} U &= \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix} \\ A &= U\Sigma V^T \\ &= \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

□

16.2.2. Geometric Perspective of SVD

As discussed above, SVD involves two orthogonal/unitary coordinate transformations indicated by the matrices U and V, one before and one after the mapping via the diagonal matrix Σ. In the case where A happens to be a square matrix so that $m = n$ and if the vector spaces $\mathcal{V}$ and $\mathcal{W}$ have the same dimension and coincide, then from a geometric viewpoint, the action of A can be separated into three steps: first a rotation/reflection, then a stretching/compression, finally followed by another rotation/reflection. We illustrate this using a small example. Consider

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

which acts as a shear. It is not hard to obtain its SVD, which is noted below but the readers can try to derive it as a small exercise.

$$A = U\Sigma V^T = \begin{bmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

First, we have V^T which constitutes a positive (counter-clockwise) rotation of the coordinate system by a degree of $\theta = \tan^{-1}(2) \approx 63.4°$ (see Section 10.1.2). Then, the image is distorted by stretching/compressing along the intermediate x/y-axis respectively by a factor of 2 through Σ. Finally, we carry out another coordinate frame rotation by a degree of $\theta = \tan^{-1}(1/2) \approx 26.6°$ in the negative (clockwise) direction. Notice that the passive rotations of the coordinates axes are equivalent to active rotations of vectors in the opposite sense. The flow of the entire procedure is shown in Figure 16.2 below.

16.2.3. Compact SVD

During the derivation of Theorem 16.2.1, we have shown that if $n > m$ then there will be at least $n - m$ zero eigenvalues for $T^\dagger T$ (or $A^* A$) in Footnote 8. Of

Figure 16.2.: *The geometric interpretation of the* 2×2 *SVD example. Red/Blue arrows represent the intermediate coordinate axes and green arrows mark the distortion of two given points indicated by the standard unit vectors initially.*

course, they are not included in the Σ matrix as the main diagonal only has the length of m so there will be m singular values only. However, we can tighten the upper bound on the number of non-zero eigenvalues and therefore that of singular values. By the same Properties 13.1.2 used in Footnote 8, we can know that if A has a rank of $k \leq \min(m,n)$, then A^*A will also have a rank of k and

thus k non-zero singular values.[10] Hence the original SVD of a rank k matrix

$$
\begin{aligned}
A &= U\Sigma V^T \\
&= \left[\hat{u}^{(1)}|\hat{u}^{(2)}|\cdots|\hat{u}^{(k)}|\cdots|\hat{u}^{(m)}\right]
\begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 & 0 & \cdots \\
0 & \sigma_2 & & & & \\
\vdots & & \ddots & & & \\
0 & & & \sigma_k & & \\
0 & & & & 0 & \\
\vdots & & & & & \ddots
\end{bmatrix}
\begin{bmatrix}
\hat{v}^{(1)*} \\ \hline \hat{v}^{(2)*} \\ \hline \vdots \\ \hline \hat{v}^{(k)*} \\ \hline \vdots \\ \hline \hat{v}^{(n)*}
\end{bmatrix}
\end{aligned}
\tag{16.23}
$$

can be reduced to the so-called ***Truncated/Compact SVD***, by only keeping entries associated with up to the first k non-zero singular values, resulting in

$$
A = U_k\Sigma_k V_k^* = \left[\hat{u}^{(1)}|\hat{u}^{(2)}|\cdots|\hat{u}^{(k)}\right]
\begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & & \\
\vdots & & \ddots & \\
0 & & & \sigma_k
\end{bmatrix}
\begin{bmatrix}
\hat{v}^{(1)*} \\ \hline \hat{v}^{(2)*} \\ \hline \vdots \\ \hline \hat{v}^{(k)*}
\end{bmatrix}
\tag{16.24}
$$

where U_k, V_k come from retaining only the first k columns of U and V, and Σ_k is a diagonal $k \times k$ square matrix that preserves the k non-zero singular values. From this point onwards, we mean the Compact SVD instead of the original one whenever we refer to SVD unless otherwise specified. If the singular values are distinct, the compact SVD of a matrix will be *unique up to a (complex) sign difference*[11] for the singular vectors because the right singular vectors are designated to be the eigenvectors of A^*A and we have required them to be arranged in the order of decreasing singular values. It can also be expanded in a

[10] The rank of k means that the nullity of A^*A is $n-k$ by the Rank-nullity Theorem 6.3.9 and thus it has $n-k$ zero eigenvalues and thus k non-zero eigenvalues (we implicitly have used Properties 10.2.4, adapted for complex matrices, can you see how?).

[11] which means that an $e^{i\theta}$ factor can be added to any singular vector.

fashion similar to Equation (10.8b), which gives

$$\left[\hat{u}^{(1)}|\hat{u}^{(2)}|\cdots|\hat{u}^{(k)}\right]\begin{bmatrix}\sigma_1 & 0 & \cdots & 0\\ 0 & \sigma_2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_k\end{bmatrix}\begin{bmatrix}\hat{v}^{(1)*}\\ \hline \hat{v}^{(2)*}\\ \hline \vdots \\ \hline \hat{v}^{(k)*}\end{bmatrix}$$
$$= \sigma_1\hat{u}^{(1)}\hat{v}^{(1)*} + \sigma_2\hat{u}^{(2)}\hat{v}^{(2)*} + \cdots + \sigma_k\hat{u}^{(k)}\hat{v}^{(k)*} \tag{16.25}$$

Example 16.2.3. Find the (Compact) SVD of the following matrix.

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

Solution. It is not hard to observe that the third row is the sum of the first two linearly independent rows so the rank of A is 2 (which can also be inferred from directly doing the Gaussian Elimination), so A^TA also has a rank of 2 and there will be two non-zero singular values for A. Now we compute the eigenvectors of A^TA as usual, where the results are simply noted below:

$$A^TA = \begin{bmatrix}6 & 0 & 3 & 3\\ 0 & 2 & -1 & 1\\ 3 & -1 & 2 & 1\\ 3 & 1 & 1 & 2\end{bmatrix}$$

and

$$\vec{v}^{(1)} = \begin{bmatrix}\frac{2}{\sqrt{6}}\\ 0\\ \frac{1}{\sqrt{6}}\\ \frac{1}{\sqrt{6}}\end{bmatrix} \qquad \text{for } \lambda_1 = 9$$

$$\vec{v}^{(2)} = \begin{bmatrix} 0 \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} \qquad \text{for } \lambda_2 = 3$$

and $\vec{v}^{(3)} = (-\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, 0)^T$, $\vec{v}^{(4)} = (-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}})^T$ for $\lambda_3 = \lambda_4 = 0$. We now only need the first two of them for the SVD, so

$$V = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} 3 & 0 \\ 0 & \sqrt{3} \end{bmatrix}$$

as $\sigma_1 = \sqrt{\lambda_1} = \sqrt{9} = 3$ and $\sigma_2 = \sqrt{\lambda_2} = \sqrt{3}$. Then we have

$$\begin{aligned} \vec{u}^{(1)} &= \frac{1}{\sigma_1} A \hat{v}^{(1)} \\ &= \frac{1}{3} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} \\ 0 \\ \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \vec{u}^{(2)} &= \frac{1}{\sigma_2} A \hat{v}^{(2)} \\ &= \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \end{aligned}$$

So

$$U = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{6}} & 0 \end{bmatrix}$$

and it is not hard to check that

$$A = U\Sigma V^*$$

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{6}} & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & \sqrt{3} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

□

16.2.4. Relation between SVD and PCA

We will end this section by talking about how SVD can be used to carry out PCA, introduced in Section 11.2.2. Given a real data matrix $[X']$, PCA requires us to compute the eigenvalues and orthonormal eigenvectors of the covariance matrix $Q = \frac{1}{m-1}[X']^T[X']$ where m is the number of data. If we know that the SVD of the data matrix is $[X'] = U\Sigma V^T$, then we have

$$\begin{aligned} Q &= \frac{1}{m-1}[X']^T[X'] \\ &= \frac{1}{m-1}(U\Sigma V^T)^T U\Sigma V^T \\ &= \frac{1}{m-1}V\Sigma U^T U\Sigma V^T && (\Sigma \text{ is diagonal}) \\ &= \frac{1}{m-1}V\Sigma^2 V^T && (U^T U = I \text{ as } U \text{ is orthogonal,} \\ & && \text{Properties 10.1.2}) \\ &= V(\frac{\Sigma^2}{m-1})V^T && (16.26) \end{aligned}$$

By comparing this to (10.6) in Definition 10.2.1, we immediately identify that the orthogonal matrix $P = V$ contains the orthonormal principal directions in columns, which coincide with the right singular vectors of $[X']$. The variances of the PCs are related to the singular values by

$$\lambda_j = \frac{\sigma_j^2}{m-1} \tag{16.27}$$

as $D = \Sigma^2/(m-1)$. In addition, as suggested by (11.35) below Theorem 11.2.8, the new PC coordinates are given

$$\vec{u}_i = [\mathbf{e}]^T \vec{x}_i = V^T \vec{x}_i \tag{16.28}$$

More generally, considering the entire data matrix $[X'] = [\vec{x}^{(1)}|\vec{x}^{(2)}|\cdots|\vec{x}^{(n)}]$ where each row represents a data point, we can compute the PCs for all data by

$$\begin{aligned} V^T X^T &= V^T [\vec{x}^{(1)}|\vec{x}^{(2)}|\cdots|\vec{x}^{(n)}]^T = V^T \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_m^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_m^{(2)} \\ \vdots & \vdots & & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_m^{(n)} \end{bmatrix} \\ &= V^T (U\Sigma V^T)^T \\ &= V^T V \Sigma U^T = \Sigma U^T \end{aligned} \tag{16.29}$$

and now each column of ΣU^T, or equivalently each row of $U\Sigma$, contains the PC coordinates of the corresponding data point. The readers are referred to this Statistics Stackexchange post (134282) for a more in-depth explanation.

16.3. Pseudoinverses

16.3.1. Definition of Pseudoinverses and Inference from SVD

The Singular Value Theorem 16.2.1 shows that a linear transformation can be described by Equation (16.15) between the orthonormal bases for $\mathcal{V}$ and $\mathcal{W}$. Particularly, for any non-zero singular value $\sigma_j > 0$, Equation (16.15) can be inverted as

$$T^+(\vec{u}^{(j)}) = \frac{1}{\sigma_j}\vec{v}^{(j)} \tag{16.30}$$

where now T^+ reverts the mapping by T from each of $\gamma = \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(k)}\}$ back to each of $\beta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}\}$ where there are k non-zero singular values. Notice that the subspaces generated by β and γ are the kernel's complement and the range, $\mathcal{N}(T)^\perp$ and $\mathcal{R}(T)$ respectively. Motivated by all of these, we may want to construct a ***"pseudoinverse"*** $T^+ : \mathcal{W} \to \mathcal{V}$ out of T even though the matrix representation of T may not have a square shape to be invertible in the usual sense: we only need Equation (16.30) to hold from $\mathcal{R}(T)$ to $\mathcal{N}(T)^\perp$ and this will be good enough to capture the invertibility between just these two subspaces. Both $\mathcal{R}(T)$ and $\mathcal{N}(T)^\perp$ have the same dimension[12] and there will always be an invertible linear transformation L between them by Properties 7.1.16. We require that this L coincides with T as long as the input vectors are in the kernel's complement $\mathcal{N}(T)^\perp$ and hence in the backward direction T^+ coincides with L^{-1} over the range $\mathcal{R}(T)$ as well. For other vectors in the complement of the range $\mathcal{R}(T)^\perp \subseteq \mathcal{W}$, we effectively ignore them by requiring that T^+ maps all of them into the zero vector. From these, we can properly define the pseudoinverse of a linear transformation as below.

[12] Let T has a rank of k so that $\dim(\mathcal{R}(T)) = k$, then by the Rank-Nullity Theorem 7.1.9, $\dim(\mathcal{N}(T)) = n - k$ and $\dim(\mathcal{N}(T)^\perp) = n - (n - k) = k$.

Definition 16.3.1 (Pseudoinverse of a Linear Transformation)**.** For two inner product spaces $\mathcal{V}$ and $\mathcal{W}$ of finite dimensions n and m respectively, and given a linear transformation $T\colon \mathcal{V} \to \mathcal{W}$ between them, let $L\colon \mathcal{N}(T)^\perp \to \mathcal{R}(T)$ be the invertible linear transformation between the two subspaces $\mathcal{N}(T)^\perp$ and $\mathcal{R}(T)$ such that

$$L(\vec{v}) = T(\vec{v}) \qquad \text{for every } \vec{v} \in \mathcal{N}(T)^\perp \tag{16.31}$$

Then the *(Moore-Penrose) pseudoinverse* $T^+ : \mathcal{W} \to \mathcal{V}$ is defined by

$$T^+(\vec{w}) = \begin{cases} L^{-1}(\vec{w}) & \text{for } \vec{w} \in \mathcal{R}(T) \\ \mathbf{0} & \text{for } \vec{w} \in \mathcal{R}(T)^\perp \end{cases} \tag{16.32}$$

Hence, the pseudoinverse always exists for any T by construction, regardless of whether T is invertible or not. Further, if T is invertible, then we simply have $T^+ = T^{-1}$ as $\mathcal{N}(T)^\perp = \mathcal{V}$ and $\mathcal{R}(T) = \mathcal{W}$. Finally, the Moore-Penrose pseudoinverse, as a linear transformation, is also unique. Since the Singular Value Theorem can be translated to SVD for matrices, we may like to know if there will be a matrix version of the pseudoinverse as well. Remember that in SVD (Theorem 16.2.2) the key is to transform the coordinate system from the standard basis to the β basis through the matrix V^*, then apply the factors σ_j through Σ, and finally transform the coordinate system again from the γ basis back to the standard basis via the matrix U. Therefore, it is logical to derive the pseudoinverse matrix by reversing these three steps. We only need to take special care when we invert the middle step involving the sigma factors Σ: In the corresponding Σ^+ matrix, the first k diagonal entries will be the reciprocals $1/\sigma_j$ for the k non-zero singular values, and beyond that they will be all zeros so that any $\vec{w} \in \mathcal{R}(T)^\perp$ is indeed mapped to the zero vector. Subsequently, the pseudoinverse matrix will look like $A^+ = V\Sigma^+U^*$

Properties 16.3.2 (Pseudoinverse of a Matrix)**.** Let A be an $m \times n$ matrix with a (full) SVD of $A = U\Sigma V^*$ and $k \leq \min(m, n)$ non-zero singular values

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$. Then, define a diagonal matrix Σ^+ where

$$\Sigma^+_{ij} = \begin{cases} \frac{1}{\sigma_j} & \text{if } i = j \leq k \\ 0 & \text{otherwise} \end{cases} \tag{16.33}$$

and the pseudoinverse of A will then be

$$A^+ = V\Sigma^+U^* \tag{16.34}$$

Note that we can also construct a "compact" pseudoinverse matrix just as the compact SVD by omitting the zero rows/columns of Σ^+, and the pseudoinverse matrix will also be unique up to a complex sign like the singular vectors themselves. In addition, the pseudoinverse of Σ is simply Σ^+.

Example 16.3.1. Find the pseudoinverse of the matrix in Example 16.2.3.

Solution. Note that U and V are the same as in the original example, with

$$\Sigma^+ = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$$

Hence the pseudoinverse matrix is

$$\begin{aligned} A^+ &= V\Sigma^+U^* \\ &= \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ 0 & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix} \end{aligned}$$

□

► Short Exercise: Check if (16.30) holds for the left singular vectors.[13]

However, be reminded that in general, the product between a matrix and its pseudoinverse may not produce the identity matrix, i.e. $A^+A \neq I$. Here, we have

$$A^+A = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix}$$

16.3.2. Minimal Solution

One major usage of pseudoinverses is to compute the "best-fit" solution $\vec{x}_f$ for the system $A\vec{x} = \vec{h}$ even if it has no solution (inconsistent) or infinitely many solutions, in the sense that the norm of $\vec{x}_f$ is minimized, and hence it is called the ***minimal solution***. As the name of pseudoinverses suggests, we can invert the system to acquire $\vec{x}_f = A^+\vec{h}$ by treating A^+ just like the inverse of A. To show the minimal properties of $\vec{x}_f = A^+\vec{h}$, we need to establish the following observations first.

Properties 16.3.3. Let $\mathcal{V}$ and $\mathcal{W}$ be two finite-dimensional inner product spaces, and $T\colon \mathcal{V} \to \mathcal{W}$ be a linear transformation between them. Then

(a) T^+T is the orthogonal projection of $\mathcal{V}$ onto $\mathcal{N}(T)^\perp$;

(b) TT^+ is the orthogonal projection of $\mathcal{W}$ onto $\mathcal{R}(T)$.

[13]We only check $\vec{u}^{(1)}$ here and leave $\vec{u}^{(2)}$ to the readers:

$$A^+\vec{u}^{(1)} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} \frac{2}{3\sqrt{6}} \\ 0 \\ \frac{1}{3\sqrt{6}} \\ \frac{1}{3\sqrt{6}} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} \frac{2}{\sqrt{6}} \\ 0 \\ \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} = \frac{1}{\sigma_1}\vec{v}^{(1)}$$

Proof. Refer to Section 10.3.2 for the definition of orthogonal projections. For (a), if $\vec{v} \in \mathcal{N}(T)^{\perp}$, then following Definition 16.3.1, we have $T(\vec{v}) = L(\vec{v}) \in \mathcal{R}(T)$ and thus $T^{+}T(\vec{v}) = L^{-1}(L(\vec{v})) = \vec{v}$. On the other hand, for $\vec{v} \in (\mathcal{N}(T)^{\perp})^{\perp} = \mathcal{N}(T)$, we simply have $T^{+}T(\vec{v}) = T^{+}(\mathbf{0}) = \mathbf{0}$. These two rules show that $T^{+}T$ is the orthogonal projection of $\mathcal{V}$ onto $\mathcal{N}(T)^{\perp}$. The proof of (b) is similar and we leave it as an exercise to the readers.[14] □

Now we are ready to show that $\vec{x}_f = A^{+}\vec{h}$ is indeed the "best-fit" solution to $A\vec{x} = \vec{h}$ with the minimum norm.

Properties 16.3.4. For a linear system $A\vec{x} = \vec{h}$ where A is an $m \times n$ matrix and $\vec{h}$ is a vector of length m, the proposed solution $\vec{x}_f = A^{+}\vec{h}$ will satisfy the following properties.

(a) If $A\vec{x} = \vec{h}$ is consistent, then $\vec{x}_f$ is the unique solution to the system such that it has the minimum norm possible, i.e. $A\vec{x}_f = \vec{h}$, and $\|\vec{x}_f\| \leq \|\vec{x}\|$ for any solution $\vec{x}$ to the system where the equality holds if and only if $\vec{x} = \vec{x}_f$;

(b) Otherwise, if $A\vec{x} = \vec{h}$ is inconsistent, then $\vec{x}_f$ is the unique best approximation to the solution of the system such that it still has the minimum norm possible, i.e. the error is minimized with $\|A\vec{x}_f - \vec{h}\| \leq \|A\vec{x} - \vec{h}\|$ for any $\vec{x}$ where the equality holds if and only if $A\vec{x} = A\vec{x}_f$. Furthermore, when there are multiple "best-fit" solutions $\vec{x}$ so that $A\vec{x} = A\vec{x}_f$, then $\|\vec{x}_f\| \leq \|\vec{x}\|$ for where the equality holds if and only if $\vec{x} = \vec{x}_f$.

Proof. For (a), given $A\vec{x} = \vec{h}$ is consistent, $\vec{h}$ will be in the column space of A, $\mathcal{R}(A)$, by Properties 6.3.7. So $A\vec{x}_f = AA^{+}\vec{h} = \vec{h}$ by part (b) of Properties 16.3.3 where we set $T = A$, and thus $\vec{x}_f$ is indeed a solution to the linear system. For any solution $\vec{x}$ to $A\vec{x} = \vec{h}$, we have

$$A^{+}A\vec{x} = A^{+}\vec{h} = \vec{x}_f$$

[14] Again by Definition 16.3.1, for $\vec{v} \in \mathcal{R}(T)$, $T^{+}(\vec{v}) = L^{-1}(\vec{v})$ and $TT^{+}(\vec{v}) = L(L^{-1}(\vec{v})) = \vec{v}$. While for $\vec{v} \in \mathcal{R}(T)^{\perp}$, we have $TT^{+}(\vec{v}) = T(\mathbf{0}) = \mathbf{0}$.

and by part (a) of Properties 16.3.3, it implies that $\vec{x}_f$ will be the orthogonal projection of $\vec{x}$ onto $\mathcal{N}(A)^\perp$. Using (10.9a) from Properties 10.3.3 immediately shows that, the shortest norm $\|\vec{x}\|$ where we fix $\vec{w} = \mathbf{0}$ is achieved by $\vec{x}_f \in \mathcal{N}(A)^\perp$ because

$$\|\vec{x}\| = \|\vec{x} - \mathbf{0}\|^2 \geq \|T(\vec{x}) - \mathbf{0}\|^2 = \left\|\vec{x}_f - \mathbf{0}\right\|^2 = \left\|\vec{x}_f\right\|^2$$

where $T = A^+A$ now. While for (b), if $A\vec{x} = \vec{h}$ is inconsistent, then $A\vec{x}_f = AA^+\vec{h}$ will be the orthogonal projection of $\vec{h}$ onto $\mathcal{R}(A)$ by part (b) of Properties 16.3.3. Therefore, using (10.9b) this time in Properties 10.3.3, the minimum error $\|A\vec{x} - \vec{h}\|$ will be achieved by $\|T(\vec{h}) - \vec{h}\| = \|AA^+\vec{h} - \vec{h}\| = \left\|A\vec{x}_f - \vec{h}\right\|$ where we now take $T = AA^+$, $\vec{v} = \vec{h}$ and $\mathcal{R}(A) = \mathcal{W}$. For any best-fit $\vec{x}$ that satisfies $A\vec{x} = A\vec{x}_f = \vec{y}$, multiplying both sides by A^+ to the left gives

$$A^+\vec{y} = A^+A\vec{x}_f = A^+A(A^+\vec{h}) = A^+\vec{h} = \vec{x}_f$$

The third equality holds because $A^+\vec{h} \in \mathcal{N}(T)^\perp$ and A^+A is the orthogonal projection onto $\mathcal{N}(T)^\perp$ by part (a) of Properties 16.3.3. Therefore, applying part (a) of the current properties on the other linear system $A\vec{x} = \vec{y}$ and the proposed solution $\vec{x}_f = A^+\vec{y}$ immediately shows that the shortest norm $\|\vec{x}\|$ with $\vec{x}$ still being the best approximation to the original equation $A\vec{x} = \vec{h}$ such that $A\vec{x} = \vec{y}$, is achieved by $\|\vec{x}_f\| \leq \|\vec{x}\|$. □

Example 16.3.2. Find the minimal solution of the system $A\vec{x} = \vec{h}$, where

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

is the matrix from Example 16.2.3, and $\vec{h} =$ (a) $(2, 0, 2)^T$ and (b) $(1, 1, 1)^T$.

Solution. By Properties 16.3.4, we simply compute $A^+\vec{h}$ for both cases, where

$$A^+ = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix}$$

by Example 16.3.1. In (a), it can be checked that the system is consistent for the given $\vec{h}$, and

$$A^+\vec{h} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \\ 0 \\ \frac{2}{3} \end{bmatrix}$$

It can be verified that this solves the linear system exactly. Meanwhile for (b), we have

$$A^+\vec{h} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{9} & \frac{2}{9} & \frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{4}{9} \\ 0 \\ \frac{2}{9} \\ \frac{2}{9} \end{bmatrix}$$

and

$$A(A^+\vec{h}) = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{4}{9} \\ 0 \\ \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \\ \frac{4}{3} \end{bmatrix} \neq \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \vec{h}$$

The minimum error is $\left\| AA^+\vec{h} - \vec{h} \right\| = \sqrt{\left(1 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2 + \left(1 - \frac{4}{3}\right)^2} = \frac{1}{\sqrt{3}}$. □

We note that the pseudoinverse solution coincides with the least-square solution in Theorem 13.1.3 when the full SVD is employed:

$$\vec{x}_f = (A^*A)^{-1}A^*\vec{h}$$

$$
\begin{aligned}
&= (V\Sigma^* U^* U \Sigma V^*)^{-1}(V\Sigma U^*)\vec{h} \\
&= (V\Sigma^* \Sigma V^*)^{-1}(V\Sigma U^*)\vec{h} && \text{(Definition 10.4.1, } U \text{ is unitary)} \\
&= (V^*)^{-1}\Sigma^+(\Sigma^+)^* V^{-1} V \Sigma U^* \vec{h} \\
&= V\Sigma^+(\Sigma^+)^*\Sigma U^*\vec{h} && \text{(Definition 10.4.1, } V \text{ is unitary)} \\
&= V\Sigma^+ U^*\vec{h} = A^+\vec{h} && \begin{aligned}&((\Sigma^+)^*\Sigma = I \text{ if } m \geq n \\ &\text{and } \Sigma \text{ is full-rank})\end{aligned}
\end{aligned}
$$

as long as A and thus Σ are full-rank. Finally, by Properties 16.3.3, we can arrive at the four equivalent defining properties of Moore-Penrose pseudoinverses as below.

Properties 16.3.5. An alternative set of definitions for the Moore-Penrose pseudoinverse A^+ of a matrix A is

1. $AA^+A = A$ (A is a *weak inverse*);
2. $A^+AA^+ = A^+$ (A^+ is also a weak inverse);
3. $(AA^+)^\dagger = AA^+$ (AA^+ is Hermitian);
4. $(A^+A)^\dagger = A^+A$ (A^+A is also Hermitian).

The first condition can be inferred as follows. $A\vec{v} \in \mathcal{R}(A)$ is in the column space of A for any $\vec{v}$, and thus $AA^+A\vec{v} = AA^+(A\vec{v}) = A\vec{v}$ as AA^+ is the orthogonal projection onto $\mathcal{R}(A)$ by (a) of Properties 16.3.3 and we must have $AA^+A = A$. The second condition is derived similarly. Finally, the last two conditions are the direct consequence of applying the part $T = T^\dagger$ of Properties 12.3.2 to Properties 16.3.3 where $T = A^+A$ and $T = AA^+$ are both orthogonal projections.

16.4. Generalization of SVD to Alternative Inner Products

In the Singular Value Theorem 16.2.1, the adjoint $T^\dagger$ of the linear transformation T is involved, but for simplicity, during the derivation of the matrix SVD, we have only discussed the case where the conjugate transpose A^* is used in place of the adjoint $A^\dagger$. However, as already shown in Section 12.2, we can define different matrix adjoints with respect to different inner products, and hence different notions of orthogonality and distance. Therefore, one may ask if we can come up with a variant of matrix SVD and pseudoinverses that incorporates the symmetric/Hermitian bilinear form associated with any inner product, and for example, at the same time still produces orthonormal bases of left/right singular vectors. The answer is possible, with appropriate changes to the previous formulae about SVD while the essence of its properties remains intact. We now state the formulation below.

Definition 16.4.1 (SVD for an Alternative Adjoint)**.** For any $m \times n$ matrix A, its adjoint with respect to the two inner products defined by the symmetric/Hermitian bilinear forms B and C for $\mathbb{C}^m$ and $\mathbb{C}^n$, is

$$A^\dagger = C^{-1}A^*B \tag{16.35}$$

(replace A^* by A^T for $\mathbb{R}^m$ and $\mathbb{R}^n$) Meanwhile, the new matrix SVD with respect to this adjoint is

$$A = U\Sigma V^*C \tag{16.36}$$

where U and V are $m \times m$ and $n \times n$ matrices, and Σ is an $m \times n$ diagonal matrix. The matrix V is constructed by

$$V = \left[\hat{v}^{(1)}|\hat{v}^{(2)}|\cdots|\hat{v}^{(n)}\right] \tag{16.37}$$

where the column vectors $\hat{v}^{(1)}, \hat{v}^{(2)}, \ldots, \hat{v}^{(n)}$ are the orthonormal (now with respect to the inner product described by C) eigenvectors of the Hermitian

matrix product $A^\dagger A$ (Properties 12.2.6, also called the ***right singular vectors*** in this case), and ordered by decreasing, real (Properties 12.2.5) eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ as well. Σ then consists of the first $k = \min(m, n)$ ***singular values*** $\sigma_j = \sqrt{\lambda_j}$ which are the square roots of the eigenvalues of $A^\dagger A$ (also in decreasing order) for these right singular vectors $\hat{v}^{(j)}$ in the (unitary, with respect to C) matrix V along the main diagonal, and has a value of zero elsewhere. Finally, U is made up of the orthonormal (with respect to the inner product described by B) column vectors that satisfy

$$\hat{u}^{(j)} = \frac{1}{\sigma_j} A \hat{v}^{(j)} \tag{16.38}$$

for any j that is associated with a non-zero singular value $\sigma_j \neq 0$. These vectors $\hat{u}^{(j)}$ are also called the ***left singular vectors*** in this case. Again, when there are not enough column vectors to construct U by this relation, use Gram-Schmidt Orthogonalization (Definition 12.3.3 adapted for the new inner product) to create orthonormal vectors that extend the basis of $\hat{u}^{(j)}$ up to $j = m$.

Definition 16.4.2 (Pseudoinverse of a Matrix for an Alternative Adjoint)**.** Let A be an $m \times n$ matrix with the (full) SVD of $A = U\Sigma V^* C$ with respect to some adjoint $A^\dagger$ as given in (16.36) of Theorem 16.4.1 that has $k \leq \min(m, n)$ non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$. Then define a diagonal matrix Σ^+ where

$$\Sigma^+_{ij} = \begin{cases} \frac{1}{\sigma_j} & \text{if } i = j \leq k \\ 0 & \text{otherwise} \end{cases} \tag{16.39}$$

and the pseudoinverse of A with respect to this adjoint will then be

$$A^+ = V\Sigma^+ U^* B \tag{16.40}$$

It means that SVD and pseudoinverses with respect to alternative inner products require two major modifications. First, the singular vectors are derived from

$A^\dagger A$ instead of $A^* A$ and will be orthonormal with respect to these inner products instead of the usual Euclidean inner products. Second, the SVD and pseudoinverse will need an extra scaling by the bilinear forms C and B first. With these appropriate modifications, all the properties proposed in Section 16.3.2 will continue to hold. Now let's see how we find the minimal solution of an inconsistent system when the norms and errors are defined with respect to other inner products different from the standard inner products.

Example 16.4.1. Find the minimal solution of the inconsistent system $A\vec{x} = \vec{h}$, where

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

where the error between $\vec{h}$ and $A\vec{x}$ is computed with respect to the inner product defined by

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$
$$\|A\vec{x} - \vec{h}\|^2 = (A\mathbf{x} - \mathbf{h})^T B(A\mathbf{x} - \mathbf{h})$$

and the norm $\|\vec{x}\|$ is computed with respect to the inner product defined by

$$C = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$
$$\|\vec{x}\|^2 = \mathbf{x}^T C\mathbf{x}$$

Solution. By (16.35), the new adjoint is

$$A^\dagger = C^{-1} A^* B$$

$$
\begin{aligned}
&= \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 \\ -1 & -2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{3}{2} & \frac{5}{2} \\ -4 & -\frac{20}{3} \\ 5 & \frac{25}{3} \end{bmatrix}
\end{aligned}
$$

Then $A^\dagger A$ is computed as

$$
\begin{aligned}
A^\dagger A &= \begin{bmatrix} \frac{3}{2} & \frac{5}{2} \\ -4 & -\frac{20}{3} \\ 5 & \frac{25}{3} \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} \\
&= \begin{bmatrix} \frac{13}{2} & -\frac{13}{2} & 13 \\ -\frac{52}{3} & \frac{52}{3} & -\frac{104}{3} \\ \frac{65}{3} & -\frac{65}{3} & \frac{130}{3} \end{bmatrix}
\end{aligned}
$$

The only non-zero eigenvalue and eigenvector of $A^\dagger A$ can be seen to be

$$
\vec{v}^{(1)} = \begin{bmatrix} 3 \\ -8 \\ 10 \end{bmatrix} \qquad \text{for } \lambda_1 = \frac{403}{6}
$$

and hence the corresponding singular value and the normalized right singular vector (with respect to C) will be

$$
\hat{v}^{(1)} = \begin{bmatrix} \frac{3}{\sqrt{186}} \\ -\frac{8}{\sqrt{186}} \\ \frac{10}{\sqrt{186}} \end{bmatrix} \qquad \text{for } \sigma_1 = \sqrt{\frac{403}{6}}
$$

The corresponding left singular vector is then derived according to (16.38) as

$$
\hat{u}^{(1)} = \frac{1}{\sigma_1} A \hat{v}^{(1)}
$$

$$= \frac{1}{\sqrt{\frac{403}{6}}} \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} \begin{bmatrix} \frac{3}{\sqrt{186}} \\ -\frac{8}{\sqrt{186}} \\ \frac{10}{\sqrt{186}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{13}} \\ \frac{2}{\sqrt{13}} \end{bmatrix}$$

Therefore, following (16.40), the pseudoinverse A^+ with respect to the adjoint is constructed by (using only the compact part of its SVD)

$$\begin{aligned} A^+ &= V\Sigma^+U^*B \\ &= \hat{v}^{(1)} \left(\frac{1}{\sigma_1}\right) \hat{u}^{(1)*}B \\ &= \begin{bmatrix} \frac{3}{\sqrt{186}} \\ -\frac{8}{\sqrt{186}} \\ \frac{10}{\sqrt{186}} \end{bmatrix} \left(\frac{1}{\sqrt{\frac{403}{6}}}\right) \begin{bmatrix} \frac{1}{\sqrt{13}} & \frac{2}{\sqrt{13}} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{9}{403} & \frac{15}{403} \\ -\frac{24}{403} & -\frac{40}{403} \\ \frac{30}{403} & \frac{50}{403} \end{bmatrix} \end{aligned}$$

and then the minimal solution as in Properties 16.3.4 will be

$$\begin{aligned} \vec{x}_f &= A^+\vec{h} \\ &= \begin{bmatrix} \frac{9}{403} & \frac{15}{403} \\ -\frac{24}{403} & -\frac{40}{403} \\ \frac{30}{403} & \frac{50}{403} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{54}{403} \\ -\frac{144}{403} \\ \frac{180}{403} \end{bmatrix} \end{aligned}$$

□

► Short Exercise: Check if (1) (as well as (2)) in Properties 16.3.5 holds in this example.[15]

16.5. Earth Science Applications

Example 16.5.1. Find the minimal solution to the ion concentrations in the seawater sample in Example 1.4.3.

Solution. The linear system can be written as $A\vec{x} = \vec{h}$ where

$$A = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \qquad \vec{h} = \begin{bmatrix} 0 \\ 34.0 \end{bmatrix}$$

We first seek the SVD of the matrix A as suggested by Theorem 16.2.2. We have

$$\begin{aligned} A^T A &= \begin{bmatrix} 1 & 23.0 \\ 2 & 24.3 \\ -1 & 35.5 \\ -2 & 96.1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \\ &= \begin{bmatrix} 530 & 560.9 & 815.5 & 2208.3 \\ 560.9 & 594.49 & 860.65 & 2331.23 \\ 815.5 & 860.65 & 1261.25 & 3413.55 \\ 2208.3 & 2331.23 & 3413.55 & 9239.21 \end{bmatrix} \end{aligned}$$

It would be wise to delegate the calculation of the eigenvalues of $A^T A$ and thus the singular values of A to a computer, and we simply note the results for the

[15]We will only show (1) and leave (2) to the readers. Both of them are routine calculations.

$$\begin{aligned} AA^+A &= \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} \begin{bmatrix} \frac{9}{403} & \frac{15}{403} \\ -\frac{24}{403} & -\frac{40}{403} \\ \frac{30}{403} & \frac{50}{403} \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} \\ &= \begin{bmatrix} \frac{3}{13} & \frac{5}{13} \\ \frac{6}{13} & \frac{10}{13} \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \end{bmatrix} = A \end{aligned}$$

two non-zero singular values below.

$$\hat{v}^{(1)} = \begin{bmatrix} 0.21325 \\ 0.22518 \\ 0.32946 \\ 0.89178 \end{bmatrix} \qquad \sigma_1 \approx 107.78$$

$$\hat{v}^{(2)} = \begin{bmatrix} -0.46577 \\ -0.82773 \\ 0.18590 \\ 0.25171 \end{bmatrix} \qquad \sigma_2 \approx 2.81$$

and

$$\begin{aligned} \hat{u}^{(1)} &= \frac{1}{\sigma_1} A\hat{v}^{(1)} \\ &= \frac{1}{107.78} \begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \begin{bmatrix} 0.21325 \\ 0.22518 \\ 0.32946 \\ 0.89178 \end{bmatrix} = \begin{bmatrix} -0.01345 \\ 0.99991 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \hat{u}^{(2)} &= \frac{1}{\sigma_2} A\hat{v}^{(2)} \\ &= \frac{1}{2.81} \begin{bmatrix} 1 & 2 & -1 & -2 \\ 23.0 & 24.3 & 35.5 & 96.1 \end{bmatrix} \begin{bmatrix} -0.46577 \\ -0.82773 \\ 0.18590 \\ 0.25171 \end{bmatrix} = \begin{bmatrix} -0.99991 \\ -0.01345 \end{bmatrix} \end{aligned}$$

Hence according to Formula (16.40) in Properties 16.3.2, the pseudoinverse of A is

$$\begin{aligned} A^+ &= V\Sigma^+U^* \\ &= \begin{bmatrix} 0.21325 & -0.46577 \\ 0.22518 & -0.82773 \\ 0.32946 & 0.18590 \\ 0.89178 & 0.25171 \end{bmatrix} \begin{bmatrix} \frac{1}{107.78} & 0 \\ 0 & \frac{1}{2.81} \end{bmatrix} \begin{bmatrix} -0.01345 & 0.99991 \\ -0.99991 & -0.01345 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 0.1657 & 0.0042 \\ 0.2944 & 0.0060 \\ -0.0662 & 0.0022 \\ -0.0897 & 0.0071 \end{bmatrix}$$

and thus the minimal solution will be

$$\begin{aligned} \vec{x}_f &= A^+\vec{h} \\ &= \begin{bmatrix} 0.1657 & 0.0042 \\ 0.2944 & 0.0060 \\ -0.0662 & 0.0022 \\ -0.0897 & 0.0071 \end{bmatrix} \begin{bmatrix} 0 \\ 34.0 \end{bmatrix} = \begin{bmatrix} 0.1430 \\ 0.2057 \\ 0.0737 \\ 0.2403 \end{bmatrix} \end{aligned}$$

We can convert the molar concentration of each ion species to mass concentration via multiplying them by their respective molar mass. For example,

$$[\mathrm{Na}^+] = (23.0\,\mathrm{g\,mol}^{-1})(0.1430\,\mathrm{mol\,kg}^{-1}) = 3.29\,\mathrm{g\,kg}^{-1}$$

and similarly $[\mathrm{Mg}^{2+}] = (24.3)(0.2057) = 5.00\,\mathrm{g\,kg}^{-1}$, $[\mathrm{Cl}^-] = (35.5)(0.0737) = 2.62\,\mathrm{g\,kg}^{-1}$, $[\mathrm{SO}_4^{2-}] = (96.1)(0.2403) = 23.09\,\mathrm{g\,kg}^{-1}$. However, while this minimal solution is a valid solution by part (a) of Properties 16.3.4, it is far from the actual observed composition (see https://www.britannica.com/science/seawater). This is because the minimal solution also tries to minimize the squared norm $\|\vec{x}\|^2$, which tends to average out the components. Therefore, we have to be very careful if we attempt to obtain a single solution for a linear system that admits infinitely many solutions using the pseudoinverse. □

16.6. Python Programming

First, we will demonstrate how to carry out Cholesky Factorization and L(D)U Decomposition. We will take the matrices in Examples 16.1.1 and 16.1.3 for those.

```
import numpy as np
from scipy import linalg

A1 = np.array([[4, 2, 0],
               [2, 2, 2],
               [0, 2, 5]])
print(linalg.cholesky(A1))
```

gives

```
[[2. 1. 0.]
 [0. 1. 2.]
 [0. 0. 1.]]
```

as expected, and

```
A2 = np.array([[1, 1, 2],
               [2, 2, 3],
               [1, 3, 0]])
P, L, U_raw = linalg.lu(A2)
D = np.diag(np.diag(U_raw)) # Factor out the diagonal elements
    in U_raw
U = U_raw/np.diag(U_raw)[:,None]
print(P, L, D, U)
print(np.allclose(P@L@D@U, A2)) # Check if the factors can be
   combined to reconstruct the matrix
```

which produces

```
[[0.  0.  1. ]
 [1.  0.  0. ]
 [0.  1.  0. ]]
[[1.  0.  0.  ]
 [0.5 1.  0.  ]
 [0.5 0.  1.  ]]
[[2.  0.  0.  ]
 [0.  2.  0.  ]
 [0.  0.  0.5 ]]
[[1.  1.  1.5 ]
 [0.  1.  -0.75]
```

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

as another valid outcome of (P)LDU Decomposition different from the one derived in Example 16.1.3. Now we will move to the central topic in this chapter, the Singular Value Decomposition (SVD). We are going to use Example 16.2.3 to illustrate how to perform that. The results of

```
A3 = np.array([[1, 1, 0, 1],
               [1,-1, 1, 0],
               [2, 0, 1, 1]])
U, Sigma, V_H = linalg.svd(A3)
print(U, Sigma, V_H)
```

are

```
[[-4.08248290e-01 -7.07106781e-01 -5.77350269e-01]
 [-4.08248290e-01  7.07106781e-01 -5.77350269e-01]
 [-8.16496581e-01  2.22044605e-16  5.77350269e-01]]
 [3.00000000e+00 1.73205081e+00 9.06493304e-17]
[[-8.1649658e-01 -1.1102230e-16 -4.0824829e-01 -4.0824829e-01]
 [ 2.2204460e-16 -8.1649658e-01  4.0824829e-01 -4.0824829e-01]
 [-4.8232914e-02 -5.7533200e-01 -5.2709909e-01  6.2356492e
    -01]]
```

which essentially matches the full SVD of the matrix where the extremely small values are actually zeros but affected by numerical round-off errors, and there are some sign differences. The pseudoinverse can be simply computed by

```
print(linalg.pinv(A3))
```

that returns

```
[[ 1.11111111e-01  1.11111111e-01  2.22222222e-01]
 [ 3.33333333e-01 -3.33333333e-01 -7.44563871e-17]
 [-1.11111111e-01  2.22222222e-01  1.11111111e-01]
 [ 2.22222222e-01 -1.11111111e-01  1.11111111e-01]]
```

which is expected by Example 16.3.1, and the minimal solution in Example 16.3.2 can be found via either

```
h = np.array([1, 1, 1])
print(linalg.pinv(A3) @ h)
print(linalg.lstsq(A3, h))
```

both will produce the same answer apart from the different, but negligible round-off errors in the second entry which should have been zero.

16.7. Exercises

Exercise 16.1 Show that the product of a sequence of lower[upper]-triangular matrices of the same extent ultimately results in a lower[upper]-triangular matrix as well.

Exercise 16.2 Find the (a) Cholesky Factorization, (b) LDU Decomposition for the symmetric matrix

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

Exercise 16.3 Find the L(D)U Decomposition for

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

and solve the linear system $A\vec{x} = \vec{h}$ where $\vec{h} =$ (a) $(1, -3, 4)^T$, (b) $(2, 3, 1)^T$ by the LU method.

Exercise 16.4 Find the SVD and pseudoinverse for

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

How about those of A^T?

Exercise 16.5 Find the minimal solution to the linear system

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

How about if the norm of the solution is computed according to the inner product defined by (numerical answer is accepted)

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

$$||\vec{x}||^2 = \mathbf{x}^T C \mathbf{x}$$

Chapter 17

Dynamical Systems

Near the end of Chapter 9, we have talked about systems of first-order linear ODEs which are actually the prototype of *linear* dynamical systems. A *dynamical system* describes the change or evolution (*dynamics*) of multiple variables (*state*) in time that can depend on the variables themselves as well as the time coordinate. Starting from a particular initial point (*particle*) in the physical/state space, a deterministic trajectory will be traced by the particle as the dynamical system operates. We will derive several important ideas like equilibrium points and their types, starting from the easier linear dynamical systems, and then generalize to the more complicated *non-linear* dynamical systems where the gap can often be bridged by the *Hartman-Grobman Theorem* if the linearized equilibrium point in question is a *hyperbolic equilibrium point*. Finally, we will discuss the prime example of non-linear dynamical systems in Earth Science, the *Lorenz-63 Model* which demonstrates the phenomenon of *chaos*.

17.1. Linear Dynamical Systems

17.1.1. Types of Equilibrium Points in Two-dimensional Linear Dynamical Systems

We will motivate the notions of equilibrium points and their types from the perspective of the relatively simpler two-dimensional linear dynamical systems. Recalling Definition 9.3.3, such dynamical systems may be succinctly described by

$$\begin{cases} \dfrac{dx}{dt} = \alpha_1^{(1)} x + \alpha_2^{(1)} y \\ \dfrac{dy}{dt} = \alpha_1^{(2)} x + \alpha_2^{(2)} y \end{cases} \tag{17.1}$$

where $\alpha_j^{(i)}$ are constant coefficients and there are no time-dependent forcing terms. Alternatively, in matrix form, it is

$$\mathbf{y}' = A\mathbf{y} \tag{17.2}$$

where $\mathbf{y} = (x, y)^T$ and

$$A = \begin{bmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} \\ \alpha_1^{(2)} & \alpha_2^{(2)} \end{bmatrix} \tag{17.3}$$

An ***equilibrium point*** of a dynamical system is where the R.H.S. vanishes (i.e. $\mathbf{y}' = \mathbf{0}$) so that the time derivatives are zeros and the variables (or physically, a *particle*) become stationary. The characteristics of the trajectories/*orbits* near these equilibrium points are important and can be readily inferred from their types.

Definition 17.1.1 (Equilibrium Points of a Dynamical System)**.** An equilibrium

point $\tilde{\mathbf{y}}$ of a *time-independent* dynamical system

$$\frac{d\mathbf{y}}{dt} = F(\mathbf{y}) \tag{17.4}$$

where F is a vector of functions that only depend on the state $\mathbf{y}$, satisfies

$$F(\tilde{\mathbf{y}}) = \mathbf{0} \tag{17.5}$$

It is not hard to see that the origin $\mathbf{0}$ is automatically an (only) equilibrium point for the linear dynamical system in (17.2) if A is invertible[1] and we will focus on this scenario. As we have seen in Section 9.3, the characteristics (growth/decay/spiral centered at the origin) are largely determined by the eigenvalues of A through diagonalization, assumed possible. Specifically, by Properties 9.3.4 the solution trajectories/orbits (***field lines***) for (17.2) will be described by $z_1 = c_1 e^{\lambda_1 t}$, $z_2 = c_2 e^{\lambda_2 t}$ where λ_1, λ_2 are the respective eigenvalues of A and $\mathbf{z} = (z_1, z_2)^T$ denotes the transformed coordinates in the basis formed by the eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$. The family of trajectories in the original physical state space will then be in the form of

$$\mathbf{y} = c_1 \mathbf{v}_1 e^{\lambda_1 t} + c_2 \mathbf{v}_2 e^{\lambda_2 t} \tag{17.6}$$

According to this, we can determine the types of equilibrium points at the origin based on the eigenvalues λ_1, λ_2 of the coefficient matrix A for any two-dimensional linear dynamical system.

Properties 17.1.2 (Types of Equilibrium Points for Two-dimensional Linear Dynamical Systems)**.** Given a two-dimensional dynamical system $\mathbf{y}' = A\mathbf{y}$ where A is a real 2×2 matrix, denote the eigenvalues of A by λ_1 and λ_2 which are exactly the roots of the characteristic polynomial of A. Further assume that

[1] If A is not invertible, it means that there exists a non-trivial null space for A (Properties 6.3.10) and by Definition 6.3.8 this null space $\mathcal{N}(A)$ will be a subspace (straight line) where for $\tilde{\mathbf{y}} \in \mathcal{N}(A)$, $\tilde{\mathbf{y}}' = A\tilde{\mathbf{y}} = \mathbf{0}$ so there will be an "equilibrium line" instead and it becomes a degenerate case.

A is invertible such that there is no eigenvalue of pure zero, i.e. $\lambda_j \neq 0$ (Theorem 9.1.6). The equilibrium point at the origin $\tilde{\mathbf{y}} = \mathbf{0}$ can be classified into different types depending on the following three situations.

1. If the characteristic polynomial has two distinct real roots, then it is
   (a) a ***source (unstable node)*** where the field lines emanate from it if both of the eigenvalues are positive, i.e. $\lambda_1, \lambda_2 > 0$;
   (b) a ***sink (stable node)*** where the field lines converge to it if both of the eigenvalues are negative, i.e. $\lambda_1, \lambda_2 < 0$;
   (c) otherwise, a ***saddle point*** where the field lines are deflected away from it if one of the eigenvalues is positive and the other one is negative. i.e. $\lambda_1 > 0, \lambda_2 < 0$.
2. Else if, the characteristic polynomial has a repeated real root $\lambda_1 = \lambda_2 = \lambda_p$, then it is
   (a) a source if $\lambda_p > 0$;
   (b) a sink if $\lambda_p < 0$;

   where for simplicity we require that the geometric multiplicity of the repeated eigenvalue is equal to its algebraic multiplicity = 2, and thus A is diagonalizable.
3. Otherwise, the characteristic polynomial has two complex conjugate roots $\lambda_1 = a + bi$, $\lambda_2 = a - bi$, $b \neq 0$, then
   (a) an ***outward/unstable spiral (source)*** if $a > 0$;
   (b) an ***inward/stable spiral (sink)*** if $a < 0$;
   (c) a ***center*** if $a = 0$;

These various types of equilibrium points in a two-dimensional space are summarized in Table 17.1 below. Again, assume that the matrix A can be

diagonalized into

$$D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

as suggested by Properties 9.2.2, then due to Properties 9.2.3, the trace $\operatorname{tr}(A) = \lambda_1 + \lambda_2$ and determinant $\det(A) = \lambda_1\lambda_2$ will remain invariant. These are exactly the sum and product of the roots for the characteristic polynomial and by high school algebra we can readily infer the signs of the two eigenvalues from them. Moreover, we can know if the roots belong to the case of real and distinct, real and repeated, or complex conjugates, by calculating the discriminant $\Delta = \operatorname{tr}(A)^2 - 4\det(A)$ from the characteristic polynomial

$$\begin{aligned} p(\lambda) &= (\lambda_1 - \lambda)(\lambda_2 - \lambda) \\ &= \lambda_1\lambda_2 - (\lambda_1 + \lambda_2)\lambda + \lambda^2 \\ &= \det(A) - \operatorname{tr}(A)\lambda + \lambda^2 \end{aligned}$$

which is also invariant. Hence we can determine the different types of equilibrium points (Situations 1 and 3 in Properties 17.1.2)[2] solely from the trace and determinant of A (see the three rightmost columns in Table 17.1). The classification and the corresponding phase portraits are visualized in Figure 17.2 as follows.

Example 17.1.1. Identify the type of equilibrium point at the origin for the three two-dimensional dynamical systems $\mathbf{y}' = A\mathbf{y}$ where the coefficient matrices $A =$

(a) $\begin{bmatrix} -1 & 4 \\ 2 & 1 \end{bmatrix}$

(b) $\begin{bmatrix} 3 & 2 \\ 1 & 3 \end{bmatrix}$

[2] Situation 2 is a bit special since our assumption may be violated where the geometric multiplicity of the repeated eigenvalue is now only 1, i.e. deficient, and it becomes degenerate.

Type of Roots	Signs of Roots	Type of Equilibrium Points	Stability	$\mathrm{tr}(A)$	$\det(A)$	Δ
Two Distinct Real Roots	Both Positive	Source (Unstable Node)	Unstable	$(+)$	$(+)$	$(+)$
	Both Negative	Sink (Stable Node)	Stable	$(-)$	$(+)$	$(+)$
	One Positive, One Negative	Saddle Point	Unstable	Any	$(-)$	$(+)$
One Repeated (Real) Root	Positive	Source (Unstable Node)	Unstable	$(+)$	$(+)$	0
	Negative	Sink (Stable Node)	Stable	$(-)$	$(+)$	0
Two Complex (Conjugate) Roots	Positive Real Part	Outward / Unstable Spiral	Unstable	$(+)$	$(+)$	$(-)$
	Negative Real Part	Inward / Stable Spiral	Stable	$(-)$	$(+)$	$(-)$
	Zero Real Part	Center	Stable	0	$(+)$	$(-)$

Table 17.1.: *Types and attributes of equilibrium points in a two-dimensional dynamical system. The discriminant is* $\Delta = \mathrm{tr}(A)^2 - 4\det(A)$.

(c) $\begin{bmatrix} -1 & -1 \\ 2 & -1 \end{bmatrix}$

Solution. We simply do the classification according to Table 17.1. For case (a), the trace is $\mathrm{tr}(A) = (-1)+1 = 0$, the determinant is $\det(A) = (-1)(1)-(4)(2) = -9 < 0$ and the discriminant is $\Delta = (0)^2 - 4(-9) = 36 > 0$, so it corresponds to a saddle point. Similarly, for case (b), we have $\mathrm{tr}(A) = 3+3 = 6 > 0$, $\det(A) = (3)(3) - (2)(1) = 7 > 0$ and $\Delta = (6)^2 - 4(7) = 8 > 0$, thus it is an unstable node. Finally, for case (c), $\mathrm{tr}(A) = (-1) + (-1) = -2 < 0$,

Figure 17.2.: *Different types of equilibrium points and their representative phase portraits categorized according to the trace and determinant of the coefficient matrices for the family of two-dimensional dynamical systems.*

$\det(A) = (-1)(-1) - (2)(-1) = 3 > 0$ and $\Delta = (-2)^2 - 4(3) = -8 < 0$ and therefore it is a stable spiral. The stream plots of these three dynamical systems are shown in Figure 17.3. □

Figure 17.3.: *The stream plots for the three two-dimensional dynamical systems in Example 17.1.1.*

17.1.2. Generalization to Time-independent Forcings and Higher-dimensional Linear Dynamical Systems

Sometimes there will be an additional forcing term on the R.H.S. of the linear dynamical system that represents external processes. Let's further assume that it is time-independent just like the dynamical system itself before, so that we have

$$\frac{d\mathbf{y}}{dt} = A\mathbf{y} + \mathbf{G} \tag{17.7}$$

where $\mathbf{G}$ is a constant vector with the same dimension as the state vector $\mathbf{y}(t)$. As in Properties 17.1.2, we require that A is invertible, and hence by Theorem 3.1.2 there exists a unique solution $\tilde{\mathbf{y}}$ for the linear system $A\mathbf{y} = -\mathbf{G}$. Subsequently, the dynamical system can be rewritten into

$$\frac{d\mathbf{y}(t)}{dt} = A\mathbf{y}(t) - A\tilde{\mathbf{y}} = A(\mathbf{y}(t) - \tilde{\mathbf{y}}) \tag{17.8}$$

where a linear change of variable $\delta\mathbf{y}(t) = \mathbf{y}(t) - \tilde{\mathbf{y}}$ will make it become

$$\frac{d(\delta\mathbf{y}(t))}{dt} = A(\delta\mathbf{y}(t)) \tag{17.9}$$

which is essentially the same as (17.2) but with $\delta\mathbf{y}(t)$ in place of $\mathbf{y}(t)$. Therefore, both the qualitative and quantitative behaviors of the dynamical system will also be the same, except that the equilibrium point is now located by $\delta\mathbf{y} = 0$, i.e. moved to $\mathbf{y} = \tilde{\mathbf{y}}$ instead of the original $\mathbf{y} = \mathbf{0}$. This effectively translates the equilibrium point and the entire corresponding phase portrait by a displacement of $\tilde{\mathbf{y}}$, and the results from the previous discussion can thus be readily applied to linear dynamical systems with a time-independent forcing. Particularly, for a two-dimensional linear dynamical system as in (17.1), if we add an additional time-independent term so that it now looks like

$$\begin{cases} \dfrac{dx}{dt} = \alpha_1^{(1)} x + \alpha_2^{(1)} y + G_x \\ \dfrac{dy}{dt} = \alpha_1^{(2)} x + \alpha_2^{(2)} y + G_y \end{cases} \tag{17.10}$$

where $(G_x, G_y)^T = \mathbf{G}$ are some fixed constants, and $\tilde{\mathbf{y}} = (\tilde{x}, \tilde{y})^T$ is the unique solution to the linear system $A\mathbf{y} = -\mathbf{G}$ with A in the form of (17.3) and invertible such that

$$\begin{cases} \alpha_1^{(1)} \tilde{x} + \alpha_2^{(1)} \tilde{y} = -G_x \\ \alpha_1^{(2)} \tilde{x} + \alpha_2^{(2)} \tilde{y} = -G_y \end{cases} \tag{17.11}$$

then the dynamical system can be recast as

$$\begin{cases} \dfrac{dx}{dt} - \dfrac{d\tilde{x}}{dt} = \alpha_1^{(1)} x + \alpha_2^{(1)} y - \alpha_1^{(1)} \tilde{x} + \alpha_2^{(1)} \tilde{y} \\ \dfrac{dy}{dt} - \dfrac{d\tilde{y}}{dt} = \alpha_1^{(2)} x + \alpha_2^{(2)} y - \alpha_1^{(2)} \tilde{x} + \alpha_2^{(2)} \tilde{y} \end{cases}$$

$$\begin{cases} \dfrac{d(x - \tilde{x})}{dt} = \alpha_1^{(1)} (x - \tilde{x}) + \alpha_2^{(1)} (y - \tilde{y}) \\ \dfrac{d(y - \tilde{y})}{dt} = \alpha_1^{(2)} (x - \tilde{x}) + \alpha_2^{(2)} (y - \tilde{y}) \end{cases}$$

$$\Rightarrow \begin{cases} \dfrac{d(\delta x)}{dt} = \alpha_1^{(1)}\delta x + \alpha_2^{(1)}\delta y \\ \dfrac{d(\delta y)}{dt} = \alpha_1^{(2)}\delta x + \alpha_2^{(2)}\delta y \end{cases} \tag{17.12}$$

which is essentially the same as (17.1) but with displacements from the newly translated equilibrium point δx and δy replacing the x and y coordinates that are initially relative to the origin.

Example 17.1.2. Locate the equilibrium point for the two-dimensional linear dynamical system

$$\begin{cases} \dfrac{dx}{dt} = -5x + 2y + 8 \\ \dfrac{dy}{dt} = x - 4y + 2 \end{cases}$$

and determined its type.

Solution. First, the position of the equilibrium point is found by solving the linear system $A\mathbf{y} = -\mathbf{G}$ where

$$A = \begin{bmatrix} -5 & 2 \\ 1 & -4 \end{bmatrix} \qquad \mathbf{G} = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} 8 \\ 2 \end{bmatrix}$$

It can be checked that the solution is

$$\tilde{\mathbf{y}} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

as

$$\begin{bmatrix} -5 & 2 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -8 \\ -2 \end{bmatrix} = -\begin{bmatrix} 8 \\ 2 \end{bmatrix}$$

and thus the equilibrium point is at $\tilde{\mathbf{y}} = (2, 1)^T$. Next, translate the coordinate system so that it is now relative to this equilibrium point by writing $\delta x = x - 2$

and $\delta y = y - 1$. The dynamical system now looks like

$$\begin{cases} \dfrac{d(x-2)}{dt} = -5(x-2) + 2(y-1) \\ \dfrac{d(y-1)}{dt} = (x-2) - 4(y-1) \end{cases}$$
$$\begin{cases} \dfrac{d(\delta x)}{dt} = -5\delta x + 2\delta y \\ \dfrac{d(\delta y)}{dt} = \delta x - 4\delta y \end{cases}$$

the one in (17.9) and the type of this equilibrium point can then be inferred from Table 17.1. The trace, determinant and discriminant are $\mathrm{tr}(A) = (-5) + (-4) = -9 < 0$, $\det(A) = (-5)(-4) - (1)(2) = 18 > 0$ and $\Delta = (-9)^2 - 4(18) = 9 > 0$, and therefore it is a stable node located at $(2, 1)^T$. □

Another generalization to be made is for higher-dimensional dynamical systems. The logic in (17.6) and Properties 17.1.2 remains the same. For a dynamical system $\mathbf{y}' = A\mathbf{y}$ of any dimension, let's say n, denote all the pair of eigenvalues and eigenvectors by λ_j and $\vec{v}_\lambda^{(j)}$, $j = 1, 2, \ldots, n$. Then the subspace spanned by those $\vec{v}_\lambda^{(j)}$ with negative real λ_j form the stable part of a node, and in contrast, those with positive real λ_j form the unstable part of a node. If there are both stable and unstable parts, then they will give rise to an overall saddle point. The subspace spanned by a pair of $\vec{v}_\lambda^{(j)}$ with the eigenvalues λ_j being complex conjugates will lead to an unstable spiral if their real part is positive, a stable spiral if their real part is negative, and a center if it is exactly zero. The characteristic of the equilibrium point can be viewed either separately or together in these different subspaces (more formally known as *"manifolds"* in this context).

17.2. Non-linear Dynamical Systems

17.2.1. Heuristic Understandings of Stability and Attractiveness

Before moving from linear dynamical systems to non-linear dynamical systems, we will take a detour to visit the concepts of stability and attractiveness. Heuristically, stability means that if the particle starts close enough to the equilibrium point, then it will always remain close to the equilibrium point. Meanwhile, attractiveness means that if it starts close enough to the equilibrium point likewise, then it will eventually be absorbed to that equilibrium point. We give the formal definitions for both of them below.

Definition 17.2.1. For a time-independent dynamical system as given by (17.4) in Definition 17.1.1, an equilibrium point $\tilde{\mathbf{y}}$ that satisfies (17.5) is

(a) ***stable***, if for all $\epsilon > 0$, there exists an $\eta(\epsilon) > 0$[3] such that $\left\|\mathbf{y}(\mathbf{y}_0, 0) - \tilde{\mathbf{y}}\right\| \leq \eta$ implies that $\left\|\mathbf{y}(\mathbf{y}_0, t) - \tilde{\mathbf{y}}\right\| \leq \epsilon$ for all $t \geq 0$;

(b) ***attractive***, if there exists an $\eta > 0$ such that $\left\|\mathbf{y}(\mathbf{y}_0, 0) - \tilde{\mathbf{y}}\right\| \leq \eta$ implies that $\left\|\mathbf{y}(\mathbf{y}_0, t) - \tilde{\mathbf{y}}\right\| \to 0$ as $t \to \infty$.

where $\mathbf{y}(\mathbf{y}_0, t)$ denotes the position of the particle at time t with an initial condition (starting point) of $\mathbf{y}_0 = \mathbf{y}(\mathbf{y}_0, 0)$. An equilibrium point is ***asymptotically stable*** if it is both stable and attractive, and simply ***unstable*** if it is not stable.

Translated to plain language, it means that the equilibrium point is stable if, for all particles starting inside a ball centered at the equilibrium point with a small radius of η, their subsequent trajectories will be contained within another ball with a small radius of ϵ at all times, whereas it is attractive if all particles starting

[3] Note that $\eta(\epsilon) \leq \epsilon$ implicitly. The definitions here and the following discussion are largely based on this Math Stackexchange post (4453028).

Figure 17.4.: *Field lines for the non-linear dynamical system (17.13).*

inside some ball with a small radius of η will converge to the equilibrium point as the time goes to infinity. It may be tempting to say that an attractive equilibrium point will be automatically (asymptotically) stable but there are some subtle differences between these two, demonstrated in the following counter-example. Consider the non-linear two-dimensional dynamical system

$$\begin{cases} \dfrac{dx}{dt} = x^2 - y^2 \\ \dfrac{dy}{dt} = 2xy \end{cases} \tag{17.13}$$

The pattern of its field lines is shown in Figure 17.4. The field lines are akin to the *magnetic dipole* type. The equilibrium point at the origin is attractive as all outgoing field lines will return to the equilibrium point, but is also unstable since we can always find a more outer field line that travels farther in both the x and y-directions.[4] Nevertheless, all stable equilibrium types in Table 17.1 except centers, are attractive and hence asymptotically stable.

[4]There is a caveat to this "counter-example" though. The "central" field line (not explicitly drawn out) that starts from $y = 0$ to the right of the equilibrium point will keep accelerating along the x-axis and diverges so it is only *almost* attractive. In fact, it is impossible to have

17.2.2. Linearization and Hartman-Grobman Theorem

Since non-linear dynamical systems are more complex, we may want to reduce them to the more familiar, linear dynamical systems. Motivated by the form (17.9), we seek to approximate the behavior of orbits near a non-linear equilibrium point by

$$\frac{d(\delta y_i(t))}{dt} = \sum_{j=1}^{n} A_{ij}(\delta y_j(t)) = \sum_{j=1}^{n} \alpha_j^{(i)}(\delta y_j(t)) \tag{17.14}$$

where δy_i represents a small displacement from the equilibrium point in the i-th direction, and each of the R.H.S. is a linear function of the displacement components with constant coefficients $\alpha_j^{(i)}$. Then, the previous (17.12) shows the fully written-out form of this approximation applicable to a two-dimensional dynamical system. Using Taylor Expansion to the first order from Multivariable Calculus, the non-linear dynamical system that takes the general form of (17.4), or written in component form

$$\frac{dy_i}{dt} = F_i(\mathbf{y}) \tag{17.15}$$

can be readily approximated by

$$\frac{d(y_i(t) - \tilde{y}_i)}{dt} \approx F_i(\tilde{\mathbf{y}}) + \sum_{j=1}^{n} \left(\left.\frac{\partial F_i}{\partial y_j}\right|_{\tilde{\mathbf{y}}} \right) (y_j(t) - \tilde{y}_j) + O(\delta {y_j}^2) \tag{17.16}$$

near the equilibrium point $\tilde{\mathbf{y}}$ where $\partial F_i/\partial y_j|_{\tilde{\mathbf{y}}}$ is the partial derivative of F_i along the j-th direction evaluated at the equilibrium point. The matrix formed by $\partial F_i/\partial y_j$ is known as the ***Jacobian (matrix)***. Due to Definition 17.1.1 of

an unstable, yet *truly* attractive equilibrium point if the dynamical system is smooth enough. However, the key takeaway is to help the readers distinguish between the concepts of stability and attractiveness, and this example should be adequate.

equilibrium points (17.5), $F_i(\tilde{\mathbf{y}}) = 0$, and if we further discard the higher-order error terms $O(\delta y_j{}^2)$ it becomes

$$\frac{d(\delta y_i(t))}{dt} = \sum_{j=1}^{n} \left(\left. \frac{\partial F_i}{\partial y_j} \right|_{\tilde{\mathbf{y}}} \right) \delta y_j(t) \tag{17.17}$$

Comparing with the proposed form in (17.14), we immediately identity the required coefficients $\alpha_j^{(i)}$ with $\partial F_i / \partial y_j|_{\tilde{\mathbf{y}}}$. For a two-dimensional non-linear dynamical system, the corresponding fully written-out form is then

$$\begin{cases} \dfrac{d(\delta x)}{dt} = \dfrac{\partial F_1}{\partial x}(\tilde{x}, \tilde{y})\delta x + \dfrac{\partial F_1}{\partial y}(\tilde{x}, \tilde{y})\delta y \\ \dfrac{d(\delta y)}{dt} = \dfrac{\partial F_2}{\partial x}(\tilde{x}, \tilde{y})\delta x + \dfrac{\partial F_2}{\partial y}(\tilde{x}, \tilde{y})\delta y \end{cases} \tag{17.18}$$

Now, the problem is whether such a ***linearization*** of the non-linear equilibrium point can successfully capture its characteristics. To proceed, we first introduce a class of equilibrium points known as ***hyperbolic equilibrium points***.

Definition 17.2.2 (Hyperbolic Equilibrium Points)**.** For a linear dynamical system as formulated by (17.9), a hyperbolic equilibrium point is an equilibrium point where all the eigenvalues of A have non-zero real parts. Similarly, for a non-linear dynamical system, a hyperbolic equilibrium point is an equilibrium point where all the eigenvalues of the Jacobian matrix $A_{ij} = \partial F_i / \partial y_j|_{\tilde{\mathbf{y}}}$ in the linearization by (17.17) have non-zero real parts, evaluated at the equilibrium point.

By this criteria, in Table 17.1, the stable/unstable nodes, saddle points, and stable/unstable spirals are all hyperbolic equilibrium points. The only exception is the centers, in addition to other omitted degenerate cases where at least one of the eigenvalues is strictly zero. Since stable equilibrium types excluding centers are all asymptotically stable as noted at the end of the last subsection, hyperbolic equilibrium points are either asymptotically stable or unstable. Now we are ready to examine the major result about hyperbolic equilibrium points, the

Hartman-Grobman (Linearization) Theorem, which states that the qualitative behavior of orbits near a non-linear hyperbolic equilibrium point will be the same as its linearization.

Theorem 17.2.3 (Hartman-Grobman Theorem)**.** A hyperbolic equilibrium point $\tilde{\mathbf{y}}$ of any non-linear dynamical system that has a Jacobian of $\partial F_i/\partial y_j|_{\tilde{\mathbf{y}}}$, belongs to the same type and has the same qualitative behavior in its neighborhood as the corresponding linear hyperbolic equilibrium point, with the matrix A in (17.9) equal to this Jacobian $\partial F_i/\partial y_j|_{\tilde{\mathbf{y}}}$.

Unfortunately, for an equilibrium point that is not hyperbolic, we need other methods to determine its behavior which are out of the scope of this book. Anyway, let's see how the Hartman-Grobman Theorem can be applied in the following example.

Example 17.2.1. Find all the equilibrium points for the two-dimensional non-linear dynamical system

$$\begin{cases} \dfrac{dx}{dt} = y - xy \\ \dfrac{dy}{dt} = x - xy \end{cases}$$

Solution. Since the dynamical system is now non-linear we can have multiple isolated equilibrium points. A simple rewriting of the dynamical system shows that

$$\begin{cases} \dfrac{dx}{dt} = y(1-x) \\ \dfrac{dy}{dt} = x(1-y) \end{cases}$$

from which we can easily deduce the two equilibrium points: $(\tilde{x}, \tilde{y}) = (0, 0)$ or

$(1, 1)$. The Jacobian of the non-linear dynamical system is

$$\begin{aligned}
\frac{\partial F_i}{\partial y_j} &= \begin{bmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} \\ \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\partial (y - xy)}{\partial x} & \dfrac{\partial (y - xy)}{\partial y} \\ \dfrac{\partial (x - xy)}{\partial x} & \dfrac{\partial (x - xy)}{\partial y} \end{bmatrix} \\
&= \begin{bmatrix} -y & 1 - x \\ 1 - y & -x \end{bmatrix}
\end{aligned}$$

At the first equilibrium point $(\tilde{x}, \tilde{y}) = (0, 0)$, the Jacobian is

$$\left.\frac{\partial F_i}{\partial y_j}\right|_{(0,0)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

the characteristic polynomial of which is simply $\lambda^2 - 1 = (\lambda - 1)(\lambda + 1)$ and thus it has two eigenvalues of -1 and 1, one positive and one negative. By the Hartman-Grobman Theorem, this means that the first equilibrium point $(\tilde{x}, \tilde{y}) = (0, 0)$ is locally a saddle point. On the other hand, at the second equilibrium point $(\tilde{x}, \tilde{y}) = (1, 1)$, the Jacobian is

$$\left.\frac{\partial F_i}{\partial y_j}\right|_{(1,1)} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

which clearly has a repeated negative eigenvalue of -1 with a multiplicity of 2, and thus by the Hartman-Grobman Theorem it is locally a sink (stable node). The phase portrait of this dynamical system is shown in Figure 17.5. □

Figure 17.5.: *The stream plot for the dynamical system in Example 17.2.1.*

17.3. Earth Science Applications: the Lorenz-63 Model

The ***Lorenz-63 Model*** is consisted of three variables X, Y, Z and their evolutions in time are described by

$$\begin{cases} \dfrac{dX}{dt} = -\sigma X + \sigma Y & \text{(17.19a)} \\ \dfrac{dY}{dt} = -XZ + \rho X - Y & \text{(17.19b)} \\ \dfrac{dZ}{dt} = XY - \beta Z & \text{(17.19c)} \end{cases}$$

where the parameters $\sigma, \rho, \beta > 0$ are all positive. It is a minimalist model that demonstrates the phenomenon of *chaos*. In the following subsections, we will

discuss the context in which the model is formulated and the characteristics of its equilibrium points.

17.3.1. Derivation of the Lorenz-63 Model

Reference Materials: [9]

The Lorenz-63 model originates from a simplified version of the *Rayleigh-Barnard Convection* problem about the convective/diffusion-driven flow. Particularly, we will restrict ourselves to a two-dimensional flow in the x/z-(zonal/meridional) directions so that the velocity vector is $\vec{u} = (u_x, 0, u_z)^T$. The governing equations for this scenario are the zonal/meridional *momentum equations*, *thermodynamics equation*, and the *continuity equation* with *Boussinesq approximation* as follows.

$$\begin{cases} \dfrac{\partial u_x}{\partial t} + u_x \dfrac{\partial u_x}{\partial x} + u_z \dfrac{\partial u_x}{\partial z} = \nu(\dfrac{\partial^2 u_x}{\partial x^2} + \dfrac{\partial^2 u_x}{\partial z^2}) - \dfrac{1}{\rho_0}\dfrac{\partial p}{\partial x} & (17.20) \\ \dfrac{\partial u_z}{\partial t} + u_x \dfrac{\partial u_z}{\partial x} + u_z \dfrac{\partial u_z}{\partial z} = \nu(\dfrac{\partial^2 u_z}{\partial x^2} + \dfrac{\partial^2 u_z}{\partial z^2}) - \dfrac{1}{\rho_0}\dfrac{\partial p}{\partial z} + \alpha T g & (17.21) \\ \dfrac{\partial T}{\partial t} - u_z \Gamma + u_x \dfrac{\partial T}{\partial x} + u_z \dfrac{\partial T}{\partial z} = \kappa(\dfrac{\partial^2 T}{\partial x^2} + \dfrac{\partial^2 T}{\partial z^2}) & (17.22) \\ \dfrac{\partial u_x}{\partial x} + \dfrac{\partial u_z}{\partial z} = 0 & (17.23) \end{cases}$$

The constants are: g, the gravitational acceleration; α, the thermal expansion coefficient of the fluid; the lapse rate $\Gamma = -\frac{T_H - T_L}{d}$ which is the difference in the temperature prescribed at the top and bottom plates over the vertical depth of the fluid d; ρ_0, the reference fluid density; ν and κ, the kinematic viscosity and thermal diffusivity. In addition to the velocities u_x and u_z, there are two other variables: T as the departure of temperature from the reference profile and p as

the pressure perturbation. By the ***Buckingham Pi Theorem***[5], we can convert these equations into dimensionless form using 9 non-dimensional variables:

$$\begin{cases} x^* = \dfrac{x}{d} & (17.24a) \\ z^* = \dfrac{z}{d} & (17.24b) \\ t^* = \dfrac{t}{d^2/\kappa} & (17.24c) \\ u_x^* = \dfrac{u_x}{\kappa/d} & (17.24d) \\ u_z^* = \dfrac{u_z}{\kappa/d} & (17.24e) \\ T^* = \dfrac{T}{\kappa\nu/g\alpha d^3} & (17.24f) \\ p^* = \dfrac{p}{\rho_0\kappa^2/d^2} & (17.24g) \\ \sigma = \dfrac{\nu}{\kappa} & (17.24h) \\ R = \dfrac{g\alpha d^3(T_H - T_L)}{\kappa\nu} & (17.24i) \end{cases}$$

The first 7 of them are the non-dimensional (indicated by the asterisk *) physical variables of displacements, time, velocities, temperature, and pressure with the characteristic length/time/velocity/temperature/pressure scales of d, d^2/κ (the *dissipation timescale*), κ/d, $\kappa\nu/g\alpha d^3$, and $\rho_0\kappa^2/d^2$. The last two of them

[5]It states that all physically consistent laws are "unit-independent" and the number of non-dimensional variables required to transform a set of equations into dimensionless form is the number of physical variables minus the number of fundamental physical dimensions. There are 13 physical variables here, time t, two displacements x and z, two velocities u_x and u_z, temperature T, kinematic viscosity ν, thermal diffusivity κ, pressure p, density ρ_0, thermal expansion coefficient α and gravity g (together counted as a single variable as they appear in the equation once and only once in combination over the term $\alpha T g$), and finally two other variables, the vertical depth d and temperature difference $T_H - T_L$ (derived from lapse rate Γ). Meanwhile, there are 4 fundamental dimensions: length, time, mass, and temperature. So there will be $13 - 4 = 9$ non-dimensional variables.

are the *Prandlt number* $\sigma = \frac{\nu}{\kappa}$ as the ratio of kinematic viscosity ν to thermal diffusivity κ, and *Rayleigh number* $R = \frac{g\alpha d^3(T_H - T_L)}{\kappa\nu}$. As we will see later, these two parameters will decide the behavior of the system. By substituting all these into (17.20), we have

$$\begin{aligned}
&\left(\frac{\kappa/d}{d^2/\kappa}\right)\frac{\partial u_x^*}{\partial t^*} + \left(\frac{(\kappa/d)^2}{d}\right)u_x^*\frac{\partial u_x^*}{\partial x^*} + \left(\frac{(\kappa/d)^2}{d}\right)u_z^*\frac{\partial u_x^*}{\partial z^*} \\
&= \nu\left(\left(\frac{\kappa/d}{d^2}\right)\frac{\partial^2 u_x^*}{\partial x^{*2}} + \left(\frac{\kappa/d}{d^2}\right)\frac{\partial^2 u_x^*}{\partial z^{*2}}\right) - \left(\frac{\kappa^2/d^2}{d}\right)\frac{\partial p^*}{\partial x^*}
\end{aligned}$$

Dividing both sides of the equation by the factor $\frac{\kappa^2}{d^3}$ leads to

$$\begin{aligned}
\frac{\partial u_x^*}{\partial t^*} + u_x^*\frac{\partial u_x^*}{\partial x^*} + u_z^*\frac{\partial u_x^*}{\partial z^*} &= \frac{\nu}{\kappa}\left(\frac{\partial^2 u_x^*}{\partial x^{*2}} + \frac{\partial^2 u_x^*}{\partial z^{*2}}\right) - \frac{\partial p^*}{\partial x^*} \\
&= \sigma\left(\frac{\partial^2 u_x^*}{\partial x^{*2}} + \frac{\partial^2 u_x^*}{\partial z^{*2}}\right) - \frac{\partial p^*}{\partial x^*}
\end{aligned} \tag{17.25}$$

Similarly, by noticing that $\alpha T g = \frac{\kappa\nu}{d^3}T^*$, we can rewrite (17.21) into

$$\frac{\partial u_z^*}{\partial t^*} + u_x^*\frac{\partial u_z^*}{\partial x^*} + u_z^*\frac{\partial u_z^*}{\partial z^*} = \sigma\left(\frac{\partial^2 u_z^*}{\partial x^{*2}} + \frac{\partial^2 u_z^*}{\partial z^{*2}}\right) - \frac{\partial p^*}{\partial z^*} + \sigma T^* \tag{17.26}$$

Meanwhile, (17.22) can be transformed into

$$\begin{aligned}
&\left(\frac{\kappa\nu/g\alpha d^3}{d^2/\kappa}\right)\frac{\partial T^*}{\partial t^*} + (\kappa/d)u_z^*\left(\frac{T_L - T_H}{d}\right) \\
&+ \left(\frac{(\kappa/d)(\kappa\nu/g\alpha d^3)}{d}\right)u_x^*\frac{\partial T^*}{\partial x^*} + \left(\frac{(\kappa/d)(\kappa\nu/g\alpha d^3)}{d}\right)u_z^*\frac{\partial T^*}{\partial z^*} \\
&= \kappa\left(\left(\frac{\kappa\nu/g\alpha d^3}{d^2}\right)\frac{\partial^2 T^*}{\partial x^{*2}} + \left(\frac{\kappa\nu/g\alpha d^3}{d^2}\right)\frac{\partial^2 T^*}{\partial z^{*2}}\right)
\end{aligned}$$

This time, dividing by the factor $\frac{\kappa^2\nu}{g\alpha d^5}$ gives

$$\frac{\partial T^*}{\partial t^*} + u_z^*\left(\frac{g\alpha d^3(T_H - T_L)}{\kappa\nu}\right) + u_x^*\frac{\partial T^*}{\partial x^*} + u_z^*\frac{\partial T^*}{\partial z^*}$$

$$= \frac{\partial T^*}{\partial t^*} + Ru_z^* + u_x^* \frac{\partial T^*}{\partial x^*} + u_z^* \frac{\partial T^*}{\partial z^*} = \frac{\partial^2 T^*}{\partial x^{*2}} + \frac{\partial^2 T^*}{\partial z^{*2}} \tag{17.27}$$

Finally, the non-dimensional form of the mass continuity equation (17.23) is simply

$$\frac{\partial u_x^*}{\partial x^*} + \frac{\partial u_z^*}{\partial z^*} = 0 \tag{17.28}$$

According to this, we can rewrite $\vec{u}^* = (u_x^*, 0, u_z^*)^T = (-\frac{\partial \psi^*}{\partial z^*}, 0, \frac{\partial \psi^*}{\partial x^*})$ where ψ^* is known as the (non-dimensional) ***stream function***. This is possible as the problem is set up to be two-dimensional and we will further require it to be bounded, such that $\vec{u} = \mathbf{0}$ and $T = 0$ ($\vec{u}^* = \mathbf{0}$ and $T^* = 0$) at $z = 0, d$ ($z^* = 0, 1$) and the boundary condition along the x-direction is periodic. With

$$\begin{cases} u_x^* = -\dfrac{\partial \psi^*}{\partial z^*} & \text{(17.29a)} \\ u_z^* = \dfrac{\partial \psi^*}{\partial x^*} & \text{(17.29b)} \end{cases}$$

(17.28) is then automatically satisfied.[6] Now substituting (17.29) into (17.25) and (17.26), and writing $\frac{\partial}{\partial[\,]} = \partial_{[\,]}$ for simplicity, we have

$$\begin{aligned} &\partial_{t^*}\partial_{z^*}\psi^* - (\partial_{z^*}\psi^*)(\partial_{x^*}\partial_{z^*}\psi^*) + (\partial_{x^*}\psi^*)((\partial_{z^*})^2\psi^*) \\ &= \sigma\left((\partial_{x^*})^2\partial_{z^*}\psi^* + (\partial_{z^*})^3\psi^*\right) + \partial_{x^*}p^* \end{aligned} \tag{17.30}$$

and

$$\begin{aligned} &\partial_{t^*}\partial_{x^*}\psi^* - (\partial_{z^*}\psi^*)((\partial_{x^*})^2\psi^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}\partial_{x^*}\psi^*) \\ &= \sigma\left((\partial_{x^*})^3\psi^* + (\partial_{z^*})^2\partial_{x^*}\psi^*\right) - \partial_{z^*}p^* + \sigma T^* \end{aligned} \tag{17.31}$$

Differentiating (17.30) with respect to z and (17.31) with respect to x, and then adding them together will eliminate the pressure term and lead to an L.H.S. of

$$[\partial_{t^*}(\partial_{z^*})^2\psi^* - ((\partial_{z^*})^2\psi^*)(\partial_{x^*}\partial_{z^*}\psi^*) - (\partial_{z^*}\psi^*)(\partial_{x^*}(\partial_{z^*})^2\psi^*)$$

[6] $\frac{\partial u_x^*}{\partial x^*} + \frac{\partial u_z^*}{\partial z^*} = -\frac{\partial \psi^*}{\partial x^* \partial z^*} + \frac{\partial \psi^*}{\partial x^* \partial z^*} \equiv 0.$

$$
\begin{aligned}
&+ (\partial_{z^*}\partial_{x^*}\psi^*)((\partial_{z^*})^2\psi^*) + (\partial_{x^*}\psi^*)((\partial_{z^*})^3\psi^*)] \\
&+ [\partial_{t^*}(\partial_{x^*})^2\psi^* - (\partial_{x^*}\partial_{z^*}\psi^*)((\partial_{x^*})^2\psi^*) - (\partial_{z^*}\psi^*)((\partial_{x^*})^3\psi^*) \\
&+ ((\partial_{x^*})^2\psi^*)(\partial_{z^*}\partial_{x^*}\psi^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}(\partial_{x^*})^2\psi^*)] \\
= \; &\partial_{t^*}\left((\partial_{x^*})^2 + (\partial_{z^*})^2\right)\psi^* - (\partial_{z^*}\psi^*)(\partial_{x^*}\left((\partial_{x^*})^2 + (\partial_{z^*})^2\right)\psi^*) \\
&+ (\partial_{x^*}\psi^*)(\partial_{z^*}\left((\partial_{x^*})^2 + (\partial_{z^*})^2\right)\psi^*)
\end{aligned}
$$

and an R.H.S. of

$$
\begin{aligned}
&[\sigma\left((\partial_{x^*})^2(\partial_{z^*})^2\psi^* + (\partial_{z^*})^4\psi^*\right) + \partial_{z^*}\partial_{x^*}p^*] \\
&+ [\sigma\left((\partial_{x^*})^4\psi^* + (\partial_{z^*})^2(\partial_{x^*})^2\psi^*\right) - \partial_{x^*}\partial_{z^*}p^* + \sigma\partial_{x^*}T^*] \\
= \; &\sigma\left((\partial_{x^*})^2 + (\partial_{z^*})^2\right)^2\psi^* + \sigma\partial_{x^*}T^*
\end{aligned}
$$

By using the *Laplacian* symbol $\Delta^* = (\partial_{x^*})^2 + (\partial_{z^*})^2$ in this equation, and also substituting the stream function (17.29) into the other equation (17.27), we can sort out a set of two equations in two unknowns ψ^* and T^*, that describes the evolution of the system state:

$$
\begin{cases}
\partial_{t^*}\Delta^*\psi^* - (\partial_{z^*}\psi^*)(\partial_{x^*}\Delta^*\psi^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}\Delta^*\psi^*) = \sigma\Delta^{*2}\psi^* + \sigma\partial_{x^*}T^* & (17.32) \\
\partial_{t^*}T^* - (\partial_{z^*}\psi^*)(\partial_{x^*}T^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}T^*) = \Delta^*T^* - R\partial_{x^*}\psi^* & (17.33)
\end{cases}
$$

As mentioned before, the statistics of the system are controlled by the two parameters σ and R here. However, we still have to simplify the problem to make it tractable while retaining the chaotic nature. For this, we will decompose the variables T^* and ψ^* into Fourier modes along the zonal and meridional directions. Since the boundary conditions are periodic and fixed to zero in the x and z-directions respectively, we assume that the solution will take the following form of infinite summations of basis functions in terms of products between

complex exponentials and sines (can refer to Section 14.1):

$$\begin{cases} T^*(x^*, z^*, t^*) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=1}^{\infty} \hat{T}_{k_1,k_2}(t^*) e^{ik_1\pi a x^*} \sin(k_2 \pi z^*) & (17.34a) \\ \psi^*(x^*, z^*, t^*) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=1}^{\infty} \hat{\psi}_{k_1,k_2}(t^*) e^{ik_1\pi a x^*} \sin(k_2 \pi z^*) & (17.34b) \end{cases}$$

where $a = d/l$ is the aspect ratio (depth to width) of the domain[7]. To proceed, we first need to understand that to generate various Fourier basis functions with new wave numbers, we have to rely on the non-linearity due to the interaction of flow with different wave numbers (scales). Otherwise, the orthogonality of the Fourier modes will prevent this from happening over linear terms. Heuristically, if we look at, for example, the portion of the complex exponentials in x^* for two ψ^* basis functions, $\hat{\psi}_k e^{ik\pi a x^*}$ and $\hat{\psi}_l e^{il\pi a x^*}$, then their product (as well as with one of them conjugated) will produce

$$\begin{cases} \psi_k^* \psi_l^* = \hat{\psi}_k \hat{\psi}_l e^{i(k+l)\pi a x^*} & (17.35a) \\ \psi_k^* \overline{\psi_l^*} = \hat{\psi}_k \overline{\hat{\psi}_l} e^{i(k-l)\pi a x^*} & (17.35b) \end{cases}$$

So the new signal can have either a wave number of $k + l$ and $k - l$. If particularly one of the original wave numbers k and l is 1, then by inductively using this property we can essentially generate all wave numbers and hence the full dynamics. Therefore, we are motivated to focus on the Fourier modes with a wave number of 1. Again we will first look at the x^* portion of the Fourier modes and we will choose the imaginary part of $\hat{\psi}_{k_1=1}$ which is $\hat{\psi}_{k_1=1} \sin(\pi a x^*)$ as the starting point. We will drop the non-linear advection term $-(\partial_{z^*}\psi^*)(\partial_{x^*}\Delta^*\psi^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}\Delta^*\psi^*)$ in the prognostic equation (17.32) of ψ^* to trim down the model. This makes the $\sigma\partial_{x^*}T^*$ term on R.H.S. become the only way to keep the non-linearity (the remaining $\sigma\Delta^{*2}\psi^*$ is only a linear damping term). For this $\sigma\partial_{x^*}T^*$ term to contribute to the tendency of $\hat{\psi}_{k_1=1} \sin(\pi a x^*)$

[7] $\frac{z}{d} = z^*$, $\frac{x}{l} = \frac{dx^*}{l} = ax^*$.

in the time derivative term $\partial_{t^*}\Delta^*\psi^*$ on L.H.S., we must also have the real part of $\hat{T}_{k_1=1}$ which is $\hat{T}_{k_1=1}\cos(\pi a x^*)$, due to the form of $\sigma\partial_{x^*}T^*$ which upon substitution becomes $-\sigma\pi a\hat{T}_{k_1=1}\sin(\pi a x^*)$, so that it is compatible with the sine part of $\hat{\psi}_{k_1=1}\sin(\pi a x^*)$ with a wave number of 1.

Since we have removed the advection term in the prognostic equation of ψ^*, we have to keep the advection term $-(\partial_{z^*}\psi^*)(\partial_{x^*}T^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}T^*)$ in the prognostic equation of T^* to ensure the non-linearity of the whole dynamical system. By the previous argument over (17.35), as long as either the sum of or difference in wave numbers k_1 and k_2 between the two scales is 1 ($k_1 + k_2 = 1$ or $k_1 - k_2 = 1$), then their non-linear interaction can generate the required wave number 1 signal (the real part of $\hat{T}_{k_1=1}$). One of the simplest ways to achieve this without involving too many new scales is to introduce the real part of the mean temperature $\hat{T}_{k_1=0}$ which is a constant in x^*, so that the forcing term for $\hat{T}_{k_1=1}\cos(\pi a x^*)$ is produced in (17.33) as we already have $\hat{\psi}_{k_1=1}$. Physically, it represents the non-linear interaction between the mean temperature field and the zonal wave number 1 flow. So we come to the third variable $\hat{T}_{k_1=0}$ finally. To verify if we can close the equations so that there are enough forcing terms for all three included wave numbers, we can check if the non-linear interaction between the imaginary part of $\hat{\psi}_{k_1=1}$ and the real part of $\hat{T}_{k_1=1}$ indeed brings us $\hat{T}_{k_1=0}$. Substituting the relevant expressions into the advection term $-(\partial_{z^*}\psi^*)(\partial_{x^*}T^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}T^*)$ in (17.33), we have (where we can ignore any part related to z^* for now)

$$
\begin{aligned}
&-(\partial_{z^*}\psi^*)(\partial_{x^*}T^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}T^*) \\
=& -\hat{\psi}_{k_1=1}\sin(\pi a x^*)\partial_{x^*}(\hat{T}_{k_1=1}\cos(\pi a x^*)) \\
&+ \partial_{x^*}(\hat{\psi}_{k_1=1}\sin(\pi a x^*))\hat{T}_{k_1=1}\cos(\pi a x^*) \\
=& -\hat{\psi}_{k_1=1}\sin(\pi a x^*)(-\pi a\hat{T}_{k_1=1}\sin(\pi a x^*)) \\
&+ (\pi a\hat{\psi}_{k_1=1}\cos(\pi a x^*))\hat{T}_{k_1=1}\cos(\pi a x^*) \\
=& \pi a\hat{\psi}_{k_1=1}\hat{T}_{k_1=1}(\sin^2(\pi a x^*) + \cos^2(\pi a x^*)) = \pi a\hat{\psi}_{k_1=1}\hat{T}_{k_1=1}
\end{aligned}
$$

which indeed gives the zonally constant temperature forcing as desired. We

repeat the same analysis for the z portion of the Fourier modes and similarly start with $\hat{\psi}_{k_2=1}\sin(\pi z^*)$ and $\hat{T}_{k_2=1}\sin(\pi z^*)$ of the same wave numbers. This time we will need the meridional wave number 2 temperature field and we will check following the same logic above:

$$
\begin{aligned}
&-(\partial_{z^*}\psi^*)(\partial_{x^*}T^*) + (\partial_{x^*}\psi^*)(\partial_{z^*}T^*) \\
&= -(\pi\hat{\psi}_{k_2=1}\cos(\pi z^*))((-)\hat{T}_{k_2=1}\sin(\pi z^*)) \\
&\quad + \hat{\psi}_{k_2=1}\sin(\pi z^*)(\pi\hat{T}_{k_2=1}\cos(\pi z^*)) \\
&\quad \text{(Notice that the differentiation of the real part of } \hat{T}_{k_1=1} \text{ with respect to } x \\
&\quad \text{in the first term will contribute an extra negative sign as above)} \\
&= 2\pi\hat{\psi}_{k_2=1}\hat{T}_{k_2=1}\sin(\pi z^*)\cos(\pi z^*) = \pi\hat{\psi}_{k_2=1}\hat{T}_{k_2=1}\sin(2\pi z^*)
\end{aligned}
$$

Now we will combine all of these where the needed three variables are the imaginary part of $\hat{\psi}_{k_1=1,k_2=1}$, and the real parts of $\hat{T}_{k_1=1,k_2=1}$ and $\hat{T}_{k_1=0,k_2=2}$ respectively. Substituting them into (17.32), we have

$$
\begin{aligned}
&\partial_{t^*}\Delta^*(\hat{\psi}_{k_1=1,k_2=1}\sin(\pi a x^*)\sin(\pi z^*)) \\
&= \sigma\Delta^{*2}(\hat{\psi}_{k_1=1,k_2=1}\sin(\pi a x^*)\sin(\pi z^*)) \\
&\quad + \sigma\partial_{x^*}((\hat{T}_{k_1=1,k_2=1}\cos(\pi a x^*)\sin(\pi z^*)) + (\hat{T}_{k_1=0,k_2=2}\sin(2\pi z^*)))
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow \quad & -\partial_{t^*}(\hat{\psi}_{k_1=1,k_2=1}\sin(\pi a x^*)\sin(\pi z^*)) \\
&= \sigma(\hat{\psi}_{k_1=1,k_2=1}\sin(\pi a x^*)\sin(\pi z^*)) - \sigma(\hat{T}_{k_1=1,k_2=1}\sin(\pi a x^*)\sin(\pi z^*)) \\
\Longrightarrow \quad & \partial_{t^*}\hat{\psi}_{k_1=1,k_2=1} = \sigma\hat{T}_{k_1=1,k_2=1} - \sigma\hat{\psi}_{k_1=1,k_2=1} \qquad (17.36)
\end{aligned}
$$

where we have ignored any scaling constant arising from differentiation except the negative sign, and the only effect of the Laplacian Δ is hence to introduce an extra negative sign.[8] Replacing $\hat{\psi}_{k_1=1,k_2=1}, \hat{T}_{k_1=1,k_2=1}, \hat{T}_{k_1=0,k_2=2}$ by X, Y, Z respectively then readily yields (17.19a). In the same essence, we can derive

[8] To illustrate, $\Delta^*((\sin(\pi a x^*)\sin(\pi z^*)) = ((\partial_{x^*})^2 + (\partial_{z^*})^2)((\sin(\pi a x^*)\sin(\pi z^*)) = -\pi^2 a^2(\sin(\pi a x^*)\sin(\pi z^*)) - \pi^2(\sin(\pi a x^*)\sin(\pi z^*)) = -(\pi^2 a^2 + \pi^2)(\sin(\pi a x^*)\sin(\pi z^*))$.

the other two equations (17.19b) and (17.19c) of the Lorenz-63 Model by substitution into (17.33) with some adjustments to the parameters.[9] Now the ρ in (17.19b) represents a rescaling of the Rayleigh number relative to the *critical* Rayleigh number which dictates the initialization of the convection and β in (17.19c) comes from the aspect ratio. With (17.19) derived properly, we are ready to examine the nature of its dynamics.

17.3.2. Stability Analysis of Each Equilibrium Point

Reference Materials: [3]

We will inspect the stability of each equilibrium point of the Lorenz-63 Model and thus its behavior under different parameter settings. First, from (17.19) it is obvious that the origin $\mathbf{0} = (0, 0, 0)^T$ is an equilibrium point. The other equilibrium points can be found by first noticing that (17.19a) requires $\tilde{X} = \tilde{Y}$. Substituting this into (17.19b) and (17.19c) gives

$$\begin{cases} -\tilde{X}\tilde{Z} + \rho\tilde{X} - \tilde{X} = 0 \implies \tilde{Z} = \rho - 1 \\ \tilde{X}^2 - \beta\tilde{Z} = 0 \end{cases}$$

Combining these two equations leads to

$$\tilde{X}^2 = \beta(\rho - 1) \implies \tilde{X} = \pm\sqrt{\beta(\rho - 1)}$$

Therefore the other two equilibrium points are

$$\tilde{\mathbf{X}}_{\pm} = (\pm\sqrt{\beta(\rho - 1)}, \pm\sqrt{\beta(\rho - 1)}, \rho - 1)^T \tag{17.37}$$

which only emerge when $\rho > 1$ due to the square roots in the X and Y components. Therefore, we can divide the system into two major cases: $0 < \rho < 1$ and

[9] Unfortunately, the algebra is too tedious to be included in the text. Any ambitious reader can try it though.

$\rho > 1$.[10] When $0 < \rho < 1$, the origin is the only equilibrium point. According to (17.17)), the linearized system will have a Jacobian of

$$\frac{\partial F_i}{\partial X_j} = \begin{bmatrix} \frac{\partial F_1}{\partial X} & \frac{\partial F_1}{\partial Y} & \frac{\partial F_1}{\partial Z} \\ \frac{\partial F_2}{\partial X} & \frac{\partial F_2}{\partial Y} & \frac{\partial F_2}{\partial Z} \\ \frac{\partial F_3}{\partial X} & \frac{\partial F_3}{\partial Y} & \frac{\partial F_3}{\partial Z} \end{bmatrix} = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho - Z & -1 & -X \\ Y & X & -\beta \end{bmatrix} \tag{17.38}$$

At the origin, it is

$$\left.\frac{\partial F_i}{\partial X_j}\right|_{\mathbf{0}} = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{bmatrix} \tag{17.39}$$

the characteristic polynomial of which is

$$\begin{aligned} \det\left(\left.\frac{\partial F_i}{\partial X_j}\right|_{\mathbf{0}} - \lambda I\right) &= \begin{vmatrix} -\sigma - \lambda & \sigma & 0 \\ \rho & -1 - \lambda & 0 \\ 0 & 0 & -\beta - \lambda \end{vmatrix} \\ &= -(\lambda + \beta)((-\sigma - \lambda)(-1 - \lambda) - \rho\sigma) \\ &= -(\lambda + \beta)(\lambda^2 + (1 + \sigma)\lambda - (\rho - 1)\sigma) \end{aligned} \tag{17.40}$$

The roots are $\lambda_1 = -\beta$ and

$$\begin{aligned} \lambda_\pm &= \frac{1}{2}\left(-(1 + \sigma) \pm \sqrt{(1 + \sigma)^2 - (-4(\rho - 1)\sigma)}\right) \\ &= \frac{1}{2}\left(-(1 + \sigma) \pm \sqrt{(1 + \sigma)^2 + (4\rho\sigma - 4\sigma)}\right) = \frac{1}{2}(-(1 + \sigma) \pm \sqrt{\mathcal{D}}) \end{aligned} \tag{17.41}$$

[10] $\rho = 1$, as well as $\rho = \rho_c$ that we will see later, belongs to the special case of a *bifurcation point* where transition happens and their treatment is out of the scope of this book.

by the quadratic equation formula where $\mathcal{D}$ denotes the discriminant. When $0 < \rho < 1$,

$$\begin{aligned}\sqrt{\mathcal{D}} &= \sqrt{(1+\sigma)^2 + (4\rho\sigma - 4\sigma)} \\ &< \sqrt{(1+\sigma)^2} = 1 + \sigma\end{aligned}$$

and therefore the two other roots $\lambda_\pm = \frac{1}{2}(-(1+\sigma) \pm \sqrt{\mathcal{D}}) < 0$ must also be negative. By the Hartman-Grobman Theorem 17.2.3, the origin is hence a hyperbolic equilibrium point which is locally a stable node. Therefore, when $0 < \rho < 1$, the entire Lorenz-63 Model is stable, which corresponds to the physical situation where the Rayleigh number does not reach the critical threshold and convection does not arise. However, when $\rho > 1$, by the same logic, now $\sqrt{D} > 1 + \sigma$ and one of the roots $\lambda_+ > 0$ will be positive, and the origin then becomes a saddle point. At the same time, as $\rho > 1$, there will appear two new equilibrium points indicated by (17.37) and we are going to determine their types as well. However, it is instructive to first note that we only need to consider either of them because the Lorenz-63 Model itself, in addition to these two equilibrium points, is symmetric in the X and Y directions.[11] We will take the positive equilibrium point $\tilde{\mathbf{X}}_+ = (\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1)^T$. The Jacobian of linearization at this equilibrium point is found by substituting it into (17.38), which gives

$$\left.\frac{\partial F_i}{\partial X_j}\right|_{\tilde{\mathbf{X}}_+} = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho - (\rho-1) & -1 & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & \sqrt{\beta(\rho-1)} & -\beta \end{bmatrix}$$

[11] Replacing X by $-X$ and Y by $-Y$ in (17.19) gives

$$\begin{cases} \dfrac{d(-X)}{dt} = -\sigma(-X) + \sigma(-Y) & \Longrightarrow \dfrac{dX}{dt} = -\sigma X + \sigma Y \\ \dfrac{d(-Y)}{dt} = -(-X)Z + \rho(-X) - (-Y) & \Longrightarrow \dfrac{dY}{dt} = -XZ + \rho X - Y \\ \dfrac{dZ}{dt} = (-X)(-Y) - \beta Z & \Longrightarrow \dfrac{dZ}{dt} = XY - \beta Z \end{cases}$$

which is exactly the same as (17.19).

$$= \begin{bmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & \sqrt{\beta(\rho-1)} & -\beta \end{bmatrix} \tag{17.42}$$

and its characteristic polynomial is

$$\begin{aligned}
&\det\left(\left.\frac{\partial F_i}{\partial X_j}\right|_{\tilde{\mathbf{X}}_+} - \lambda I \right) \\
&= \begin{vmatrix} -\sigma-\lambda & \sigma & 0 \\ 1 & -1-\lambda & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & \sqrt{\beta(\rho-1)} & -\beta-\lambda \end{vmatrix} \\
&= (-\sigma-\lambda)\begin{vmatrix} -1-\lambda & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & -\beta-\lambda \end{vmatrix} - \sigma \begin{vmatrix} 1 & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & -\beta-\lambda \end{vmatrix} \\
&= (-\sigma-\lambda)((-1-\lambda)(-\beta-\lambda) - (\sqrt{\beta(\rho-1)})(-\sqrt{\beta(\rho-1)}) \\
&\quad - \sigma((1)(-\beta-\lambda) - (\sqrt{\beta(\rho-1)})(-\sqrt{\beta(\rho-1)})) \\
&= (-\sigma-\lambda)(\beta + (\beta+1)\lambda + \lambda^2 + \beta(\rho-1)) - \sigma(-\beta-\lambda+\beta(\rho-1)) \\
&= (-\sigma-\lambda)(\beta\rho + (\beta+1)\lambda + \lambda^2) + \beta\sigma + \sigma\lambda - \beta\sigma\rho + \beta\sigma \\
&= (-\beta\sigma\rho - (\beta+1)\sigma\lambda - \sigma\lambda^2) + (-\beta\rho\lambda - (\beta+1)\lambda^2 - \lambda^3) \\
&\quad + \beta\sigma + \sigma\lambda - \beta\sigma\rho + \beta\sigma \\
&= -\lambda^3 - (\beta+\sigma+1)\lambda^2 - \beta(\sigma+\rho)\lambda - 2\beta\sigma(\rho-1)
\end{aligned} \tag{17.43}$$

by Cofactor Expansion along the first row. As the coefficients of the characteristic polynomial are all negative as long as $\rho > 1$, the roots must be negative if they are real.[12] So the problem is reduced to determine the nature of its roots. Particularly, we are interested in the case when there are one real negative root and two complex roots, and the real part of the two complex roots changes from negative to positive. The required critical value of $\rho = \rho_c$ where this transition happens can be found by considering purely imaginary roots $\lambda = i\mu$ as the real

[12] If any real root is positive, then $-\lambda^3 - (\beta+\sigma+1)\lambda^2 - \beta(\sigma+\rho)\lambda - 2\beta\sigma(\rho-1) < 0 \neq 0$.

part vanishes. Substituting this into (17.43) gives

$$\begin{aligned} &-(i\mu)^3 - (\beta+\sigma+1)(i\mu)^2 - \beta(\sigma+\rho_c)(i\mu) - 2\beta\sigma(\rho_c-1) \\ &= i\mu^3 + (\beta+\sigma+1)\mu^2 - i\beta(\sigma+\rho_c)\mu - 2\beta\sigma(\rho_c-1) \end{aligned} \tag{17.44}$$

By equating real and imaginary parts, we have

$$(\beta+\sigma+1)\mu^2 = 2\beta\sigma(\rho_c-1) \tag{17.45a}$$

$$\mu^3 = \beta(\sigma+\rho_c)\mu \tag{17.45b}$$

Combining these two equations (dividing (a) by (b)) then yields

$$\begin{aligned} \beta+\sigma+1 &= \frac{2\beta\sigma(\rho_c-1)}{\beta(\sigma+\rho_c)} \\ (\beta+\sigma+1)(\sigma+\rho_c) &= 2\sigma(\rho_c-1) \\ \rho_c(\beta+\sigma+1)+\sigma(\beta+\sigma+1) &= 2\sigma\rho_c - 2\sigma \\ \rho_c(\beta-\sigma+1) &= -\sigma(\beta+\sigma+3) \\ \rho_c &= \frac{\sigma(\beta+\sigma+3)}{\sigma-\beta-1} \end{aligned} \tag{17.46}$$

However, this does not tell us which side, $\rho < \rho_c$ or $\rho > \rho_c$, corresponds to a positive/negative real part of the complex roots. We will simply state the conclusion for that.[13] If $\rho < \rho_c$, the complex roots will have a negative real part. So by the Hartman-Grobman Theorem 17.2.3, the equilibrium points $\tilde{\mathbf{X}}_\pm$ will be stable, behaving as a one-dimensional manifold of a stable node and a two-dimensional manifold of a stable spiral combined together. Any particle will spiral inwards and be absorbed into either of the two equilibrium points. However, when $\rho > \rho_c$, the complex roots will now have a positive real part and by the same argument the equilibrium points $\tilde{\mathbf{X}}_\pm$ will be unstable locally, akin to a one-dimensional manifold of a stable node fused with a two-dimensional manifold of an unstable spiral. The effect is to cause the particle to gradually spiral outwards about one of the equilibrium points, and when it moves far

[13] Again, the algebra to explicitly show this is too horrifying to be written down.

enough away from it, it will be attracted and gravitate towards the another equilibrium point. However, as the particle get closer to it, the same scenario will occur again where it will now gradually spiral outwards about the new equilibrium point. This alternating pattern will keep going on and on, where the particle will revolve about and be flung between these two equilibrium points. Hence, as $\rho > \rho_c$, the Lorenz-63 Model is known as a ***strange attractor*** and ***chaos*** arises: the system is very sensitive to any small perturbation in the initial position of a particle and it can lead to great deviation in future trajectories (the ***Butterfly Effect***). Physically, the Rayleigh number is sufficiently high to produce chaos. Usually, we set $\sigma = 10$ and $\beta = \frac{8}{3}$. The critical value of ρ_c will then be

$$\rho_c = \frac{(10)\left(\frac{8}{3} + 10 + 3\right)}{10 - \frac{8}{3} - 1} \approx 24.74 \tag{17.47}$$

according to (17.46). The following two figures (Figures 17.6 and 17.7) then show the evolution of the system with $\rho = 14$ and $\rho = 28$[14], when ρ is under and above the critical ρ_c.

17.4. Python Programming

In this section, we will show the necessary codes to produce Figures 17.6 and 17.7. First, we need to use the `solve_ivp` function from `scipy.integrate` to integrate the dynamical system. We will define the equations of the Lorenz-63 Model as a function of t and $\mathbf{X}$ (which is denoted by the variable `y` in the argument).

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.integrate import solve_ivp
```

[14] $\rho = 28$ is a very common configuration picked to demonstrate the chaotic behavior.

Figure 17.6.: *The evolution of the Lorenz-63 Model when* $\sigma = 10, \beta = 8/3, \rho = 14$ *where the initial position is* $\boldsymbol{X} = (-10, 10, 30)^T$.

Figure 17.7.: *The evolution of the Lorenz-63 Model when* $\sigma = 10, \beta = 8/3, \rho = 28$ *where the initial position is* $\boldsymbol{X} = (5, -5, 20)^T$. *An extra of 15 particles (overlapping green dots) are generated nearby and their positions after integration over 2 units of time are shown as red dots.*

```
def Lorenz63(t, y, sigma=10, beta=8/3, rho=14):
    X, Y, Z = y[0], y[1], y[2]
    dXdt = -sigma*X + sigma*Y
    dYdt = -X*Z + rho*X - Y
    dZdt = X*Y - beta*Z
    return([dXdt, dYdt, dZdt])
```

Then we will set the period of integration and the initial position of the particle.

```
t_span = [0,25]
Y_0 = [-10,10,30]
```

Finally, we simply provide the defined `Lorenz63` function for `solve_ivp` to integrate.

```
sol14 = solve_ivp(Lorenz63, t_span, Y_0, t_eval=np.linspace(
    t_span[0], t_span[1], 10000))
```

where `t_eval` controls the frequency of sampling (10000 evenly spaced time steps here). To plot the results, we need to use the 3D projection and extract `sol.y`:

```
fig14 = plt.figure()
ax14 = plt.subplot(projection='3d')
ax14.set_title("Lorenz-63, " r"$\sigma=10, \beta=8/3, \rho=14$
    ")
ax14.set_xlabel("X")
ax14.set_ylabel("Y")
ax14.set_zlabel("Z")
ax14.scatter(*Y_0, c="g", label="Initial Position")
ax14.plot(*sol14.y)
ax14.legend()
fig14.savefig("Lorenz63_14", dpi=300, bbox_inches="tight")
```

To run this with different parameters, we can give `args` in `solve_ivp`. Now we set $\rho = 28$ to enable chaotic behavior. The asterisk * effectively expands the arrays into the three-directional components.

```
Y_0 = [5,-5,20]
```

```
sol28 = solve_ivp(Lorenz63, t_span, Y_0, t_eval=np.linspace(
    t_span[0], t_span[1], 10000), args=(10,8/3,28))

# Other plotting commands are the same and hence omitted.
ax28.scatter(*Y_0, c="g", label="Initial Position")
ax28.plot(*sol28.y)
```

We will add some perturbations to the initial position of the particle via Gaussian noises. 15 new particles are generated in this way.

```
n = 15
noises = np.random.multivariate_normal([0,0,0], np.diag
    ([0.1]*3), n)
Ys_perturb = noises + Y_0
```

We repeat the integration procedure for all particles. Their positions after 2 units of time are then plotted as well.

```
Ys_t2 = np.full(Ys_perturb.shape, np.nan)
for ii in np.arange(n):
    sol_p = solve_ivp(Lorenz63, [0,2], Ys_perturb[ii,:], args
        =(10,8/3,28))
    Ys_t2[ii,:] = sol_p.y[:,-1]

# Other plotting commands are the same and hence omitted.
#ax28.scatter(*Ys_perturb.T, c="g")
ax28.scatter(*Ys_t2.T, c="r", label="Positions after 2 Units
    of Time")
fig28.savefig("Lorenz63_28", dpi=300, bbox_inches="tight")
```

17.5. Exercises

Exercise 17.1 Determine the type of equilibrium point at the origin for the two-dimensional linear dynamical system $\mathbf{y}' = A\mathbf{y}$ where $A =$

(a) $\begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$;

(b) $\begin{bmatrix} 0 & -1 \\ 2 & 2 \end{bmatrix}$;

(c) $\begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}$;

(d) $\begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix}$

and find their general solution.

Exercise 17.2 Describe the equilibrium point for the three-dimensional linear dynamical system $\mathbf{y}' = A\mathbf{y} + \mathbf{G}$ where

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

Exercise 17.3 Find all equilibrium points for the following two-dimensional non-linear dynamical system

$$\begin{cases} \dfrac{dx}{dt} = x^2 - 2y^2 + xy \\ \dfrac{dy}{dt} = x^2 + y^2 - 1 \end{cases}$$

and discuss their stability.

Exercise 17.4 Examine the chaotic behavior of the *Rössler attractor*:

$$\begin{cases} \dfrac{dx}{dt} = -y - z & (17.48a) \\ \dfrac{dy}{dt} = x + ay & (17.48b) \\ \dfrac{dz}{dt} = b + z(x - c) & (17.48c) \end{cases}$$

with three parameters $a = 0.2$, $b = 0.2$, $c = 5.7$ and plot it.

Chapter 18

Index Notation and Introduction to Tensors

This is the final chapter of this book. (Congratulations!) It will be revealed that matrices (or to be accurate, linear transformations), along with vectors, actually can be generalized to a broader class of objects known as *tensors*. In the simplest sense, tensors are quantities with subscripts. To facilitate the manipulation of these subscripts, we will introduce the handy *Index Notation*, as well as the *Kronecker delta/epsilon symbols*. The next step is to develop *matrix/tensor calculus* where several important differentiation rules and integral theorems will be given. Tensors can have physical meanings, allowing us to apply them in many scenarios. One of their usages in Earth Science is to serve as the language in *Continuum Mechanics*. On the other hand, the related matrix calculus is much needed in *Data Assimilation*, which is a technique in Atmospheric Sciences to rectify the output of a forecast from a weather model with available observations.

18.1. Index Notation and Tensors

18.1.1. Rules of Index Notation

Previously, we have usually been writing matrix multiplications along the lines of $AB = C$ or $A\vec{x} = \vec{h}$ without explicitly showing the subscripts and summation symbol. However, if we choose to do so, then they can be expressed as

$$\sum_{k=1}^{r} A_{ik}B_{kj} = C_{ij} \qquad\qquad \sum_{j=1}^{n} A_{ij}x_j = h_i$$

However, the summation symbol is clumsy to be included every time, especially when it involves a chain of products like $ABCD$. The ***Index Notation*** (also known as the ***Einstein Convention***) solves this problem by assuming that any subscript appearing exactly twice in an additive term will be summed (*contracted*) implicitly over all values that the subscript can take. This also means that any such pair of repeated subscripts must occur only in positions with the same range of values. Consequentially, the above examples are simplified into

$$A_{ik}B_{kj} = C_{ij} \qquad\qquad A_{ij}x_j = h_i$$

To be more clear, subscripts that appear exactly twice in any additive term are called ***summation indices*** (or ***dummy indices***) and are summed over specific values which are often automatically inferred from the situation (for example, in three-dimensional physical problems, they will usually take the explicit values of $1, 2, 3$.). On the other hand, subscripts that only appear once in *every* additive term of an expression or equation are known as ***free indices***. In the first example above, k is the dummy index, while i, j are free indices. (How about the second example?[1]) There are some other important points about the Index Notation to be emphasized: First, it is not allowed for a subscript to appear more than

[1] j is the dummy index and i is the free index.

twice within a single additive term (so we cannot write $A_{ik}B_{kj}C_{kl}$, although it is okay for a dummy subscript to appear twice in each of multiple additive terms, e.g. $A_{ik}B_{kj} + C_{ij}D_{kk}$); Second, any pair of dummy indices can be replaced by any other alphabets (e.g. $A_{ik}B_{kj}$ is equivalent to $A_{il}B_{lj}$) but the free indices cannot be changed arbitrarily; Finally, as just mentioned, all free indices have to appear once and only once in every additive term (on both sides if it is an equation) (so $A_i + B_{ij}C_j + D_{ikk}$ is acceptable but not $A_{ij} + B_{kkj}$). As long as the quantities have their subscripts written out, the order of the quantities (but not their indices) is unimportant. Below we give some examples of the Index Notation in a three-dimensional setting:

1. $a_ib_i = \sum_{i=1}^{3} a_ib_i = a_1b_1 + a_2b_2 + a_3b_3$ (Dot Product)
2. $A_{kk} = \sum_{k=1}^{3} A_{kk} = A_{11} + A_{22} + A_{33}$ (Trace)
3. $A_{ik}B_{jk} = \sum_{k=1}^{3} A_{ik}B_{jk} = A_{i1}B_{j1} + A_{i2}B_{j2} + A_{i3}B_{j3}$ (AB^T)
4. $$\begin{aligned} A_{ijk}B_{jl}C_{km} &= \sum_{j=1}^{3}\sum_{k=1}^{3} A_{ijk}B_{jl}C_{km} \\ &= A_{i11}B_{1l}C_{1m} + A_{i12}B_{1l}C_{2m} + A_{i13}B_{1l}C_{3m} \\ &\quad + A_{i21}B_{2l}C_{1m} + A_{i22}B_{2l}C_{2m} + A_{i23}B_{2l}C_{3m} \\ &\quad + A_{i31}B_{3l}C_{1m} + A_{i32}B_{3l}C_{2m} + A_{i33}B_{3l}C_{3m} \\ &:= Y_{ilm} \end{aligned}$$

   will be a subscripted quantity with 3 free indices i, l, m where each additive term contains $3^2 = 9$ smaller terms.
5. $$\begin{aligned} A_{ikl}B_{jkm}C_{lm} &= \sum_{k=1}^{3}\sum_{l=1}^{3}\sum_{m=1}^{3} A_{ikl}B_{jkm}C_{lm} \\ &:= Z_{ij} \end{aligned}$$

will be a subscripted quantity with 2 free indices i, j where each additive term contains $3^3 = 27$ smaller terms.

6. $A_{ij}B_{ijkl}C_{kl} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3}A_{ij}B_{ijkl}C_{kl} := X$ will be a scalar quantity without free indices, consisting of $3^4 = 81$ smaller terms.

► Short Exercise: Determine whether the following equations are valid or not, and justify your answers.[2]

1. $x_i = u_j v_{ijj}$;
2. $A_{ij} = B_{ik}C_{kj} + D_{jl}x_i y_l$;
3. $A_{ij}x_j = B_{ij}$;
4. $A_{ii} + u_j v_j = B_{kl}C_{km}w_{lm} - 1$;
5. $A_{ij} + B_{ij}C_{ij} = D_{ij}$.

18.1.2. Kronecker Delta and Epsilon Symbols

Before getting to know about the *Kronecker delta/epsilon symbols*, it is helpful to denote the three-dimensional Cartesian standard unit vectors $\hat{\imath} = \hat{e}^{(1)}$, $\hat{\jmath} = \hat{e}^{(2)}$, $\hat{k} = \hat{e}^{(3)}$ by $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to switch and ensure the consistency of the notation. We will assume that the problem is always three-dimensional so that every index always runs through $1, 2, 3$ unless otherwise specified. Therefore, given any three-dimensional vector $\mathbf{u} = (u_1, u_2, u_3)^T$, it can be expressed by

$$\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3 = u_i\mathbf{e}_i \tag{18.1}$$

using the Index Notation. Now, the ***Kronecker delta*** symbol is defined as

Definition 18.1.1 (Kronecker Delta Symbol)**.** The Kronecker delta symbol is a

quantity with 2 subscripts such that

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{18.2}$$

Compared with Definition 2.1.1, it works like and can be treated as an identity matrix, although there are some subtle differences between them. Due to the definition of the delta symbol, its effect is to replace a dummy index with the other index. For example,

$$A_{ij}\delta_{jk} = A_{ik} \qquad\qquad A_{ij}B_{kl}\delta_{jl} = A_{ij}B_{kj}$$

since only when $j = k$, the delta symbol will be equal to 1 and otherwise 0 in the first case (how about the second case?). Recall that for the standard unit vectors, we have

$$\mathbf{e}_1 \cdot \mathbf{e}_1 = \mathbf{e}_2 \cdot \mathbf{e}_2 = \mathbf{e}_3 \cdot \mathbf{e}_3 = 1 \tag{18.3a}$$

$$\mathbf{e}_1 \cdot \mathbf{e}_2 = \mathbf{e}_1 \cdot \mathbf{e}_3 = \mathbf{e}_2 \cdot \mathbf{e}_3 = 0 \tag{18.3b}$$

or more succinctly,

$$\mathbf{e}_i \cdot \mathbf{e}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{18.4}$$

Hence we can define the Kronecker delta symbol via

$$\delta_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j \tag{18.5}$$

With this, we can derive (4.4) in Definition 4.2.1 for the real dot product in a new way:

$$\mathbf{u} \cdot \mathbf{v} = (u_i\mathbf{e}_i) \cdot (v_j\mathbf{e}_j)$$

[2]Incorrect (j appears three times in the single additive term on R.H.S.), Correct (i, j are free indices and k, l are dummy indices), Incorrect (j appears twice on L.H.S. but once on R.H.S), Correct (All subscripts are dummy indices and all additive terms are scalars), Incorrect (i, j cannot be free and dummy indices at the same time).

$$
\begin{aligned}
&= u_i v_j (\mathbf{e}_i \cdot \mathbf{e}_j) \\
&= u_i v_j \delta_{ij} && \text{(by (18.5))} \\
&= u_i v_i && (18.6)
\end{aligned}
$$

On the other hand, the ***epsilon (permutation)*** symbol (or the ***Levi-Civita symbol***) is defined as

Definition 18.1.2 (Levi-Civita Epsilon Symbol)**.** The epsilon permutation symbol is a quantity with 3 subscripts such that

$$
\epsilon_{ijk} = \begin{cases} 0 & \text{if any two indices are equal} \\ +1 & \begin{array}{l}\text{if } (i,j,k) \text{ is an even permutation of } (1,2,3), \\ \text{i.e. } (1,2,3), (2,3,1), (3,1,2)\end{array} \\ -1 & \begin{array}{l}\text{if } (i,j,k) \text{ is an odd permutation of } (1,2,3), \\ \text{i.e. } (2,1,3), (3,2,1), (1,3,2)\end{array} \end{cases} \tag{18.7}
$$

Similarly, recall that from Definition 4.2.7, we have

$$\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3, \mathbf{e}_2 \times \mathbf{e}_1 = -\mathbf{e}_3 \tag{18.8a}$$

$$\mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1, \mathbf{e}_3 \times \mathbf{e}_2 = -\mathbf{e}_1 \tag{18.8b}$$

$$\mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2, \mathbf{e}_1 \times \mathbf{e}_3 = -\mathbf{e}_2 \tag{18.8c}$$

$$\mathbf{e}_1 \times \mathbf{e}_1 = \mathbf{e}_2 \times \mathbf{e}_2 = \mathbf{e}_3 \times \mathbf{e}_3 = \mathbf{0} \tag{18.8d}$$

and thus by going through all the cases, they can be summarized as

$$\mathbf{e}_i \times \mathbf{e}_j = \epsilon_{ijk} \mathbf{e}_k \tag{18.9}$$

By taking a dot product with $\mathbf{e}_l$ on both sides and adjusting the indices, we can arrive at an equivalent definition of the epsilon symbol:

$$
\begin{aligned}
\mathbf{e}_l \cdot (\mathbf{e}_i \times \mathbf{e}_j) &= \epsilon_{ijk} (\mathbf{e}_l \cdot \mathbf{e}_k) \\
\mathbf{e}_l \cdot (\mathbf{e}_i \times \mathbf{e}_j) &= \epsilon_{ijk} \delta_{lk} \\
\mathbf{e}_l \cdot (\mathbf{e}_i \times \mathbf{e}_j) &= \epsilon_{ijl}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{e}_k \cdot (\mathbf{e}_i \times \mathbf{e}_j) &= \epsilon_{ijk} && \text{(Replacing a free index } l \to k \\
& && \text{on both sides simultaneously is allowed)} \\
\epsilon_{jki} &= \mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k) && \text{(Rotating the free indices} \\
& && k \to i, i \to j, j \to k) \\
\epsilon_{ijk} &= \mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k) && ((j, k, i) \text{ is an even permutation of } (i, j, k) \\
& && \text{and can be rotated within the epsilon symbol)}
\end{aligned}
\tag{18.10}
$$

In the same spirit, we can derive the formula of cross product in Properties 4.2.9 by the Index Notation as follows.

$$
\begin{aligned}
\mathbf{u} \times \mathbf{v} &= u_i \mathbf{e}_i \times v_j \mathbf{e}_j \\
&= u_i v_j (\mathbf{e}_i \times \mathbf{e}_j) \\
&= u_i v_j \epsilon_{ijk} \mathbf{e}_k \\
&\text{(by (18.9))} \\
&= (u_2 v_3 - u_3 v_2)\mathbf{e}_1 + (u_3 v_1 - u_1 v_3)\mathbf{e}_2 + (u_1 v_2 - u_2 v_1)\mathbf{e}_3 \\
&\text{(Direct evaluation using (18.7) in Definition 18.1.2)}
\end{aligned}
\tag{18.11}
$$

Comparing with the determinant form of cross product (4.9) in Properties 4.2.9, we have

$$
\mathbf{u} \times \mathbf{v} = u_i v_j \epsilon_{ijk} \mathbf{e}_k = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \tag{18.12}
$$

Replacing $\mathbf{e}_k$ by an arbitrary vector w_k, we arrive at an alternative formula for determinant using the Index Notation in three-dimensional vector spaces:

$$
u_i v_j \epsilon_{ijk} w_k = \begin{vmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}
$$

$$\epsilon_{ijk}u_i v_j w_k = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \quad \begin{array}{l}\text{(Swapping rows twice} \\ \text{leaves the sign of determinant unchanged)}\end{array} \tag{18.13}$$

These two symbols are closely related by the ***Epsilon-delta Identity***.

Properties 18.1.3 (ϵ-δ Identity)**.** Contracting over the pair of first subscripts between two epsilon symbols gives

$$\epsilon_{ijk}\epsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km} \tag{18.14}$$

Proof. We will use a brute-force approach and enumerate all the outcomes. Note that the R.H.S. will have values

$$\begin{cases} +1 & \text{if } j = m,\ k = n,\ j \neq n \text{ (and thus } k \neq m) \\ -1 & \text{if } j = n,\ k = m,\ j \neq m \text{ (and thus } k \neq n) \\ 0 & \text{Otherwise} \end{cases} \tag{18.15}$$

Now consider the L.H.S. As the first indices of the two epsilons are equal and every index can only take the three values of $1, 2, 3$, for their product to give a non-zero value, the only two possibilities are (1) $j = m \neq i$, $k = n \neq i$, $j \neq n$; and (2) $j = n \neq i$, $k = m \neq i$, $j \neq m$ where the four free indices j, k, m, n cannot take the same value as the dummy index i. Moreover, j, k (also m, n) cannot be equal simultaneously. These are consistent with the first and second situations in (18.15). Now we only need to justify if they really lead to the outputs of $+1$ and -1 respectively. In the first case, it is not hard to see that ϵ_{ijk} and ϵ_{imn} must be $+1$ or -1 at the same time so their product indeed gives $+1$. Similarly, in the second case, one of the epsilon symbols will be $+1$ and the other will be -1 ($\epsilon_{ijk} = -\epsilon_{ikj} = -\epsilon_{imn}$) so it will produce -1. The identity is hence established. □

Note that contracting one more index (e.g. $j = m$) in the identity gives

$$\begin{aligned}\epsilon_{ijk}\epsilon_{ijn} &= \delta_{jj}\delta_{kn} - \delta_{jn}\delta_{kj} \\ &= 3\delta_{kn} - \delta_{kn} = 2\delta_{kn}\end{aligned} \tag{18.16}$$

as $\delta_{jj} = \delta_{11} + \delta_{22} + \delta_{33} = 1 + 1 + 1 = 3$ and in the second term the dummy index j is replaced by k. Going one step further, contracting all three indices simply gives

$$\epsilon_{ijk}\epsilon_{ijk} = 2\delta_{kk} = 2(3) = 6 \tag{18.17}$$

Now let's see how this identity can be used in showing other vector identities.

Example 18.1.1. Prove Formula (5.8) for vector triple product in Properties 5.4.1 using the Index Notation and Epsilon-delta Identity.

Solution. Applying (18.11) twice, we have

$$\begin{aligned}
&\vec{u} \times (\vec{v} \times \vec{w}) \\
&= (u_j \mathbf{e}_j) \times (v_m w_n \epsilon_{mni} \mathbf{e}_i) \\
&= (u_j \epsilon_{jik} (v_m w_n \epsilon_{mni})) \mathbf{e}_k \\
&= \epsilon_{ikj}\epsilon_{imn} u_j v_m w_n \mathbf{e}_k && \text{(Rotating the epsilon symbols)} \\
&= (\delta_{km}\delta_{jn} - \delta_{kn}\delta_{jm})(u_j v_m w_n \mathbf{e}_k) && \text{((18.14) in Properties 18.1.3)} \\
&= u_j v_k w_j \mathbf{e}_k - u_j v_j w_k \mathbf{e}_k \\
&= (u_j w_j)(v_k \mathbf{e}_k) - (u_j v_j)(w_k \mathbf{e}_k) \\
&= (\vec{u} \cdot \vec{w})\vec{v} - (\vec{u} \cdot \vec{v})\vec{w} && \text{(by (18.6))}
\end{aligned}$$

□

When there is no ambiguity, we will omit the standard unit vector symbol $\mathbf{e}_k$ so that, for example, the second-to-last line above is simply $(u_j w_j)v_k - (u_j v_j)w_k$ and a vector $\vec{v}$ is simply denoted by something like v_i.

18.1.3. Transformation and Quotient Laws for Tensors

Now we will formally introduce ***tensors***. A scalar is a rank-0 tensor and a vector is a rank-1 tensor. The ***rank*** of a tensor (not to be confused with the rank of a matrix) is the number of its subscripts. Linear transformations or bilinear forms will be rank-2 tensors where the associated matrices serve as their representations in certain coordinate systems. Note that this explanation is a largely simplified one, and for Mathematicians the definition will be much more rigorous.[3] But for Physicists or Earth Scientists, it is adequate to know that a rank-r tensor *"transforms like a tensor"*, obeying the so-called *Transformation Law*. To this end, we first need to revisit how a coordinate transformation is carried out. As in (7.8), the change of coordinates matrix is

$$P_{\beta}^{\beta'} = \left[[\vec{v}_{\beta}^{(1)}]_{\beta'} | [\vec{v}_{\beta}^{(2)}]_{\beta'} | \cdots | [\vec{v}_{\beta}^{(n)}]_{\beta'} \right] \tag{18.18}$$

In most of Earth Science usages, the coordinate systems used will be orthonormal and we will assume this is so for the rest of the chapter. This makes the tensors *Cartesian*. By Properties 7.2.6, the coordinates of any vector in the new orthonormal coordinate system relative to the old orthonormal coordinate system can be found by direct projection and vice versa. Therefore, if we denote the old and new orthonormal basis by $\beta = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ and $\beta' = \{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$, the change of coordinates matrix will become

$$P_{\beta}^{\beta'} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}'_1 & \mathbf{e}_2 \cdot \mathbf{e}'_1 & \mathbf{e}_3 \cdot \mathbf{e}'_1 \\ \mathbf{e}_1 \cdot \mathbf{e}'_2 & \mathbf{e}_2 \cdot \mathbf{e}'_2 & \mathbf{e}_3 \cdot \mathbf{e}'_2 \\ \mathbf{e}_1 \cdot \mathbf{e}'_3 & \mathbf{e}_2 \cdot \mathbf{e}'_3 & \mathbf{e}_3 \cdot \mathbf{e}'_3 \end{bmatrix} = \mathbf{e}_j \cdot \mathbf{e}'_i \tag{18.19}$$

[3] Going into depth, the rank of a tensor is actually given by a 2-tuple (r, s) which means that it receives r *covectors* and s vectors to output a scalar. A covector takes one vector to produce a scalar and hence by definition is a $(0, 1)$-tensor, while vice versa a vector will take one covector to also produce a scalar and is a $(1, 0)$-tensor. Linear transformations are usually treated as $(1, 1)$-tensors while bilinear forms are $(0, 2)$-tensors. r and s indicates the number of *contravariant/covariant* indices.

This is an orthogonal matrix[4], and by (7.7), any vector $\vec{v}$ (that is also a rank-1 tensor) is transformed according to

$$\begin{aligned}
[\vec{v}]' &= P_{\beta}^{\beta'}[\vec{v}] \\
v'_i &= (\mathbf{e}_j \cdot \mathbf{e}'_i) v_j \\
v'_j &= (\mathbf{e}_i \cdot \mathbf{e}'_j) v_i
\end{aligned} \tag{18.20}$$

where we only keep the prime to denote the new coordinate frame as a shorthand and have swapped the indices in the last line. From now on, we will write

$$a_{ij} = \mathbf{e}_i \cdot \mathbf{e}'_j \tag{18.21}$$

for convenience, so the above becomes $v'_j = a_{ij} v_i$. Similarly, for a rank-2 tensor F, be it a linear transformation or bilinear form, it is transformed as

$$F'_{kl} = a_{ik} a_{jl} F_{ij} \tag{18.22}$$

This is consistent with both Properties 7.2.2 and Definition 11.1.6 at the same time, because the change of coordinates matrix is orthogonal and so $[a^{-1}]_{ki} = [a^T]_{ki} = a_{ik}$ by Properties 10.1.2. In general, the ***Transformation Law*** says

[4] Extending Properties 7.2.6, we have $\mathbf{e}'_j = (\mathbf{e}'_j \cdot \mathbf{e}_k)\mathbf{e}_k$ which expresses any new basis vector in terms of the sum of projections onto the old orthonormal basis. By considering $\mathbf{e}'_i \cdot \mathbf{e}'_j = \delta_{ij}$ due to (18.5) (valid for any orthonormal basis), we have

$$\begin{aligned}
(\mathbf{e}'_i \cdot \mathbf{e}_k)\mathbf{e}_k \cdot (\mathbf{e}'_j \cdot \mathbf{e}_l)\mathbf{e}_l &= \delta_{ij} \\
(\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_j \cdot \mathbf{e}_l)(\mathbf{e}_k \cdot \mathbf{e}_l) &= \delta_{ij} \\
(\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_j \cdot \mathbf{e}_l)\delta_{kl} &= \delta_{ij} \qquad (\mathbf{e}_k \cdot \mathbf{e}_l = \delta_{kl} \text{ from (18.5) as well}) \\
(\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}'_j \cdot \mathbf{e}_k) &= \delta_{ij}
\end{aligned}$$

where the L.H.S. is essentially $(\mathbf{e}'_i \cdot \mathbf{e}_k)(\mathbf{e}_k \cdot \mathbf{e}'_j) = [(P_{\beta}^{\beta'})^T]_{ik}[P_{\beta}^{\beta'}]_{kj}$ and R.H.S. is effectively an identity matrix, so by Properties 10.1.2 we are done.

Properties 18.1.4. A rank-r *Cartesian* tensor T is a quantity that transforms according to the rule[5]

$$T'_{ijkl\ldots} = a_{mi}a_{nj}a_{pk}a_{ql}\cdots T_{mnpq\ldots} \tag{18.23}$$

where a_{**} invokes the orthogonal transformation matrix as in (18.21).

Physically, it means that tensors are coordinate-independent quantities even though they can have different coordinate representations: These different representations have to be consistent and abide by the Transformation Law so that they still reflect the same physical quantity no matter the choice of coordinate system. Moreover, it implies that to consider if a quantity is a tensor requires a physical context as hinted by the example below.

Example 18.1.2. Show that the following quantities $T = (v_1, v_2) =$

(a) (x_2, x_1)

(b) (x_1^2, x_1x_2)

are actually not rank-1 Cartesian tensors despite looking like vectors (where the index only takes the values of $1, 2$ here).

Solution. For (a), the usual coordinate transformations on x and y respectively by an orthogonal rotation over the plane of a degree θ are given by (10.4):

$$\begin{cases} x'_1 = (\cos\theta)x_1 + (\sin\theta)x_2 = v'_2 \\ x'_2 = (-\sin\theta)x_1 + (\cos\theta)x_2 = v'_1 \end{cases}$$

[5]As we are sticking with orthonormal coordinate systems and orthogonal transformations only, the difference between contravariant/covariant indices (as mentioned in Footnote 3) and hence superscripts/subscripts can be safely ignored for Cartesian tensors. If you don't understand any of what I am saying, ignore this footnote as well.

However, the Transformation Law (Properties 18.1.4) requires that

$$\begin{aligned} v_1' = T_1' &= a_{11}T_1 + a_{21}T_2 \\ &= (\cos\theta)x_2 + (\sin\theta)x_1 \\ v_2' = T_2' &= a_{12}T_1 + a_{22}T_2 \\ &= (-\sin\theta)x_2 + (\cos\theta)x_1 \end{aligned}$$

where

$$\begin{aligned} a_{ij} &= \mathbf{e}_i \cdot \mathbf{e}_j' \\ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} &= \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}_1' & \mathbf{e}_1 \cdot \mathbf{e}_2' \\ \mathbf{e}_2 \cdot \mathbf{e}_1' & \mathbf{e}_2 \cdot \mathbf{e}_2' \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \end{aligned}$$

These two do not match. For (b), we similarly have

$$\begin{aligned} v_1' = (x_1')^2 &= ((\cos\theta)x_1 + (\sin\theta)x_2)^2 \\ &= (\cos\theta)^2 x_1^2 + (\cos\theta)(\sin\theta)x_1x_2 + (\sin\theta)^2 x_2^2 \end{aligned}$$

for the first component, but the Transformation Law demands that

$$\begin{aligned} v_1' = T_1' = (x_1^2)' &= a_{11}T_1 + a_{21}T_2 \\ &= (\cos\theta)x_1^2 + (\sin\theta)x_1x_2 \end{aligned}$$

They are clearly incompatible. Nevertheless, it is quite easy to see that the intuitive displacement vector $\vec{x} = (x_1, x_2)^T$, as well as $(x_2, -x_1)^T$, will be a rank-1 tensor (we leave the verification to the readers)[6]. □

It is not hard to show that the product of two tensors is a tensor.[7] Therefore, kinetic energy (per unit mass) in Newtonian Mechanics $\frac{1}{2}(\vec{v} \cdot \vec{v}) = \frac{1}{2}v_i v_i$ is a rank-0 tensor (scalar) if we gladly accept the intuitive premise that the velocity $\vec{v} = v_i$ is a rank-1 tensor (vector). In addition, there is another way to quickly verify if a quantity is a tensor by the ***Quotient Law*** which works the other way around.

Theorem 18.1.5 (Quotient Law)**.** If $A_{ij\ldots n}$ is a n-subscripted quantity, $B_{pq\ldots m}$ is known to be a rank-m tensor, and the product

$$C_{ij\ldots npq\ldots m} = A_{ij\ldots k\ldots n}B_{pq\ldots k\ldots m}$$

is also a tensor of rank $n+m-2k$ where $k \geq 1$ is the amount of index contraction, then $A_{ij\ldots n}$ will be a rank-n tensor.

With this, we can easily show that some other commonly seen physical quantities are tensors. It will be of use later.

18.1.4. Isotropic Tensors

Here we will revisit the Kronecker delta and epsilon symbols. In fact, they qualify as a rank-2 and rank-3 tensor respectively. Moreover, they are further ***isotropic*** tensors in the sense that they have the same representation (i.e. no directionality) in any coordinate system under the Transformation Law. From another perspective, it has to be isotropic since their components are all constants (0, 1, or -1) and the usual coordinate transformation should have no effect on

[6]For the second case $T = (x_2, -x_1)$, the usual coordinate transformation is the essentially same as in part (a) of the example but with an extra negative sign for the second component, whereas the Transformation Law works as

$$\begin{aligned}
v_1' = T_1' &= a_{11}T_1 + a_{21}T_2 \\
&= (\cos\theta)x_2 + (\sin\theta)(-x_1) \\
&= (\cos\theta)x_2 + (-\sin\theta)x_1 = x_2' \\
v_2' = T_2' &= a_{12}T_1 + a_{22}T_2 \\
&= (-\sin\theta)x_2 + (\cos\theta)(-x_1) \\
&= -[(\cos\theta)x_2 + (\sin\theta)x_1] = -x_1'
\end{aligned}$$

so they are consistent.

them. Now we will explicitly show the isotropic attribute for the Kronecker delta first. The Transformation Law (Properties 18.1.4) requires that

$$\delta'_{ij} = \delta_{ij} \tag{18.24}$$

and the L.H.S. will transform as

$$\begin{aligned}\delta'_{ij} &= a_{ki}a_{lj}\delta_{kl} \\ &= a_{ki}a_{kj} \\ &= [a^T]_{ik}a_{kj} = \delta_{ij}\end{aligned}$$

since a_{**} acts as an orthogonal matrix, and by Properties 10.1.2, it will equal to R.H.S. which is like an identity (also see Footnote 4). This indeed shows that the Kronecker delta is a rank-2 isotropic tensor. In a similar vein, we want to show that

$$\epsilon'_{ijk} = \epsilon_{ijk} \tag{18.25}$$

For this, it is beneficial to first establish that for a 3×3 matrix L, we can extend (18.13) as

$$|L|\epsilon_{lmn} = L_{li}L_{mj}L_{nk}\epsilon_{ijk} \tag{18.26}$$

where $|L|$ is the determinant of L. This can be derived by noting that from (18.13) we have

$$|L| = L_{1i}L_{2j}L_{3k}\epsilon_{ijk} \tag{18.27}$$

[7]For example, if $C_{ij\ldots pq\ldots} = A_{ij\ldots}B_{pq\ldots}$ where A and B are tensors, then each of the smaller tensors will transform accordingly on its own: $A'_{mn\ldots} = a_{im}a_{jn}\cdots A_{ij\ldots}$, $B'_{sr\ldots} = a_{ps}a_{qr}\cdots B_{pq\ldots}$, and thus

$$\begin{aligned}C'_{mn\ldots sr\ldots} = A'_{mn\ldots}B'_{sr\ldots} &= a_{im}a_{jn}\cdots a_{ps}a_{qr}\cdots A_{ij\ldots}B_{pq\ldots} \\ &= a_{im}a_{jn}\cdots a_{ps}a_{qr}\cdots C_{ij\ldots pq\ldots}\end{aligned}$$

will also transform as indicated by Properties 18.1.4.

and swapping or rotating indices, through the properties of the epsilon symbol (or equivalently those of the determinant) will give the other five non-zero possibilities

$$
\begin{aligned}
|L| &= L_{2i}L_{3j}L_{1k}\epsilon_{ijk} \\
|L| &= L_{3i}L_{1j}L_{2k}\epsilon_{ijk} \\
-|L| &= L_{2i}L_{1j}L_{3k}\epsilon_{ijk} \\
-|L| &= L_{3i}L_{2j}L_{1k}\epsilon_{ijk} \\
-|L| &= L_{1i}L_{3j}L_{2k}\epsilon_{ijk}
\end{aligned}
$$

Using the epsilon symbol to combine all six cases then yields (18.26). Now the Transformation Law reads

$$
\begin{aligned}
\epsilon'_{ijk} &= a_{il}a_{jm}a_{kn}\epsilon_{lmn} \\
&= |a|\epsilon_{ijk} && \text{(by (18.26))}
\end{aligned}
$$

and it equals to ϵ_{ijk} as we assume that the coordinate transformation is orthogonal and by Properties 10.1.4 the determinant will take the value $|a| = 1$ or -1. We further place the restriction that it is a *proper* rotation (compared to the *improper* rotation, that is, reflection) so that $|a| = 1$, and hence the desired result follows. Finally, we note that in converse, the Kronecker delta and epsilon tensors are the only possible isotropic tensors of rank 2 and 3, up to a multiplicative constant. The steps of the proof are provided in the Appendix A.5. Although we will not use rank-4 tensors in this book, we also note in passing that the general form of an isotropic rank-4 tensor is

$$
T_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + \nu(\delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}) \tag{18.28}
$$

where λ, μ and ν are some parameters that can take any value.

18.2. Matrix/Tensor Calculus

18.2.1. Differentiating Tensors

The power of tensors is accompanied by the possibility of doing calculus on them. We will first look at differentiation before getting into integrals, just like any introductory Calculus course. Given a tensor $T_{ijk\ldots}(\mathbf{x}, t)$ depending on the position $\mathbf{x} = x_p$ (a vector) and time t (a scalar), which is often the case for physical problems, then differentiating it with respect to t simply means that

$$\left(\frac{\partial T}{\partial t}\right)_{ijk\ldots} = \frac{\partial T_{ijk\ldots}}{\partial t} \tag{18.29}$$

every component of the tensor is differentiated with respect to t, one by one. Meanwhile, to differentiate the tensor by the position vector $\mathbf{x}$ (a ***coordinate derivative***) will lead to

$$\left(\frac{\partial T}{\partial \mathbf{x}}\right)_{ijk\ldots p} = \frac{\partial T_{ijk\ldots}}{\partial x_p} \tag{18.30}$$

a new tensor which is one order higher, where each component of the original tensor is differentiated with each of the coordinates to give a vector and the results are encapsulated in the newly produced subscript. For example, for a vector $\vec{v} = v_i$, differentiation with respect to $\mathbf{x}$ will produced a rank-2 tensor of

$$\left(\frac{\partial \mathbf{v}}{\partial \mathbf{x}}\right)_{ij} = \frac{\partial v_i}{\partial x_j} \tag{18.31}$$

writing out in full matrix form, we have

$$\frac{\partial v_i}{\partial x_j} = \begin{bmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \frac{\partial v_1}{\partial x_3} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_3}{\partial x_1} & \frac{\partial v_3}{\partial x_2} & \frac{\partial v_3}{\partial x_3} \end{bmatrix} \tag{18.32}$$

The Jacobian $\partial F_i/\partial y_j$ previously introduced in Section 17.2.2 is an example of this. For a scalar u, its differentiation with respect to the position is simply a vector:

$$\left(\frac{\partial u}{\partial \mathbf{x}}\right)_i = \frac{\partial u}{\partial x_i} \tag{18.33}$$

while applying that over a rank-2 tensor T_{ij} will similarly produce a rank-3 tensor:

$$\left(\frac{\partial T}{\partial \mathbf{x}}\right)_{ijk} = \frac{\partial T_{ij}}{\partial x_k} \tag{18.34}$$

A special case is that the first index of the tensor being differentiated is contracted with the index of the position derivative. Particularly, (18.31) becomes

$$\frac{\partial v_i}{\partial x_i} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} \tag{18.35}$$

which is exactly the *divergence* $\nabla \cdot \vec{v}$ commonly seen in multivariable Calculus. For simplicity, we will write

$$\nabla = \frac{\partial}{\partial x_i} = \partial_i \tag{18.36}$$

such that

$$\nabla \cdot \vec{v} = \partial_i v_i \tag{18.37}$$

Going one step back, for (18.31) without the index contraction, it can be expressed as a vector gradient:

$$\nabla \vec{v} = \partial_j v_i \tag{18.38}$$

For the scalar case, it reduces to the usual *gradient*:

$$\nabla u = \partial_i u \tag{18.39}$$

It is not hard to extend this for the *curl* which is another common operation appearing in multivariable Calculus. In the same essence, referring to (18.12) with the indices rotated suitably and plugging in (18.36), we have

$$\nabla \times \vec{v} = \epsilon_{ijk}\partial_j v_k \tag{18.40}$$

The usual product rule is still applicable to the ∂_i operator, for example

$$\begin{aligned}\nabla(u\vec{v}) = \partial_j(uv_i) &= u\partial_j(v_i) + v_i\partial_j(u) \\ &= u\nabla\vec{v} + \vec{v}\nabla u\end{aligned}$$

Finally, we note that if $\vec{v} = \vec{x}$, then (18.31) will become

$$\frac{\partial x_i}{\partial x_j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} = \delta_{ij} \tag{18.41}$$

which is just how the usual partial derivatives work and this can be compared to (18.2).

Example 18.2.1. Show that differentiating a quadratic form $\mathbf{x}^T A\mathbf{x}$ with respect to the position vector $\mathbf{x}$ where A does not depend on $\mathbf{x}$ yields, in matrix form, $A\mathbf{x} + A^T\mathbf{x}$.

Solution. The quadratic form written in the Index Notation is

$$x_i A_{ij} x_j$$

Differentiation then gives

$$\begin{aligned}\partial_k(x_i A_{ij} x_j) &= \partial_k(x_i A_{ij} x_j) \\ &= \partial_k(x_i)A_{ij}x_j + x_i A_{ij}\partial_k(x_j) && \text{(Product Rule)} \\ &= \delta_{ik}A_{ij}x_j + x_i A_{ij}\delta_{jk} && \text{(by (18.41))} \\ &= A_{kj}x_j + x_i A_{ik} \\ &= A_{kj}x_j + (A^T)_{ki}x_i = A\mathbf{x} + A^T\mathbf{x}\end{aligned}$$

□

► Short Exercise: What is the result if the expression above is differentiated with respect to **x** once more?[8]

Example 18.2.2. Prove the following formula for the curl of cross product:

$$\nabla \times (\vec{u} \times \vec{v}) = (\nabla \cdot \vec{v})\vec{u} - (\nabla \cdot \vec{u})\vec{v} + \vec{v} \cdot \nabla\vec{u} - \vec{u} \cdot \nabla\vec{v} \tag{18.42}$$

Solution.

$$\begin{aligned}
&\nabla \times (\vec{u} \times \vec{v}) \\
&= \nabla \times (\epsilon_{ijk} u_j v_k) && \text{(by (18.12))} \\
&= \epsilon_{lmi} \partial_m (\epsilon_{ijk} u_j v_k) && \text{(by (18.40))} \\
&= \epsilon_{ilm} \epsilon_{ijk} [u_j \partial_m (v_k) + \partial_m (u_j) v_k] && \text{(Product Rule)} \\
&= (\delta_{lj}\delta_{mk} - \delta_{lk}\delta_{mj})[u_j \partial_m (v_k) + \partial_m (u_j) v_k] && \text{(by (18.14))} \\
&= [u_l \partial_m (v_m) + \partial_m (u_l) v_m] - [u_m \partial_m (v_l) + \partial_m (u_m) v_l] \\
&= (\nabla \cdot \vec{v})\vec{u} + \vec{v} \cdot \nabla\vec{u} - \vec{u} \cdot \nabla\vec{v} - (\nabla \cdot \vec{u})\vec{v}
\end{aligned}$$

□

18.2.2. Tensor Integral Theorem

On the other hand, for integrals involving tensors, the overarching result is the ***(Tensor) Divergence Theorem*** from Differential Geometry, the development of which is too advanced to be included and hence will be stated directly.

Theorem 18.2.1 (Tensor Divergence Theorem)**.** For a Cartesian tensor $T_{ij\ldots k}$ of an arbitrary rank and over a convex regular region with a volume V and surface S, its surface flux integral is equal to the volume integral of its coordinate

[8] It is simply $A + A^T$.

derivative.

$$\int_S T_{ij\ldots k} n_l dS = \int_V \frac{\partial T_{ij\ldots k}}{\partial x_l} dV \tag{18.43}$$

where $\hat{n} = n_l$ is the unit outward normal vector.

Consequentially, the usual ***Divergence Theorem*** and ***Stokes' Theorem*** can be readily seen to be a special case of the Tensor Divergence Theorem:

$$\begin{aligned}
\int_S \vec{v} \cdot \hat{n} dS = \int_S v_l n_l dS &= \int_V \frac{\partial v_l}{\partial x_l} dV && \text{(by (18.43))} \\
&= \int_V (\nabla \cdot \vec{v}) dV && \text{(by (18.37))} \qquad (18.44) \\
\int_S \hat{n} \times \vec{v} dS = \int_S \epsilon_{klm} v_m n_l dS &= \int_V \frac{\partial(\epsilon_{klm} v_m)}{\partial l} dV && \text{(by (18.43))} \\
&= \int_V \epsilon_{klm} \partial_l v_m dV \\
&= \int_V (\nabla \times \vec{v}) dV && \text{(by (18.40))} \qquad (18.45)
\end{aligned}$$

18.3. Earth System Applications: Continuum Mechanics

Reference Materials: [10]

18.3.1. Stress Tensor and Mohr's Circle

Due to the scope limit, we will only briefly go through the stress tensor's derivation. For a *mass element* that is infinitesimally small, the traction force on a (possibly oblique) surface with a normal orientation of $\hat{n} = n_i$ can be

reasonably expected to have contributions from three stress components along each direction. For instance, in the x-direction, using the xyz coordinates in the indices for now, the traction will be

$$T_x = \sigma_{xx}n_x + \sigma_{yx}n_y + \sigma_{zx}n_z \tag{18.46a}$$

and similarly

$$T_y = \sigma_{xy}n_x + \sigma_{yy}n_y + \sigma_{zy}n_z \tag{18.46b}$$

$$T_z = \sigma_{xz}n_x + \sigma_{yz}n_y + \sigma_{zz}n_z \tag{18.46c}$$

where the subscripts in σ_{**} follow the *"on-in" rule*, e.g. σ_{zx} means the stress on the z-plane in the x-direction, σ_{xz} means the stress on the x-plane in the z-direction and σ_{yy} means the stress on the y-plane in the y-direction. A schematic is provided in Figure 18.1 where we now replace the xyz indices by $1, 2, 3$.[9] (18.46) can be summarized in the matrix form of

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \sigma_{31} \\ \sigma_{12} & \sigma_{22} & \sigma_{32} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}^T \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{18.47}$$

or more compactly by Index Notation, the ***Cauchy's Formula*** becomes

$$T_j = \sigma_{ij}n_i \tag{18.48}$$

Since the traction, as a force, and the direction are both vectors, by the Quotient Law (Theorem 18.1.5), σ_{ij} must be a rank-2 tensor. It is subsequently known as the ***Cauchy Stress Tensor***. Now, we assume any mass element within a continuum (e.g. a rock layer) has to be in *static equilibrium* so it is not moving. This specifically requires that there is no net torque that will rotate the mass element (presented as a parallelepiped in Figure 18.2). The stresses on the opposite faces of the parallelepiped must then be equal in magnitude and opposite

[9] The complete derivation requires us to consider the *Cauchy Tetrahedron*.

Figure 18.1.: *The components of stress acting on a mass element.*

in direction. Take the 2, 3 directions as an illustration, the shear stresses on any pair of opposite faces (e.g. the component σ_{23} on the 2-planes) in another axis (e.g. in the 3-direction) will generate a couple that tends to rotate the parallelepiped. To prevent the mass element from rotating, the shear stresses on the corresponding side faces (e.g. the component σ_{32} on the 3-planes) must balance this torque out, and thus $\sigma_{23} = \sigma_{32}$. By the same logic, $\sigma_{13} = \sigma_{31}$ and $\sigma_{12} = \sigma_{21}$, or more generally

$$\sigma_{ij} = \sigma_{ji} \tag{18.49}$$

which means that the stress tensor is a symmetric one (this also means that the transpose in (18.47) is not important). Moreover, as it is a tensor, it obeys the Transformation Law (Properties 18.1.4) so that the change of coordinates will be

$$\sigma'_{ij} = a_{ki} a_{lj} \sigma_{kl} \tag{18.50}$$

when viewed from the matrix perspective, it easily reminds us about orthogonal diagonalization (10.6) in Definition 10.2.1 since a_{**} represents an orthogonal matrix and Theorem 10.2.2 is applicable to the symmetric stress tensor.

Figure 18.2.: *The balance between shear stresses along the* $2, 3$ *(yz) cross-section for the mass element in Figure 18.1.*

Therefore, we can always compute the orthonormal eigenvectors for any matrix representation of the stress tensor. After a change of coordinates via orthogonal diagonalization, these orthonormal eigenvectors will then become the ***principal axes/directions*** (akin to Theorem 11.2.8) used in the new coordinate frame. The corresponding eigenvalues are naturally called the ***principal stresses*** (denoted by σ_* with one subscript only) and there will be no shearing components in the transformed representation of the stress tensor. Hence, the traction along any principal direction will be aligned with the normal to the surface and is equal to the respective principal stress. The geological convection is that compression/tension is taken to be positive/negative. For the ease of demonstration, we will consider plane stress with only 2 directions first.

Example 18.3.1. For a two-dimensional stress tensor that has the components

of (unit: MPa)

$$\sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 1 & 7 \end{bmatrix}$$

find its principal stresses and principal directions.

Solution. We can directly compute the eigenvalues and eigenvectors, but we can do it in a more tricky way by recalling the results indicated by (11.12) and (11.13). The angle θ required to align the coordinate system is immediately inferred from

$$\cot(2\theta) = \frac{5-7}{2} = -1$$

which translates to $\theta = -\frac{\pi}{8}$, $\cos(\theta) = \frac{\sqrt{2+\sqrt{2}}}{2}$ and $\sin(\theta) = -\frac{\sqrt{2-\sqrt{2}}}{2}$. So the principal directions are at an angle of $\frac{\pi}{8}$ to the original axes. The principal stresses are then found by (11.12) as well:

$$\begin{bmatrix} \cos\left(-\frac{\pi}{8}\right) & \sin\left(-\frac{\pi}{8}\right) \\ -\sin\left(-\frac{\pi}{8}\right) & \cos\left(-\frac{\pi}{8}\right) \end{bmatrix} \begin{bmatrix} 5 & 1 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} \cos\left(-\frac{\pi}{8}\right) & -\sin\left(-\frac{\pi}{8}\right) \\ \sin\left(-\frac{\pi}{8}\right) & \cos\left(-\frac{\pi}{8}\right) \end{bmatrix} = \begin{bmatrix} 6-\sqrt{2} & 0 \\ 0 & 6+\sqrt{2} \end{bmatrix}$$

So the principal stresses are $6-\sqrt{2} \approx 4.59\,\text{MPa}$ and $6+\sqrt{2} \approx 7.41\,\text{MPa}$. We will usually arrange the index from the largest to smallest stress so we will write them as $\sigma_1 = 7.41\,\text{MPa}$ and $\sigma_2 = 4.59\,\text{MPa}$ and this will require us to rotate the coordinate frame by an extra 90°. □

By reversing and extending the logic in (11.12), a stress tensor that has been aligned in the principal directions and only has non-zero stress components along the main diagonal, can also undergo any rotation of degree θ, resulting in

$$\sigma'_{ij} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_1\cos\theta & -\sigma_1\sin\theta \\ \sigma_2\sin\theta & \sigma_2\cos\theta \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1\cos^2\theta + \sigma_2\sin^2\theta & -\sigma_1\sin\theta\cos\theta + \sigma_2\cos\theta\sin\theta \\ -\sigma_1\sin\theta\cos\theta + \sigma_2\sin\theta\cos\theta & \sigma_1\sin^2\theta + \sigma_2\cos^2\theta \end{bmatrix}
\end{aligned} \tag{18.51}
$$

Rewriting the components using trigonometric identities simplifies them into

$$
\begin{aligned}
\sigma_{x'} = \sigma_1' &= \sigma_1\cos^2\theta + \sigma_2\sin^2\theta \\
&= \sigma_1\left(\frac{1+\cos(2\theta)}{2}\right) + \sigma_2\left(\frac{1-\cos(2\theta)}{2}\right) \\
&= \frac{\sigma_1+\sigma_2}{2} + \frac{\sigma_1-\sigma_2}{2}\cos(2\theta) && (18.52a) \\
\sigma_{y'} = \sigma_2' &= \sigma_1\sin^2\theta + \sigma_2\cos^2\theta \\
&= \sigma_1\left(\frac{1-\cos(2\theta)}{2}\right) + \sigma_2\left(\frac{1+\cos(2\theta)}{2}\right) \\
&= \frac{\sigma_1+\sigma_2}{2} - \frac{\sigma_1-\sigma_2}{2}\cos(2\theta) && (18.52b) \\
\tau_{x'y'} = \sigma_{12}' = \sigma_{21}' &= -\sigma_1\sin\theta\cos\theta + \sigma_2\cos\theta\sin\theta \\
&= -\frac{\sigma_1-\sigma_2}{2}\sin(2\theta) && (18.52c)
\end{aligned}
$$

where $\sigma_{x'}$ and $\sigma_{y'}$ are the normal tractions along the rotated axes and $\tau_{x'y'}$ is the shear traction. Notice that $\sigma_{x'}$ (or $\sigma_{y'}$) and $\tau_{x'y'}$ together form a circle parameterized by the double angle 2θ with its center at $(\sigma, \tau) = (\frac{\sigma_1+\sigma_2}{2}, 0)$ and a radius of $\frac{\sigma_1-\sigma_2}{2}$.[10] Hence the information about normal/shear stresses on a surface of any orientation can be condensed into a circle centered at the horizontal axis that also cuts the horizontal axis at $\sigma = \sigma_1$ and σ_2 graphically. This is known as the ***Mohr's Circle*** and is illustrated in Figure 18.3. Bear in mind that when the geological convention is used, (18.52c) will have no negative sign, and an anti-clockwise [clockwise] rotation of the surface by a positive [negative] degree of θ will appear as an anti-clockwise [clockwise] rotation of

[10] $(\sigma_{x'} - \frac{\sigma_1+\sigma_2}{2})^2 + \tau_{x'y'}^2 = (\frac{\sigma_1-\sigma_2}{2})^2\cos^2(2\theta) + (\frac{\sigma_1-\sigma_2}{2})^2\sin^2(2\theta) = (\frac{\sigma_1-\sigma_2}{2})^2$

Figure 18.3.: *The schematic diagram of a Mohr's Circle.*

the diameter within the Mohr's Circle by a positive [negative] degree 2θ (double of the actual surface tilt). The largest shear traction will be obtained at a tilt of 45°.

Example 18.3.2. For a two-dimensional plane with the principal stresses of $\sigma_1 = 10\,\text{MPa}$ and $\sigma_2 = 5\,\text{MPa}$, find the normal and shear tractions over the surface with a tilt of 25° from the principal axes in the positive direction.

Solution. The Mohr's Circle will have a radius of $\frac{\sigma_1-\sigma_2}{2} = \frac{10-5}{2} = 2.5$, centered at $(\frac{\sigma_1+\sigma_2}{2}, 0) = (\frac{10+5}{2}, 0) = (7.5, 0)$ along the horizontal axis representing the normal direction. By (18.52a) and (18.52b) the normal tractions (in MPa) will be

$$\begin{aligned}\sigma_{x'}, \sigma_{y'} = 7.5 \pm 2.5\cos(2(25°)) &= 7.5 \pm 2.5\cos(50°)\\ &= 7.5 \pm 1.61 = 9.11, 5.89\end{aligned}$$

and similarly the shear stress will be $\tau_{x'y'} = 2.5\sin(2(25°)) = 1.92\,\text{MPa}$ from (18.52c). □

► Short Exercise: What are the new normal/shear tractions if the surface is further rotated by 37°?[11]

The generalization of Mohr's Circle to a three-dimensional volume can be done by first assuming that the coordinate system has been properly rotated to coincide with the three principal directions. Denote the respective principal stresses by $\sigma_1, \sigma_2, \sigma_3$ where they are arranged in descending order $\sigma_1 \geq \sigma_2 \geq \sigma_3$. Then given a normal $\hat{n} = n_i$ to the plane, the net traction is given by Cauchy's Formula (18.47):

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} = \begin{bmatrix} \sigma_1 n_1 \\ \sigma_2 n_2 \\ \sigma_3 n_3 \end{bmatrix} \tag{18.53}$$

and considering the squared magnitude of the traction, which is simply the sum of the squared magnitudes of the normal traction σ_n and shear traction τ, we have

$$\sigma_n^2 + \tau^2 = \|T\|^2 = T_1^2 + T_2^2 + T_3^2 = \sigma_1^2 n_1^2 + \sigma_2^2 n_2^2 + \sigma_3^2 n_3^2 \tag{18.54}$$

where the normal traction is simply

$$\sigma_n = (\sigma_1 n_1, \sigma_2 n_2, \sigma_3 n_3)^T \cdot (n_1, n_2, n_3)^T = \sigma_1 n_1^2 + \sigma_2 n_2^2 + \sigma_3 n_3^2 \tag{18.55}$$

Finally, we have a quite trivial condition of $\hat{n}$ being a unit normal vector:

$$1 = n_1^2 + n_2^2 + n_3^2 \tag{18.56}$$

(18.54), (18.55), (18.56) together form a linear system of three equations in three unknowns n_1^2, n_2^2, n_3^2. This can be efficiently dealt with with Cramer's Rule (Theorem 3.2.2). Considering the first unknown n_1^2, we have

$$n_1^2 = \frac{\begin{vmatrix} \sigma_n^2 + \tau^2 & \sigma_2^2 & \sigma_3^2 \\ \sigma_n & \sigma_2 & \sigma_3 \\ 1 & 1 & 1 \end{vmatrix}}{\begin{vmatrix} \sigma_1^2 & \sigma_2^2 & \sigma_3^2 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ 1 & 1 & 1 \end{vmatrix}}$$

$$
\begin{aligned}
&= \frac{[(\sigma_n^2+\tau^2)\sigma_2+\sigma_2^2\sigma_3+\sigma_n\sigma_3^2]-[\sigma_2\sigma_3^2+\sigma_3(\sigma_n^2+\tau^2)+\sigma_n\sigma_2^2]}{(\sigma_1^2\sigma_2+\sigma_2^2\sigma_3+\sigma_3^2\sigma_1)-(\sigma_2\sigma_3^2+\sigma_3\sigma_1^2+\sigma_1\sigma_2^2)} \\
&= \frac{(\sigma_n^2+\tau^2)(\sigma_2-\sigma_3)-\sigma_n(\sigma_2^2-\sigma_3^2)+(\sigma_2^2\sigma_3-\sigma_2\sigma_3^2)}{(\sigma_1^2\sigma_2+\sigma_2^2\sigma_3+\sigma_3^2\sigma_1+\sigma_1\sigma_2\sigma_3)-(\sigma_2\sigma_3^2+\sigma_3\sigma_1^2+\sigma_1\sigma_2^2+\sigma_1\sigma_2\sigma_3)} \\
&= \frac{(\sigma_n^2+\tau^2)(\sigma_2-\sigma_3)-\sigma_n(\sigma_2+\sigma_3)(\sigma_2-\sigma_3)+\sigma_2\sigma_3(\sigma_2-\sigma_3)}{(\sigma_1-\sigma_2)(\sigma_2-\sigma_3)(\sigma_1-\sigma_3)} \\
&= \frac{(\sigma_n^2+\tau^2)-\sigma_n(\sigma_2+\sigma_3)+\sigma_2\sigma_3}{(\sigma_1-\sigma_2)(\sigma_1-\sigma_3)} \qquad (18.57)
\end{aligned}
$$

Since $n_1^2 \geq 0$, $\sigma_1-\sigma_2 \geq 0$ and $\sigma_1-\sigma_3 \geq 0$, we have

$$
\begin{aligned}
(\sigma_n^2+\tau^2)-\sigma_n(\sigma_2+\sigma_3)+\sigma_2\sigma_3 &\geq 0 \\
\left(\sigma_n-\frac{\sigma_2+\sigma_3}{2}\right)^2+\tau^2+\sigma_2\sigma_3-\left(\frac{\sigma_2+\sigma_3}{2}\right)^2 &\geq 0 \\
\left(\sigma_n-\frac{\sigma_2+\sigma_3}{2}\right)^2+\tau^2 &\geq \left(\frac{\sigma_2+\sigma_3}{2}\right)^2-\sigma_2\sigma_3 \\
&= \frac{\sigma_2^2}{4}+\frac{\sigma_2\sigma_3}{2}+\frac{\sigma_3^2}{4}-\sigma_2\sigma_3 \\
&= \frac{\sigma_2^2}{4}-\frac{\sigma_2\sigma_3}{2}+\frac{\sigma_3^2}{4} \\
&= \left(\frac{\sigma_2-\sigma_3}{2}\right)^2 \qquad (18.58)
\end{aligned}
$$

by completing squares. Therefore, the normal and shear traction must lie outside the Mohr's Circle centered at $(\frac{\sigma_2+\sigma_3}{2}, 0)$ with a radius of $\frac{\sigma_2-\sigma_3}{2}$, which represents a circle that lies along the horizontal (normal traction) axis and cuts it at $\sigma_n = \sigma_2$ and σ_3. Similarly, we leave it to the readers to check that

$$
n_2^2 = -\frac{(\sigma_n^2+\tau^2)-\sigma_n(\sigma_1+\sigma_3)+\sigma_1\sigma_3}{(\sigma_1-\sigma_2)(\sigma_2-\sigma_3)} \qquad (18.59)
$$

[11] This amounts to a net rotation of $(25+37)^\circ = 62^\circ$ so we simply repeat the calculation with $\theta = 62^\circ$. The results should be $\sigma_{x'}, \sigma_{y'} = 7.5 \pm 2.5\cos(2(62^\circ)) = 7.5 \pm 2.5\cos(124^\circ) = 7.5 \mp 1.40 = 6.10, 8.90$ and $\tau_{x'y'} = 2.5\sin(124^\circ) = 2.07$. Graphically, we can just rotate the previous diameter by an extra of $2 \times 37^\circ = 74^\circ$ on the Mohr's Circle.

Figure 18.4.: *Mohr's Circles extended to the three-dimensional space, adapted to Example 18.3.3. The normal/shear tractions* (σ_n, τ) *have to lie within the shaded region.*

$$n_3^2 = \frac{(\sigma_n^2 + \tau^2) - \sigma_n(\sigma_1 + \sigma_2) + \sigma_1\sigma_2}{(\sigma_2 - \sigma_3)(\sigma_1 - \sigma_3)} \tag{18.60}$$

(notice the negative sign for n_2^2, why?) and thus along the same lines, we have

$$\left(\sigma_n - \frac{\sigma_1 + \sigma_3}{2}\right)^2 + \tau^2 \leq \left(\frac{\sigma_1 - \sigma_3}{2}\right)^2 \tag{18.61}$$

$$\left(\sigma_n - \frac{\sigma_1 + \sigma_2}{2}\right)^2 + \tau^2 \geq \left(\frac{\sigma_1 - \sigma_2}{2}\right)^2 \tag{18.62}$$

So the normal and shear traction must also fall outside the Mohr's Circle centered at $(\frac{\sigma_1+\sigma_2}{2}, 0)$ with a radius of $\frac{\sigma_1-\sigma_2}{2}$ that cuts the horizontal axis at $\sigma_n = \sigma_1$ and σ_2, but inside the larger Mohr's Circle also centered at the horizontal axis that cuts it at $\sigma_n = \sigma_1$ and σ_3. These are demonstrated in Figure 18.4.

Example 18.3.3. For a three-dimensional volume with principal stresses of $\sigma_1 = 12$, $\sigma_2 = 7$, $\sigma_3 = 4$ (in MPa), check if the following pairs of normal/shear tractions (σ_n, τ) are valid.

(a) $(9, -1)$

(b) $(6.5, 2.5)$

(c) $(5, 3.2)$

Solution. Figure 18.4 has been drawn for the situation in this question. We simply mark the locations for the traction pairs and see if the dots fall within the shaded region (green) bounded by the three Mohr's Circles. The direct visual inspection shows that the second pair is possible but the first and last pairs are not valid. □

We note that the largest shear traction is acquired when the surface is inclined at 45° just between the first and third principal planes, similar to the plane stress scenario.

18.3.2. Eulerian and Lagrangian Views, Spatial and Material Derivatives

In Continuum Mechanics, the continuum may deform and move along the flow. Due to this, we have two descriptions of the continuum: *Eulerian* and *Lagrangian* descriptions. Given a property P of the continuum, in the ***Eulerian (spatial) description***, we retrieve the value of P by referring to the local coordinates $\mathbf{x} = x_i$ at the current time t, so that $P = P(x_i, t)$. This is more commonly used in many situations because of convenience. However, sometimes we also need the ***Lagrangian (material) description*** that essentially follows a mass particle. We need to "label" the particle with reference coordinates $\mathbf{X} = X_i$ (in capital letters) which are the local coordinates recorded at some reference time t_0, such that $X_i = x_i(t_0)$. In this case, the value of P is evaluated by going back to the fixed

reference coordinates, and hence we have $P = P(X_i, t)$ instead. It is possible to convert between these two descriptions by

$$x_i = x_i(X_i, t) \tag{18.63a}$$

which extracts the current local coordinates using labels from the Lagrangian description, and vice versa

$$X_i = X_i(x_i, t) \tag{18.63b}$$

which returns the labeled coordinates using the Eulerian description. This leads to the ***displacement field***

$$u_i = x_i - X_i \tag{18.64}$$

which is the displacement between the current position of particles and their initial reference position. The ***material displacement gradient*** is the displacement differentiated with respect to the material coordinates:

$$\frac{\partial u_i}{\partial X_j} = \frac{\partial x_i}{\partial X_j} - \frac{\partial X_i}{\partial X_j} = \frac{\partial x_i}{\partial X_j} - \delta_{ij} \tag{18.65}$$

Similarly, we also have the ***spatial displacement gradient*** that comes from differentiating the displacement with respect to the spatial coordinates:

$$\frac{\partial u_i}{\partial x_j} = \frac{\partial x_i}{\partial x_j} - \frac{\partial X_i}{\partial x_j} = \delta_{ij} - \frac{\partial X_i}{\partial x_j} \tag{18.66}$$

Both of them are rank-2 tensors.[12] A comparison of these two descriptions is given in Figure 18.5. Notice that the quantity $\partial x_i / \partial X_j$ (also the closely related

[12] We only show that for the first one but the second one is very similar. Note that by the multivariable Chain Rule

$$(\frac{\partial u_i}{\partial X_j}) dX_j = du_i$$

Since dX_j and du_i are (differentials of) displacement and velocity, they are vectors and an immediate use of the Quotient Law (Theorem 18.1.5) shows that $\partial u_i / \partial X_j$ has to be a rank-2 tensor.

Figure 18.5.: *The conversion between the Eulerian/Lagrangian views for coordinates and displacements.*

$\partial X_i/\partial x_j$) appeared in the displacement gradient. This is called the ***Jacobian*** of the deformation. To see this, by the multivariable Chain Rule, the transformation between the Eulerian and Lagrangian displacements is explicitly

$$\begin{cases} d\mathbf{x}_1 = \dfrac{\partial x_1}{\partial X_1} dX_1 \mathbf{e}_1 + \dfrac{\partial x_1}{\partial X_2} dX_2 \mathbf{e}_2 + \dfrac{\partial x_1}{\partial X_3} dX_3 \mathbf{e}_3 & \text{(18.67a)} \\ d\mathbf{x}_2 = \dfrac{\partial x_2}{\partial X_1} dX_1 \mathbf{e}_1 + \dfrac{\partial x_2}{\partial X_2} dX_2 \mathbf{e}_2 + \dfrac{\partial x_2}{\partial X_3} dX_3 \mathbf{e}_3 & \text{(18.67b)} \\ d\mathbf{x}_3 = \dfrac{\partial x_3}{\partial X_1} dX_1 \mathbf{e}_1 + \dfrac{\partial x_3}{\partial X_2} dX_2 \mathbf{e}_2 + \dfrac{\partial x_3}{\partial X_3} dX_3 \mathbf{e}_3 & \text{(18.67c)} \end{cases}$$

The deformed Eulerian volume dV for an infinitesimally small mass is then given by (5.5) from Properties 5.3.2.

$$
\begin{aligned}
(d\mathbf{x}_1 \times d\mathbf{x}_2) \cdot d\mathbf{x}_3 &= \begin{vmatrix} \frac{\partial x_1}{\partial X_1} dX_1 & \frac{\partial x_1}{\partial X_2} dX_2 & \frac{\partial x_1}{\partial X_3} dX_3 \\ \frac{\partial x_2}{\partial X_1} dX_1 & \frac{\partial x_2}{\partial X_2} dX_2 & \frac{\partial x_2}{\partial X_3} dX_3 \\ \frac{\partial x_3}{\partial X_1} dX_1 & \frac{\partial x_3}{\partial X_2} dX_2 & \frac{\partial x_3}{\partial X_3} dX_3 \end{vmatrix} \\
&= \det\left(\frac{\partial x_i}{\partial X_j}\right) dX_1 dX_2 dX_3 = J dV_0
\end{aligned} \tag{18.68}
$$

where $dV_0 = dX_1 dX_2 dX_3$ is the infinitesimal reference Lagrangian volume differential and

$$
J = \det\left(\frac{\partial x_i}{\partial X_j}\right) \tag{18.69}
$$

is the determinant of the Jacobian. Alternatively, using the Index Notation and epsilon symbol, we can also arrive at

$$
\begin{aligned}
dV &= \epsilon_{lmn} dx_l dx_m dx_n \\
&= \epsilon_{lmn} \frac{\partial x_l}{\partial X_i} \frac{\partial x_m}{\partial X_j} \frac{\partial x_n}{\partial X_k} dX_i dX_j dX_k && \text{(Multivariable Chain Rule)} \\
&= \epsilon_{ijk} J dX_i dX_j dX_k && \text{(by (18.26) and (18.69))} \\
&= J dV_0
\end{aligned} \tag{18.70}
$$

Consequentially, J is the ratio of the current to the reference volume. Due to the conservation of matter, the deformed volume cannot vanish if the initial reference volume is finite and vice versa. This implies that the determinant of the Jacobian (or simply called Jacobian from now on) cannot vanish, i.e. $J = \det\left(\partial x_i / \partial X_j\right) \neq 0$. This ensures that the conversion between the Eulerian and Lagrangian descriptions is one-to-one.

Now, we are going to examine the time derivative of a tensor quantity $P = P_{ij\ldots}$ from both the Eulerian and Lagrangian views. Using the Lagrangian description, it can be expressed as $P = P(\mathbf{X}, t)$ where $\mathbf{X}$ represents the reference coordinates. The ***material derivative*** dP/dt is the rate of change in P following a particle with respect to time t. Since the reference coordinates of the tracked particle are essentially labels that do not change with time, the material derivative of $P(\mathbf{X}, t)$ in the Lagrangian framework is simply

$$\frac{d}{dt}P(\mathbf{X}, t) = \frac{\partial}{\partial t}P(\mathbf{X}, t) \tag{18.71}$$

its partial derivative with respect to time. However, in the Eulerian description, we use the current coordinates of the particle that will change in time such that $P = P(\mathbf{x}, t)$ and $\mathbf{x} = \mathbf{x}(\mathbf{X}, t)$, and thus the material derivative has to be obtained using the multivariable Chain Rule as

$$\begin{aligned}\underbrace{\frac{d}{dt}P(\mathbf{x}, t)}_{\text{Material}} &= \frac{\partial}{\partial t}P(\mathbf{x}, t) + \frac{\partial}{\partial x_k}[P(\mathbf{x}, t)]\frac{dx_k}{dt} \\ &= \underbrace{\frac{\partial}{\partial t}P(\mathbf{x}, t)}_{\text{Spatial}} + \underbrace{v_k\frac{\partial}{\partial x_k}[P(\mathbf{x}, t)]}_{\text{Advective}}\end{aligned} \tag{18.72}$$

where $d\mathbf{x}/dt = dx_k/dt = v_k = \mathbf{v}$ are the velocities of the particle. This shows that the material derivative is the combined effects of the ***spatial derivative*** which is the local rate of change of P fixed at the present position of the particle, and the ***advective derivative*** which is the rate of change of P attributed to the particle being advected along the gradient of the field of P due to the flow. Accordingly, the material derivative operator expressed in the Eulerian spatial description is given by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + v_k\frac{\partial}{\partial x_k} = \frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla \tag{18.73}$$

Going in the reverse direction, it implies that the evolution of some property measured at a fixed point in space (spatial) is the actual change imparted to the

particle (material) subtracted by the apparent advective effect due to its motion. The relationship between these three terms is demonstrated in Figure 18.6 in a one-dimensional setting.

Figure 18.6.: *The relationship between the spatial (blue) and material derivative (red) alongside the advective term (green), where we use temperature T as the variable and only look at a single direction x (and velocity u) for demonstration.*

18.3.3. Field Equations

In this subsection, we will proceed to derive several ***field equations*** which govern the dynamics of physical fields like temperature and motion. The central ingredient for the derivation is the *conservation of mass*, and we will consider the mass contained in a volume bounded by an arbitrary closed surface. The

boundary of this volume will be assumed to move alongside the particles on it and thus it is a ***material volume***. Therefore, it will contain the same set of particles for all time. Considering this material volume, the amount of mass inside it can be expressed as an integral over the deformed volume in the current configuration as

$$m = \int_{V(t)} \rho(\mathbf{x}, t) dV \tag{18.74}$$

where $V(t)$ represents the current volume and $\rho(\mathbf{x}, t)$ is the density field in the Eulerian frame, both of which can change with time. However, we can also switch to the Lagrangian frame where the mass integral is taken over a fixed reference volume $V_0 = V(t_0)$, which coincides with the material volume of the mass at some reference time. Moreover, at the reference time, the density will no longer depend on time and will only vary in the reference coordinates, i.e. $\rho_0 = \rho_0(\mathbf{X}) = \rho(\mathbf{x}(\mathbf{X}, t_0), t_0)$. Therefore, an alternative form of the mass integral will be

$$m = \int_{V_0} \rho_0(\mathbf{X}) dV_0 \tag{18.75}$$

The principle of ***conservation of mass*** requires that mass can neither be created nor destroyed and thus the mass within a material volume will remain the same at any time. As a result, the rate of change in mass must vanish, i.e. $dm/dt = 0$. Using this on (18.74) then leads to

$$\frac{d}{dt}\left(\int_{V(t)} \rho(\mathbf{x}, t) dV\right) = 0 \tag{18.76}$$

The problem with this formulation is that the integration volume, in addition to the density, depends on the time and the computation will be difficult. However, we can carry out a coordinate transformation from the Eulerian to the Lagrangian frame by (18.70) to convert the integral from a moving volume to a reference volume as

$$\frac{d}{dt}\left(\int_{V_0} \rho(\mathbf{x}(\mathbf{X}, t), t) J dV_0\right) = 0 \tag{18.77}$$

Compared to (18.75), the density can now change with time as the condition is relaxed using (18.63a) to compute the current density using the reference coordinates. To proceed, we would like to do the time derivative under the integral sign and this requires us to know how to evaluate dJ/dt first, which is the time derivative of the Jacobian. The Jacobian as a determinant, can be expressed in the Index Notation

$$J = \epsilon_{ijk} \frac{\partial x_1}{\partial X_i} \frac{\partial x_2}{\partial X_j} \frac{\partial x_3}{\partial X_k} \tag{18.78}$$

via (18.27) where $\partial x_l / \partial X_i$ is the Jacobian in the matrix/tensor form. Since the partial derivatives in the Jacobian are defined with respect to the material reference coordinates, the time derivative of the Jacobian can be done directly. Using the Product Rule, we have

$$\frac{dJ}{dt} = \epsilon_{ijk} \left[\frac{d}{dt} \left(\frac{\partial x_1}{\partial X_i} \right) \frac{\partial x_2}{\partial X_j} \frac{\partial x_3}{\partial X_k} + \frac{\partial x_1}{\partial X_i} \frac{d}{dt} \left(\frac{\partial x_2}{\partial X_j} \right) \frac{\partial x_3}{\partial X_k} + \frac{\partial x_1}{\partial X_i} \frac{\partial x_2}{\partial X_j} \frac{d}{dt} \left(\frac{\partial x_3}{\partial X_k} \right) \right] \tag{18.79}$$

Note that in the material coordinates, by (18.71) the time derivative can be treated as a partial derivative, and by ***Clairaut's Theorem*** in Multivariable Calculus we can exchange the order of partial derivatives if they are smooth enough. Therefore, we have

$$\frac{d}{dt} \frac{\partial x_l}{\partial X_i} = \frac{\partial}{\partial t} \frac{\partial x_l}{\partial X_i} = \frac{\partial}{\partial X_i} \frac{\partial x_l}{\partial t} = \frac{\partial}{\partial X_i} \left(\frac{dx_l}{dt} \right) = \frac{\partial v_l}{\partial X_i} \tag{18.80}$$

Now, we invoke the Eulerian description and introduce the current coordinates, such that

$$\frac{\partial v_l}{\partial X_i} = \frac{\partial v_l}{\partial x_p} \frac{\partial x_p}{\partial X_i} \tag{18.81}$$

Take the first term in (18.79) to illustrate, it becomes

$$\epsilon_{ijk} \frac{d}{dt} \left(\frac{\partial x_1}{\partial X_i} \right) \frac{\partial x_2}{\partial X_j} \frac{\partial x_3}{\partial X_k} = \epsilon_{ijk} \frac{\partial v_1}{\partial X_i} \frac{\partial x_2}{\partial X_j} \frac{\partial x_3}{\partial X_k} = \epsilon_{ijk} \frac{\partial v_1}{\partial x_p} \frac{\partial x_p}{\partial X_i} \frac{\partial x_2}{\partial X_j} \frac{\partial x_3}{\partial X_k} \tag{18.82}$$

Notice that

$$\epsilon_{ijk}\frac{\partial x_p}{\partial X_i}\frac{\partial x_2}{\partial X_j}\frac{\partial x_3}{\partial X_k}$$

will only give the non-zero Jacobian determinant if $p = 1$, since when $p = 2, 3$ the rows of the determinant will be identical and hence

$$\epsilon_{ijk}\frac{\partial v_1}{\partial x_p}\frac{\partial x_p}{\partial X_i}\frac{\partial x_2}{\partial X_j}\frac{\partial x_3}{\partial X_k} = \frac{\partial v_1}{\partial x_1}\left(\epsilon_{ijk}\frac{\partial x_1}{\partial X_i}\frac{\partial x_2}{\partial X_j}\frac{\partial x_3}{\partial X_k}\right) = \frac{\partial v_1}{\partial x_1}J \tag{18.83}$$

where we have used the definition of J in (18.78). We can also show this by writing out the determinants in full as in (18.27):

$$\begin{aligned}
&\epsilon_{ijk}\left(\frac{\partial v_1}{\partial x_p}\frac{\partial x_p}{\partial X_i}\right)\frac{\partial x_2}{\partial X_j}\frac{\partial x_3}{\partial X_k}\\
&= \begin{vmatrix} \frac{\partial v_1}{\partial x_p}\frac{\partial x_p}{\partial X_1} & \frac{\partial v_1}{\partial x_p}\frac{\partial x_p}{\partial X_2} & \frac{\partial v_1}{\partial x_p}\frac{\partial x_p}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix}\\
&= \frac{\partial v_1}{\partial x_1}\begin{vmatrix} \frac{\partial x_1}{\partial X_1} & \frac{\partial x_1}{\partial X_2} & \frac{\partial x_1}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix} + \frac{\partial v_1}{\partial x_2}\begin{vmatrix} \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix} + \frac{\partial v_1}{\partial x_3}\begin{vmatrix} \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \\ \frac{\partial x_2}{\partial X_1} & \frac{\partial x_2}{\partial X_2} & \frac{\partial x_2}{\partial X_3} \\ \frac{\partial x_3}{\partial X_1} & \frac{\partial x_3}{\partial X_2} & \frac{\partial x_3}{\partial X_3} \end{vmatrix}\\
&= \frac{\partial v_1}{\partial x_1}J + \frac{\partial v_1}{\partial x_2}(0) + \frac{\partial v_1}{\partial x_3}(0) = \frac{\partial v_1}{\partial x_1}J\\
&\quad \text{(Properties 2.3.4)}
\end{aligned}$$

By the same reasoning, the second and third terms in (18.79) will equal to $(\partial v_2/\partial x_2)J$ and $(\partial v_3/\partial x_3)J$. So finally (18.79) is reduced to simply

$$\frac{dJ}{dt} = \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}\right)J = \frac{\partial v_q}{\partial x_q}J = (\nabla\cdot\mathbf{v})J \tag{18.84}$$

which becomes the conservation of mass in Jacobian form. Going back to (18.77), we can take the time derivative inside the integral as the reference volume is independent of time, leading to

$$\begin{aligned}
\frac{d}{dt}\left(\int_{V_0} \rho(\mathbf{x}(\mathbf{X},t),t) J dV_0\right) &= 0 \\
\int_{V_0} \frac{d}{dt}\left(\rho(\mathbf{x}(\mathbf{X},t),t) J\right) dV_0 &= 0 \\
\int_{V_0} (J\frac{d\rho}{dt} + \rho\frac{dJ}{dt}) dV_0 &= 0 \\
\int_{V_0} (J\frac{d\rho}{dt} + \rho(\nabla\cdot\mathbf{v})J) dV_0 &= 0 \qquad \text{(by (18.84))} \\
\int_{V_0} (\frac{d\rho}{dt} + \rho(\nabla\cdot\mathbf{v})) J dV_0 = \int_{V(t)} (\frac{d\rho}{dt} + \rho(\nabla\cdot\mathbf{v})) dV &= 0 \qquad (18.85)
\end{aligned}$$

where we move from the Lagrangian frame back to the Eulerian frame in the last line via (18.70). Since we can enclose the material volume arbitrarily, the integrand has to be identically zero, and hence the conservation of mass is expressed by the ***continuity equation***:

$$\frac{d\rho}{dt} + \rho(\nabla\cdot\mathbf{v}) = 0 \qquad (18.86)$$

By (18.73), we have the local form of the continuity equation:

$$\begin{aligned}
\frac{\partial\rho}{\partial t} + \mathbf{v}\cdot\nabla\rho + \rho(\nabla\cdot\mathbf{v}) &= 0 \\
\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) &= 0 \qquad (18.87)
\end{aligned}$$

[13] The physical interpretation of the second term $\nabla\cdot(\rho\mathbf{v})$ is the mass flux moving out of a locally fixed volume, which leads to a reduction in the local density via the first term. This can be seen by putting $\nabla\cdot(\rho\mathbf{v})$ inside an arbitrary volume

[13] $\nabla\cdot(\rho\mathbf{v}) = \partial_i(\rho v_i) = v_i(\partial_i\rho) + \rho(\partial_i v_i) = \mathbf{v}\cdot\nabla\rho + \rho(\nabla\cdot\mathbf{v})$.

integral and use (18.43):

$$\int_V \nabla \cdot (\rho \mathbf{v}) dV = \int_V \partial_i (\rho v_i) dV$$
$$= \int_S \rho v_i n_i dS = \int_S \rho \mathbf{v} \cdot \hat{n} dS \tag{18.88}$$

The continuity equation can also be cast in the Jacobian form. By multiplying (18.86) by J and use (18.84), we have

$$\frac{d\rho}{dt} J + \rho (\nabla \cdot \mathbf{v}) J = \frac{d\rho}{dt} J + \rho \frac{dJ}{dt} = \frac{d(\rho J)}{dt} = 0 \tag{18.89}$$

With this, we can derive a very powerful result that describes the change of the total amount of a quantity inside a volume, the ***Reynold's Transport Theorem***.

Theorem 18.3.1 (Reynold's Transport Theorem)**.** For an *extensive* property $A(\mathbf{x}, t)$ that is proportional to the mass and hence density $\rho(\mathbf{x}, t)$, the time derivative of its integral within a material volume can be moved inside and will operate only over the quantity $A(\mathbf{x}, t)$, that is

$$\frac{d}{dt} \int_{V(t)} \rho(\mathbf{x}, t) A(\mathbf{x}, t) dV = \int_{V(t)} \rho(\mathbf{x}, t) \frac{d}{dt} (A(\mathbf{x}, t)) dV \tag{18.90}$$

Proof. We will follow the same logic as in the previous derivation for the conservation of mass by first formulating the integral over a reference volume:

$$\frac{d}{dt} \int_{V(t)} \rho(\mathbf{x}, t) A(\mathbf{x}, t) dV = \frac{d}{dt} \int_{V_0} \rho(\mathbf{x}(\mathbf{X}, t), t) A(\mathbf{x}(\mathbf{X}, t), t) J dV_0 \tag{18.91}$$

And again the time derivative can be moved inside the integral to get

$$\frac{d}{dt} \int_{V_0} \rho(\mathbf{x}(\mathbf{X}, t), t) A(\mathbf{x}(\mathbf{X}, t), t) J dV_0$$
$$= \int_{V_0} \frac{d}{dt} (\rho J A) dV_0$$

$$
\begin{aligned}
&= \int_{V_0} \left(\frac{d}{dt}(\rho J)A + \rho J \frac{dA}{dt} \right) dV_0 \\
&= \int_{V_0} \rho J \frac{dA}{dt} dV_0 && (\frac{d}{dt}(\rho J) = 0 \text{ by } (18.89)) \quad (18.92)
\end{aligned}
$$

Finally, by the familiar (18.70) we can convert back to the current volume:

$$
\int_{V_0} \rho \frac{dA}{dt} J dV_0 = \int_{V(t)} \rho(\mathbf{x}, t) \frac{d}{dt}(A(\mathbf{x}, t)) dV \tag{18.93}
$$

so the result is established. □

There are two major variants of Reynold's Transport Theorem. First, consider $G(\mathbf{x}, t) = \rho A$ as a single quantity, then

$$
\begin{aligned}
\frac{d}{dt} \int_{V(t)} G(\mathbf{x}, t) dV &= \int_{V(t)} \left[\frac{d}{dt}(\rho A) - \frac{d\rho}{dt} A \right] dV && \text{(Product Rule)} \\
&= \int_{V(t)} \left[\frac{dG}{dt} + \rho(\nabla \cdot \mathbf{v}) A \right] dV && \text{(by (18.86))} \\
&= \int_{V(t)} \left[\frac{\partial G}{\partial t} + \mathbf{v} \cdot \nabla G + (\nabla \cdot \mathbf{v}) G \right] dV && \text{(by (18.73))} \\
&= \int_{V(t)} \left[\frac{\partial G}{\partial t} + \nabla \cdot (G\mathbf{v}) \right] dV && \text{(Footnote 13)} \\
&= \int_{V(t)} \frac{\partial G}{\partial t} dV + \int_{S(t)} G\mathbf{v} \cdot \hat{n} dS && \text{(by (18.44))}
\end{aligned}
\tag{18.94}
$$

The two terms on R.H.S. mean the local change of G within the material volume and the spread of G as the boundary expands respectively. They give rise to the net change of G in the material volume. Second, if the volume is not material but a so-called ***control volume*** (denoted by CV) that can itself deform and its boundary will change with time but is independent of the mass, i.e. does not follow the particles, then we can treat it as a material volume/surface

that coincides with the control volume/surface at time t but the velocity of the boundary is properly replaced. (18.94) then becomes

$$\frac{d}{dt}\int_{CV(t)} G(\mathbf{x},t)dV = \int_{CV(t)} \frac{\partial G}{\partial t}dV + \int_{S(t)} G\mathbf{v}_b \cdot \hat{n}dS \tag{18.95a}$$

where $\mathbf{v}_b$ is the velocity of the control surface. Meanwhile, for the original material volume (now denoted by MV), the integral is simply

$$\frac{d}{dt}\int_{MV(t)} G(\mathbf{x},t)dV = \int_{MV(t)} \frac{\partial G}{\partial t}dV + \int_{S(t)} G\mathbf{v}_m \cdot \hat{n}dS \tag{18.95b}$$

As we demand the material volume to match with the control volume at time t, subtracting the first equation from the second one will eliminate the local term, and gives

$$\begin{aligned}\frac{d}{dt}\int_{MV(t)} G(\mathbf{x},t)dV &= \frac{d}{dt}\int_{CV(t)} G(\mathbf{x},t)dV + \int_{S(t)} G(\mathbf{v}_m - \mathbf{v}_b)\cdot \hat{n}dS \\ &= \frac{d}{dt}\int_{CV(t)} G(\mathbf{x},t)dV + \int_{S(t)} G\mathbf{v}_r \cdot \hat{n}dS \end{aligned} \tag{18.96}$$

with $\mathbf{v}_r = \mathbf{v}_m - \mathbf{v}_b$ being the relative velocity of the material to control boundary.

Now we will use Reynold's Transport Theorem 18.3.1 to derive the *equation of motion*. Let $A = \mathbf{v}$ be the velocity field so that $\rho\mathbf{v}$ now represents the momentum per unit mass. The change in the total momentum of a material volume is caused by two major effects: the external traction $\mathbf{T}$ on the material boundary and the *body force* (exerted at a distance, per unit mass) $\mathbf{b}$ over the entire volume:

$$\frac{d}{dt}\int_V \rho\mathbf{v}dV = \int_S \mathbf{T}dS + \int_V \rho\mathbf{b}dV \tag{18.97}$$

By the Reynold's Transport Theorem (18.90) and Cauchy's formula (18.48), it becomes

$$\int_V \rho\frac{dv_j}{dt}dV = \int_S \sigma_{ij}n_i dS + \int_V \rho b_j dV \tag{18.98}$$

written in the Index Notation. By (18.43), we can rewrite the traction term to obtain

$$\int_V \rho \frac{dv_j}{dt} dV = \int_V \frac{\partial \sigma_{ij}}{\partial x_i} dV + \int_V \rho b_j dV$$

$$\int_V \left(\rho \frac{dv_j}{dt} - \left(\frac{\partial \sigma_{ij}}{\partial x_i} + \rho b_j \right) \right) dV = 0 \tag{18.99}$$

Again, as this holds for any material volume, the integrand has to be identically zero, which gives us the desired ***Equation(s) of Motion***:

$$\rho \frac{dv_j}{dt} - \left(\frac{\partial \sigma_{ij}}{\partial x_i} + \rho b_j \right) = 0$$

$$\rho \frac{dv_j}{dt} = \frac{\partial \sigma_{ij}}{\partial x_i} + \rho b_j \tag{18.100}$$

In the ocean or atmosphere, we often assume the fluid is *inviscid* so it cannot support shear stress, although it may be necessary to include viscous forces otherwise. The stress tensor is then reduced to the *hydrostatic pressure* $\sigma_{ij} = -p\delta_{ij}$ which has the same magnitude acting in all directions. Subsequently, (18.100) becomes

$$\rho \frac{dv_j}{dt} = -\frac{\partial (p\delta_{ij})}{\partial x_i} + \rho b_j$$

$$\rho \frac{dv_j}{dt} = -\frac{\partial p}{\partial x_j} + \rho b_j \implies \rho \frac{d\mathbf{v}}{dt} = -\nabla p + \rho \mathbf{b} \tag{18.101}$$

which are commonly referred to as the *Euler's equations* that can be explicitly expressed in full form for each component:

$$\begin{cases} \rho \left(\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} \right) = -\frac{\partial p}{\partial x} + \rho g_x & \text{(18.102a)} \\ \rho \left(\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} \right) = -\frac{\partial p}{\partial y} + \rho g_y & \text{(18.102b)} \\ \rho \left(\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} \right) = -\frac{\partial p}{\partial z} + \rho g_z & \text{(18.102c)} \end{cases}$$

where we denote the three-dimensional flow velocities as $\mathbf{v} = (u, v, w)^T$, the body force is the gravity $\mathbf{b} = \mathbf{g} = (g_x, g_y, g_z)^T$, and use (18.73) to expand the material time derivative. Note that we are doing the above analysis in a non-rotating inertial coordinate frame so the usual Coriolis force does not appear.

Finally, we will use Euler's equations to derive ***Kelvin's Circulation Theorem***. *Circulation* along a closed material path is the line integral of velocity along it:

$$\Gamma_c = \oint_{C(t)} v_i dx_i \tag{18.103}$$

Its time derivative is

$$\begin{aligned}
\frac{d\Gamma_c}{dt} &= \frac{d}{dt}\left(\oint_C v_i dx_i\right) \\
&= \oint_C \frac{d}{dt}(v_i dx_i) \\
&\quad \text{(Take the material derivative inside and enclose the material} \\
&\quad \text{line elements to account for the moving material loop)} \\
&= \oint_C \frac{dv_i}{dt} dx_i + \oint_C v_i d\left(\frac{dx_i}{dt}\right) \\
&\quad \text{(Material line elements move with the flow} \\
&\quad \text{and thus } \frac{d}{dt}(dx_i) = d\left(\frac{dx_i}{dt}\right)) \\
&= \oint_C \frac{dv_i}{dt} dx_i + \oint_C v_i dv_i \\
&= \oint_C \frac{dv_i}{dt} dx_i + \oint_C d\left(\frac{\|\mathbf{v}\|^2}{2}\right)
\end{aligned} \tag{18.104}$$

The second integral is zero because the value of a scalar in a closed path must be unchanged after a loop. Substitute (18.101) into the first integral yields

$$\oint_C \frac{dv_i}{dt} dx_i = \oint_C \left(-\frac{1}{\rho}\frac{\partial p}{\partial x_i} + g_i\right) dx_i$$

$$
\begin{aligned}
&= -\oint_C \frac{1}{\rho}\frac{\partial p}{\partial x_i} dx_i + \oint_C g_i dx_i \\
&= -\oint_C \frac{dp}{\rho} + (0)
\end{aligned}
$$

where the second smaller integral vanishes because gravity is treated as a constant vector and the closed path integral will cancel out itself and Kelvin's Circulation Theorem is established:

$$
\frac{d\Gamma_c}{dt} = -\oint_C \frac{dp}{\rho} \tag{18.105}
$$

Another form of the theorem is

$$
\frac{d\Gamma_c}{dt} = \int_S \frac{\nabla\rho \times \nabla p}{\rho^2} \cdot \hat{n} dS \tag{18.106}
$$

due to Stokes's Theorem adapted from (18.45)[14]:

$$
\begin{aligned}
-\oint_C \frac{dp}{\rho} &= -\oint_C \frac{\nabla p \cdot dl}{\rho} \\
&= -\int_S \nabla \times (\frac{\nabla p}{\rho}) \cdot \hat{n} dS
\end{aligned}
$$

[14]We reduce the conversion of integrals to that between a curve and surface. For a vector field $\mathbf{u}$, denote the normal vectors to the surface element and line element as $\mathbf{n}$ and $\mathbf{m}$ respectively, and the tangential unit vector as $\mathbf{t}$, then $\mathbf{n} \times \mathbf{m} = \mathbf{t}$ and (18.43) can be appropriately turned into

$$
\begin{aligned}
\int_S (\nabla \times \mathbf{u}) \cdot \hat{n} dS &= \int_S \epsilon_{ijk} \partial_j u_k n_i dS \\
&= \oint_C \epsilon_{ijk} u_k n_i m_j dl \\
&= \oint_C u_k (\epsilon_{kij} n_i m_j) dl = \oint_C u_k t_k dl \\
&= \oint_C u_k dl_k = \oint_C \mathbf{u} \cdot d\mathbf{l}
\end{aligned}
$$

Now take $\mathbf{u} = \nabla p / \rho$.

$$= \int_S \frac{\nabla\rho \times \nabla p}{\rho^2} \cdot \hat{n} dS$$

[15] The quantity

$$\frac{\nabla\rho \times \nabla p}{\rho^2} \tag{18.107}$$

is called the *baroclinicity* which characterizes the tilting between the pressure and density contours. It often arises from a temperature gradient in a cross-section, which will generate circulation along the corresponding path.

18.4. Earth System Applications: Data Assimilation

Reference Materials: [5]

18.4.1. Basic Ideas behind Data Assimilation

Data Assimilation (DA) is a technique to integrate *observations* into the *background* state of a weather model to produce a more accurate *analysis* field. This is needed as the error of the model will inevitably grow in time and become useless if it is not supplied with real information. To motivate, we will use air

[15] By Product Rule,

$$\begin{aligned}\nabla \times (\frac{\nabla p}{\rho}) &= \epsilon_{ijk}\partial_j(\frac{\partial_k p}{\rho}) \\ &= \epsilon_{ijk}(\frac{\partial_j\partial_k p}{\rho}) + \epsilon_{ijk}(-\frac{\partial_k p}{\rho^2})\partial_j\rho\end{aligned}$$

The first term is the zero vector since the partial derivatives are symmetric but the epsilon symbol is antisymmetric:

temperature at a fixed grid point as a single variable (*univariate*) example. Let the true temperature be T_t, and the background temperature from the model output be

$$T_b = T_t + \varepsilon_b \tag{18.108a}$$

where ε_b is the background error with a variance of

$$E[\varepsilon_b^2] = \sigma_b^2 \tag{18.108b}$$

The variance here is formulated by integrating over the *probability distribution function (PDF)* and is a generalization of that in Definition 11.2.1. Similarly, the observation temperature is

$$T_o = T_t + \varepsilon_o \tag{18.109a}$$

and the observation error has a variance of

$$E[\varepsilon_o^2] = \sigma_o^2 \tag{18.109b}$$

Further, we assume that they are *unbiased* such that

$$E[T_b] = T_t \qquad \text{or} \qquad E[\varepsilon_b] = 0 \tag{18.110a}$$

$$E[T_o] = T_t \qquad \text{or} \qquad E[\varepsilon_o] = 0 \tag{18.110b}$$

Moreover, the background and observation errors have to be *uncorrelated*:

$$E[\varepsilon_b \varepsilon_o] = 0 \tag{18.111}$$

$$\epsilon_{ijk}\left(\frac{\partial_j \partial_k p}{\rho}\right) = \epsilon_{ikj}\left(\frac{\partial_k \partial_j p}{\rho}\right)$$
$$= -\epsilon_{ijk}\left(\frac{\partial_k \partial_j p}{\rho}\right) = -\epsilon_{ijk}\left(\frac{\partial_j \partial_k p}{\rho}\right) \implies \epsilon_{ijk}\left(\frac{\partial_j \partial_k p}{\rho}\right) = \mathbf{0}$$

and the second term leads to

$$-\epsilon_{ijk}\left(\frac{\partial_k p}{\rho^2}\right)\partial_j \rho = -\epsilon_{ijk}\frac{\partial_j \rho \partial_k p}{\rho^2} = -\frac{\nabla \rho \times \nabla p}{\rho^2}$$

We set out to find an optimal linear combination of T_b and T_o as a new unbiased estimation, referred to as the *analysis*:

$$T_a = a_1 T_b + a_2 T_o \tag{18.112}$$

where $a_1, a_2 > 0$ are the weights. The analysis will also have an error of ε_a such that

$$T_a = T_t + \varepsilon_a \tag{18.113}$$

The condition of being unbiased requires that

$$\begin{aligned}
E[\varepsilon_a] = E[T_a - T_t] &= 0 \\
E[a_1 T_b + a_2 T_o - T_t] &= 0 \qquad \text{(By (18.112))} \\
E[a_1(T_t + \varepsilon_b) + a_2(T_t + \varepsilon_o) - T_t] &= 0 \\
\text{(By (18.108a) and (18.109a))} & \\
a_1(E[T_t] + E[\varepsilon_b]) + a_2(E[T_t] + E[\varepsilon_o]) - E[T_t] &= 0 \\
(a_1 + a_2 - 1)E[T_t] &= 0 \qquad \text{(By (18.110))} \\
\implies a_1 + a_2 &= 1 \qquad (18.114)
\end{aligned}$$

i.e. the sum of weights has to be unity (1). To obtain the best analysis estimation T_a, we want the weighting to minimize the mean squared error of T_a, a.k.a. the analysis error variance

$$\sigma_a^2 = E[\varepsilon_a^2] = E[(T_a - T_t)^2] \tag{18.115}$$

We proceed by differentiating it with respect to one of the weights, let's say a_1, and the minimum will occur when the partial derivative is zero:

$$\frac{\partial \sigma_a^2}{\partial a_1} = 0 \tag{18.116}$$

Now

$$\varepsilon_a = T_a - T_t = a_1 T_b + a_2 T_o - (a_1 + a_2) T_t \qquad (a_1 + a_2 = 1)$$

$$
\begin{aligned}
&= a_1(T_b - T_t) + a_2(T_o - T_t) \\
&= a_1\varepsilon_b + a_2\varepsilon_o
\end{aligned} \tag{18.117}
$$

and hence

$$
\begin{aligned}
\sigma_a^2 = E[\varepsilon_a^2] &= E[(T_a - T_t)^2] \\
&= E[(a_1\varepsilon_b + a_2\varepsilon_o)^2] \\
&= E[a_1^2\varepsilon_b^2 + a_1a_2\varepsilon_b\varepsilon_o + a_2^2\varepsilon_o^2] \\
&= a_1^2E[\varepsilon_b^2] + a_1a_2E[\varepsilon_b\varepsilon_o] + a_2^2E[\varepsilon_o^2] \\
&= a_1^2\sigma_b^2 + a_2^2\sigma_o^2 && \text{(By (18.111))} \\
&= a_1^2\sigma_b^2 + (1 - a_1)^2\sigma_o^2 && (18.118)
\end{aligned}
$$

Substituting this into (18.116), we obtain

$$
\begin{aligned}
\frac{\partial\sigma_a^2}{\partial a_1} = \frac{\partial}{\partial a_1}(a_1^2\sigma_b^2 + (1 - a_1)^2\sigma_o^2) &= 0 \\
2a_1\sigma_b^2 - 2(1 - a_1)\sigma_o^2 &= 0 \\
2a_1(\sigma_b^2 + \sigma_o^2) - 2\sigma_o^2 &= 0 \\
a_1 = \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2} &= \frac{\sigma_b^{-2}}{\sigma_b^{-2} + \sigma_o^{-2}} && (18.119)
\end{aligned}
$$

where the second equality in the last line is from dividing by $\sigma_b^2\sigma_o^2$. The reciprocals of the background/observation error variances $\sigma_b^{-2}, \sigma_o^{-2}$ represent their respective *precisions*. The best analysis is then given by

$$
\begin{aligned}
T_a &= a_1T_b + (1 - a_1)T_o \\
&= \frac{\sigma_o^2}{\sigma_b^2 + \sigma_o^2}T_b + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}T_o = \frac{\sigma_b^{-2}T_b + \sigma_o^{-2}T_o}{\sigma_b^{-2} + \sigma_o^{-2}}
\end{aligned} \tag{18.120}
$$

Therefore, the analysis is a weighted average of the background and the observation where the weights are proportional to their precisions. The analysis error variance in (18.118) is subsequently

$$
\sigma_a^2 = a_1^2\sigma_b^2 + (1 - a_1)^2\sigma_o^2
$$

$$
\begin{aligned}
&= (\frac{\sigma_o^2}{\sigma_b^2+\sigma_o^2})^2\sigma_b^2 + (\frac{\sigma_b^2}{\sigma_b^2+\sigma_o^2})^2\sigma_o^2 \\
&= \frac{\sigma_b^2\sigma_o^2(\sigma_b^2+\sigma_o^2)}{(\sigma_b^2+\sigma_o^2)^2} \\
&= \frac{\sigma_b^2\sigma_o^2}{\sigma_b^2+\sigma_o^2} = \frac{1}{\sigma_b^{-2}+\sigma_o^{-2}}
\end{aligned} \tag{18.121}
$$

or alternatively,

$$
\frac{1}{\sigma_a^2} = \frac{1}{\sigma_b^2} + \frac{1}{\sigma_o^2} \tag{18.122}
$$

which means that the precision of the analysis is the sum of the precisions of the background and observation. An illustration is given in Figure 18.7 below. By rearranging the expressions in (18.120) and (18.121), the data assimilation prototype is described by

$$
W = \frac{\sigma_b^2}{\sigma_b^2+\sigma_o^2} \tag{18.123a}
$$

$$
T_a = T_b + W(T_o - T_b) \tag{18.123b}
$$

$$
\sigma_a^2 = (1-W)\sigma_b^2 \tag{18.123c}
$$

18.4.2. Optimal Interpolation

In the last subsection, we only use a single variable as the illustrating example. Obviously, in actual weather models, there will be multiple variables (let's say N) like temperature, pressure, density, and three-dimensional winds, to be simulated in a large number of grid points (M). Hence there will be a total of $L = MN$ variables in the model output and we will place all of them into a *state*

Figure 18.7.: *The probability density distribution of the background/observation/analysis temperatures in the section example where the errors are assumed to have a Gaussian shape.*

vector of length $L = MN$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix} \tag{18.124}$$

with a corresponding error vector of

$$\boldsymbol{\epsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_L \end{bmatrix} \tag{18.125}$$

which is assumed to be unbiased too, such that the expected values of the errors $E(\boldsymbol{\epsilon}) = \mathbf{0}$ is always a zero vector. The *error covariance matrix* is then defined as

$$\begin{aligned} P &= E[(\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon}))(\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon}))^T] = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &= \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & \cdots & E[\varepsilon_1\varepsilon_L] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & & E[\varepsilon_1\varepsilon_L] \\ \vdots & & \ddots & \vdots \\ E[\varepsilon_L\varepsilon_1] & E[\varepsilon_L\varepsilon_2] & \cdots & E[\varepsilon_L\varepsilon_L] \end{bmatrix} \end{aligned} \tag{18.126}$$

which is generalized from Properties 11.2.6. We will start with the simplest case where the observations are exactly the model variables. Then the background, analysis, and observation states can be expressed as

$$\mathbf{x}_b = \mathbf{x}_t + \boldsymbol{\epsilon}_b \tag{18.127a}$$

$$\mathbf{x}_a = \mathbf{x}_t + \boldsymbol{\epsilon}_a \tag{18.127b}$$

$$\mathbf{x}_o = \mathbf{x}_t + \boldsymbol{\epsilon}_o \tag{18.127c}$$

relative to the truth $\mathbf{x}_t$. The error covariance matrices for the background, analysis, and observation are

$$P_b = B = E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] \tag{18.128a}$$

$$P_a = A = E[\boldsymbol{\epsilon}_a\boldsymbol{\epsilon}_a^T] \tag{18.128b}$$

$$P_o = R = E[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T] \tag{18.128c}$$

As before, we assume that the background and observation errors are uncorrelated so that

$$E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_o^T] = E[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_b^T] = \mathbf{0} \tag{18.129}$$

We also require the analysis $\mathbf{x}_a$ to be an unbiased linear combination of $\mathbf{x}_b$ and $\mathbf{x}_o$. This allows us to write

$$\mathbf{x}_a = \mathbf{x}_b + K(\mathbf{x}_o - \mathbf{x}_b) \tag{18.130}$$

akin to (18.123b) where K is the *Kalman gain* matrix that represents the optimal weight. Therefore, the analysis will be the background plus the optimal weight times the observation increment (*innovation*). Subsequently,

$$\begin{aligned} \mathbf{x}_t + \boldsymbol{\epsilon}_a &= \mathbf{x}_t + \boldsymbol{\epsilon}_b + K(\mathbf{x}_o - \mathbf{x}_t - \boldsymbol{\epsilon}_b) \\ \boldsymbol{\epsilon}_a &= \boldsymbol{\epsilon}_b + K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b) \end{aligned} \tag{18.131}$$

The total error variance $E[\boldsymbol{\epsilon}_a^T \boldsymbol{\epsilon}_a]$, or equivalently the trace of A, by Index Notation, is

$$E[\boldsymbol{\epsilon}_a^T \boldsymbol{\epsilon}_a] = E[(\varepsilon_a)_i (\varepsilon_a)_i)] \tag{18.132}$$

Similar to (18.116), we want to differentiate this quantity but with respect to the Kalman gain matrix K and demand that

$$\frac{\partial E[\boldsymbol{\epsilon}_a^T \boldsymbol{\epsilon}_a]}{\partial K} = E\left(\frac{\partial [\boldsymbol{\epsilon}_a^T \boldsymbol{\epsilon}_a]}{\partial K}\right) = [\mathbf{0}] \tag{18.133}$$

To proceed, we first note that it works very much like (18.41) but with two Kronecker deltas now:

$$\frac{\partial K_{ij}}{\partial K_{pq}} = \delta_{ip}\delta_{jq} \tag{18.134}$$

Now, writing out the subscripts in (18.131), we have

$$(\varepsilon_a)_i (\varepsilon_a)_i = [(\varepsilon_b)_i + K_{ij}((\varepsilon_o)_j - (\varepsilon_b)_j)][(\varepsilon_b)_i + K_{ik}((\varepsilon_o)_k - (\varepsilon_b)_k)] \tag{18.135}$$

where we have to use two different dummy indices for the product due to the rule of Index Notation. Then the derivative is computed to be

$$\begin{aligned} &\frac{\partial [\boldsymbol{\epsilon}_a^T \boldsymbol{\epsilon}_a]}{\partial K} \\ &= \frac{\partial}{\partial K_{pq}} \left([(\varepsilon_b)_i + K_{ij}((\varepsilon_o)_j - (\varepsilon_b)_j)][(\varepsilon_b)_i + K_{ik}((\varepsilon_o)_k - (\varepsilon_b)_k)]\right) \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\partial}{\partial K_{pq}}[(\varepsilon_b)_i + K_{ij}((\varepsilon_o)_j - (\varepsilon_b)_j)]\right)[(\varepsilon_b)_i + K_{ik}((\varepsilon_o)_k - (\varepsilon_b)_k)] \\
&\quad + [(\varepsilon_b)_i + K_{ij}((\varepsilon_o)_j - (\varepsilon_b)_j)]\left(\frac{\partial}{\partial K_{pq}}[(\varepsilon_b)_i + K_{ik}((\varepsilon_o)_k - (\varepsilon_b)_k)]\right) \\
&= \delta_{ip}\delta_{jq}((\varepsilon_o)_j - (\varepsilon_b)_j)[(\varepsilon_b)_i + K_{ik}((\varepsilon_o)_k - (\varepsilon_b)_k)] \\
&\quad + ((\varepsilon_b)_i + K_{ij}((\varepsilon_o)_j - (\varepsilon_b)_j))\delta_{ip}\delta_{kq}((\varepsilon_o)_k - (\varepsilon_b)_k) \\
&= ((\varepsilon_o)_q - (\varepsilon_b)_q)[(\varepsilon_b)_p + K_{pk}((\varepsilon_o)_k - (\varepsilon_b)_k)] \\
&\quad + [(\varepsilon_b)_p + K_{pj}((\varepsilon_o)_j - (\varepsilon_b)_j)]((\varepsilon_o)_q - (\varepsilon_b)_q) \\
&= 2((\varepsilon_b)_p + K_{pk}((\varepsilon_o)_k - (\varepsilon_b)_k))((\varepsilon_o)_q - (\varepsilon_b)_q) \\
&\quad \text{(One of the dummy indices can be changed and the terms will be grouped)} \\
&= 2(\varepsilon_b)_p((\varepsilon_o)_q - (\varepsilon_b)_q) + 2K_{pk}((\varepsilon_o)_k - (\varepsilon_b)_k)((\varepsilon_o)_q - (\varepsilon_b)_q) \\
&= 2\boldsymbol{\epsilon}_b(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T + 2K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T
\end{aligned} \tag{18.136}
$$

where in the last step we convert it back to matrix form. Plugging it into (18.133) gives the desired expression for the Kalman gain matrix.

$$
\begin{aligned}
E\left(\frac{\partial[\boldsymbol{\epsilon}_a^T\boldsymbol{\epsilon}_a]}{\partial K}\right) &= E[2\boldsymbol{\epsilon}_b(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T + 2K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T] = [\mathbf{0}] \\
E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_o^T] - E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] + K(E[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T] - E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_o^T] - E[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_b^T] + E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T]) &= [\mathbf{0}] \\
-E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] + K(E[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T] + E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T]) &= [\mathbf{0}] \\
&\text{(Uncorrelated errors by (18.129))} \\
-B + K(B + R) &= [\mathbf{0}] \\
\implies K &= B(B + R)^{-1}
\end{aligned} \tag{18.137}
$$

The analysis covariance is

$$
\begin{aligned}
A &= E[\boldsymbol{\epsilon}_a\boldsymbol{\epsilon}_a^T] \\
&= E[(\boldsymbol{\epsilon}_b + K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b))(\boldsymbol{\epsilon}_b + K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b))^T] \\
&= E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] + E[\boldsymbol{\epsilon}_b(K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b))^T]
\end{aligned}
$$

$$
\begin{aligned}
&+ E[K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)\boldsymbol{\epsilon}_b^T] + E[K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)(K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b))^T] \\
&= E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] + E[\boldsymbol{\epsilon}_b(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T K^T] \\
&\quad + E[K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)\boldsymbol{\epsilon}_b^T] + E[K(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)(\boldsymbol{\epsilon}_o - \boldsymbol{\epsilon}_b)^T K^T] \\
&= B - BK^T - KB + K(B+R)K^T \qquad \text{(Uncorrelated errors by (18.129))} \\
&= B - BK^T - KB + B(B+R)^{-1}(B+R)K^T \qquad \text{(by (18.137))} \\
&= B - BK^T - KB + BK^T = B - KB = (I-K)B
\end{aligned} \tag{18.138}
$$

(18.130), (18.137), and (18.138) then constitute the *multivariate* ***Optimal Interpolation*** scheme that can be compared to (18.123). An alternative form for K and A can be derived as follows.

$$
\begin{aligned}
K = B(B+R)^{-1} &= B[(RR^{-1})B + R(B^{-1}B)]^{-1} \\
&= B[R(R^{-1} + B^{-1})B]^{-1} \\
&= BB^{-1}(B^{-1} + R^{-1})^{-1}R^{-1} = (B^{-1} + R^{-1})^{-1}R^{-1}
\end{aligned} \tag{18.139}
$$

$$
\begin{aligned}
A = B - KB &= B - (B^{-1} + R^{-1})^{-1}R^{-1}B \\
&= (B^{-1} + R^{-1})^{-1}(B^{-1} + R^{-1})B - (B^{-1} + R^{-1})^{-1}R^{-1}B \\
&= (B^{-1} + R^{-1})^{-1}((B^{-1} + R^{-1})B - R^{-1}B) \\
&= (B^{-1} + R^{-1})^{-1}(B^{-1}B + R^{-1}B - R^{-1}B) \\
&= (B^{-1} + R^{-1})^{-1}I = (B^{-1} + R^{-1})^{-1} \\
\implies A^{-1} &= B^{-1} + R^{-1}
\end{aligned} \tag{18.140}
$$

Same as the univariate case, the inverses of covariance matrices represent the precisions and the analysis precision is just the sum of those of the background and observation.

However, most of the time, the observations are not exactly the model variables or measured at different locations from the grid points. Let there be G observations stored in a vector $\mathbf{y}_o$, and it is related to the model variables by the observation operator H according to $\mathbf{y} = H(\mathbf{x})$. Then (a) and (b) of (18.127) remain the

same but (c) is now

$$\mathbf{y}_o = H(\mathbf{x}_t) + \boldsymbol{\epsilon}_o \tag{18.141}$$

where $\mathbf{y}_o$ and $\boldsymbol{\epsilon}_o$ are vectors of length G. Due to the scope limit, we will only discuss the case where the observation operator H is linear and hence a matrix such that

$$\mathbf{y} = H(\mathbf{x}) = H\mathbf{x} \tag{18.142}$$

Then the innovation vector will be

$$\begin{aligned}\mathbf{d} &= \mathbf{y}_o - H\mathbf{x}_b \\ &= \mathbf{y}_o - H(\mathbf{x}_t + \boldsymbol{\epsilon}_b) \\ &= \mathbf{y}_o - H\mathbf{x}_t - H\boldsymbol{\epsilon}_b \\ &= \boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b \end{aligned} \tag{18.143}$$

Similar to the previous formulation, we compute the analysis as the background plus the optimal weight times the innovation:

$$\mathbf{x}_a = \mathbf{x}_b + K(\mathbf{y}_o - H\mathbf{x}_b) = \mathbf{x}_b + K\mathbf{d} \tag{18.144}$$

$$\mathbf{x}_t + \boldsymbol{\epsilon}_a = \mathbf{x}_t + \boldsymbol{\epsilon}_b + K\mathbf{d} \implies \boldsymbol{\epsilon}_a = \boldsymbol{\epsilon}_b + K\mathbf{d} \tag{18.145}$$

We demand the same as in (18.133). However, note that the only difference of the above expression from (18.131) is that $\mathbf{x}_o - \mathbf{x}_b$ is replaced by $\mathbf{d}$ and (18.136) simply becomes

$$\frac{\partial(\boldsymbol{\epsilon}_a^T\boldsymbol{\epsilon}_a)}{\partial K} = 2\boldsymbol{\epsilon}_b\mathbf{d}^T + 2K\mathbf{d}\mathbf{d}^T \tag{18.146}$$

As before, we require the expected value of this derivative to be zero, and hence

$$\frac{\partial E[\boldsymbol{\epsilon}_a^T\boldsymbol{\epsilon}_a]}{\partial K} = E[2\boldsymbol{\epsilon}_b\mathbf{d}^T + 2K\mathbf{d}\mathbf{d}^T] = [\mathbf{0}]$$

$$
\begin{aligned}
E[\boldsymbol{\epsilon}_b(\boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b)^T + K(\boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b)(\boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b)^T] &= [\mathbf{0}] \\
-E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T H^T] + KE[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T] + KE[H\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T H^T] &= [\mathbf{0}] \\
\text{(Uncorrelated errors by (18.129))} \\
-BH^T + K(R + HBH^T) &= [\mathbf{0}] \\
\implies K &= BH^T(HBH^T + R)^{-1}
\end{aligned}
\tag{18.147}
$$

The analysis covariance is now

$$
\begin{aligned}
A &= E[\boldsymbol{\epsilon}_a\boldsymbol{\epsilon}_a^T] \\
&= E[(\boldsymbol{\epsilon}_b + K(\boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b))(\boldsymbol{\epsilon}_b + K(\boldsymbol{\epsilon}_o - H\boldsymbol{\epsilon}_b))^T] \\
&= E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] - E[KH\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] - E[\boldsymbol{\epsilon}_b(KH\boldsymbol{\epsilon}_b)^T] \\
&\quad + E[(K\boldsymbol{\epsilon}_o)(K\boldsymbol{\epsilon}_o)^T] + E[KH\boldsymbol{\epsilon}_b(KH\boldsymbol{\epsilon}_b)^T] \\
&\quad \text{(Uncorrelated errors by (18.129))} \\
&= E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] - E[KH\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] - E[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T H^T K^T] \\
&\quad + E[K\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T K^T] + E[KH\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T H^T K^T] \\
&= B - KHB - BH^T K^T + KRK^T + KHBH^T K^T \\
&= B - KHB - BH^T K^T + K(HBH^T + R)K^T \\
&= B - KHB - BH^T K^T + BH^T(HBH^T + R)^{-1}(HBH^T + R)K^T \\
&\quad \text{(by (18.147))} \\
&= B - KHB - BH^T K^T + BH^T K^T \\
&= B - KHB = (I - KH)B
\end{aligned}
\tag{18.148}
$$

Similar to (18.139), we can express K in an alternative form. We may be tempted to write

$$
\begin{aligned}
K &= BH^T(HBH^T + R)^{-1} \\
&= BH^T[(H^T R^{-1})^{-1}(H^T R^{-1})HBH^T + R(BH^T)^{-1}(BH^T)]^{-1}
\end{aligned}
$$

just like (18.139). However, H will be a non-square $G \times L$ matrix, hence $H^T R^{-1}$ and BH^T are also not square, and their inverses are not defined. To proceed, we

need the ***Sherman-Morrison-Woodbury Formula***:

$$(A + SXC)^{-1} = A^{-1} - A^{-1}S(X^{-1} + CA^{-1}S)^{-1}CA^{-1} \tag{18.149}$$

This can be derived by comparing the upper-left block of (8.14) and (8.16) with D replaced by $-X^{-1}$ and B replaced by S. Now substitute $A = R$, $S = H$, $X = B$, $C = H^T$ to get

$$(HBH^T + R)^{-1} = R^{-1} - R^{-1}H(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \tag{18.150}$$

Plugging it into (18.147), we have

$$\begin{aligned}
K &= BH^T(HBH^T + R)^{-1} \\
&= BH^T(R^{-1} - R^{-1}H(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1}) \\
&= BH^T R^{-1} - BH^T R^{-1}H(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \\
&= B(B^{-1} + H^T R^{-1}H)(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \\
&\quad - BH^T R^{-1}H(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \\
&= (B(B^{-1} + H^T R^{-1}H) - BH^T R^{-1}H)(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \\
&= (BB^{-1} + BH^T R^{-1}H - BH^T R^{-1}H)(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} \\
&= I(B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1} = (B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1}
\end{aligned} \tag{18.151}$$

and the analysis covariance can be rewritten as

$$\begin{aligned}
A = B - KHB &= B - (B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1}HB \\
&= (B^{-1} + H^T R^{-1}H)^{-1}(B^{-1} + H^T R^{-1}H)B \\
&\quad - (B^{-1} + H^T R^{-1}H)^{-1}H^T R^{-1}HB \\
&= (B^{-1} + H^T R^{-1}H)^{-1}((B^{-1} + H^T R^{-1}H)B - H^T R^{-1}HB) \\
&= (B^{-1} + H^T R^{-1}H)^{-1}(B^{-1}B + H^T R^{-1}HB - H^T R^{-1}HB) \\
&= (B^{-1} + H^T R^{-1}H)^{-1} \implies A^{-1} = B^{-1} + H^T R^{-1}H
\end{aligned} \tag{18.152}$$

18.4.3. Kalman Filter

Notice that the technique of Optimal Interpolation introduced in the last subsection does not depend on time and can be used at any time step in an isolated fashion. However, in physical models, we need to integrate forwards in time for the forecast. We may ask if there is an improved version of Optimal Interpolation such that it will take the evolution of the model state into account. Particularly, the background error is previously assumed to be static but should actually change in time, e.g. different climatologies in different seasons. ***Kalman Filter*** solves this by further cycling the background error covariance using the model. It is essentially Optimal Interpolation but with the error covariances evolved forward in time as well. No matter how the background error is initially determined, it will become more and more optimal after several cycles as long as the model and observations are good. Kalman Filter consists of two steps: the *analysis step* which is identical to Optimal Interpolation, and the *forecast step* where we carry out time integration for the state vector and the error covariance matrix in the model.

In Kalman filter, a short-term model forecast is used as the background and the subscript b is now replaced by f, and similar to (18.127) plus (18.141), we have

$$\mathbf{x}_f^{[i]} = \mathbf{x}_t^{[i]} + \boldsymbol{\epsilon}_f^{[i]} \tag{18.153a}$$

$$\mathbf{x}_a^{[i]} = \mathbf{x}_t^{[i]} + \boldsymbol{\epsilon}_a^{[i]} \tag{18.153b}$$

$$\mathbf{y}_o^{[i]} = H\mathbf{x}_t^{[i]} + \boldsymbol{\epsilon}_o^{[i]} \tag{18.153c}$$

where the superscript $^{[i]}$ indicates the i-th time step. Also (18.128) is now

$$P_f^{[i]} = E[\boldsymbol{\epsilon}_f^{[i]}\boldsymbol{\epsilon}_f^{[i]T}] \tag{18.154a}$$

$$P_a^{[i]} = E[\boldsymbol{\epsilon}_a^{[i]}\boldsymbol{\epsilon}_a^{[i]T}] \tag{18.154b}$$

$$R^{[i]} = E[\boldsymbol{\epsilon}_o^{[i]}\boldsymbol{\epsilon}_o^{[i]T}] \tag{18.154c}$$

The forward integration model M is assumed to be linear such that the forecast at the next time step is given by

$$\mathbf{x}_f^{[i+1]} = M^{[i]}\mathbf{x}_a^{[i]} \tag{18.155a}$$

$$\mathbf{x}_t^{[i+1]} = M^{[i]}\mathbf{x}_t^{[i]} - \boldsymbol{\eta}^{[i]} \tag{18.155b}$$

where $M^{[i]}$ is the corresponding forward mapping at the i-th time step and $\boldsymbol{\eta}^{[i]}$ represents the inherent model error vector. We assume that

$$E[\boldsymbol{\eta}^{[i]}] = \mathbf{0} \tag{18.156a}$$

$$E[\boldsymbol{\epsilon}_a^{[i]}\boldsymbol{\eta}^{[i]T}] = E[\boldsymbol{\eta}^{[i]}\boldsymbol{\epsilon}_a^{[i]T}] = [\mathbf{0}] \tag{18.156b}$$

$$Q^{[i]} = E[\boldsymbol{\eta}^{[i]}\boldsymbol{\eta}^{[i]T}] \tag{18.156c}$$

the model error is unbiased and uncorrelated to that of the analysis, with an error covariance matrix of Q. As said before, the analysis step remains identical, and (18.144), (18.147), (18.148) will be the same except the matrices are now differently denoted:

$$\mathbf{x}_a^{[i]} = \mathbf{x}_f^{[i]} + K^{[i]}(\mathbf{y}_o^{[i]} - H\mathbf{x}_f^{[i]}) \tag{18.157a}$$

$$K^{[i]} = P_f^{[i]}H^T(HP_f^{[i]}H^T + R^{[i]})^{-1} \tag{18.157b}$$

$$P_a^{[i]} = (I - K^{[i]}H)P_f^{[i]} \tag{18.157c}$$

Meanwhile, the forecast equations are (18.155a) for the model state, and for the error covariance, we have

$$\begin{aligned}
\boldsymbol{\epsilon}_f^{[i+1]} &= \mathbf{x}_f^{[i+1]} - \mathbf{x}_t^{[i+1]} \\
&= M^{[i]}\mathbf{x}_a^{[i]} - (M^{[i]}\mathbf{x}_t^{[i]} - \boldsymbol{\eta}^{[i]}) \qquad \text{(by (18.155))} \\
&= M^{[i]}(\mathbf{x}_a^{[i]} - \mathbf{x}_t^{[i]}) + \boldsymbol{\eta}^{[i]} \\
&= M^{[i]}\boldsymbol{\epsilon}_a^{[i]} + \boldsymbol{\eta}^{[i]}
\end{aligned} \tag{18.158}$$

and thus

$$P_f^{[i+1]} = E[\boldsymbol{\epsilon}_f^{[i+1]}\boldsymbol{\epsilon}_f^{[i+1]T}]$$

$$
\begin{aligned}
&= E[(M^{[i]}\boldsymbol{\epsilon}_a^{[i]} + \boldsymbol{\eta}^{[i]})(M^{[i]}\boldsymbol{\epsilon}_a^{[i]} + \boldsymbol{\eta}^{[i]})^T] \\
&= M^{[i]}E[\boldsymbol{\epsilon}_a^{[i]}\boldsymbol{\epsilon}_a^{[i]T}]M^{[i]T} + E[\boldsymbol{\eta}^{[i]}\boldsymbol{\eta}^{[i]T}] \qquad \text{(by (18.156b))} \\
&= M^{[i]}P_a^{[i]}M^{[i]T} + Q^{[i]} \qquad (18.159)
\end{aligned}
$$

The forecast step then consists of (18.155a) and (18.159). There are two ways to implement Kalman Filter. The standard scheme is to compute the Kalman gain first in the analysis step while the other scheme is to compute the analysis error covariance first. The flowcharts for these two implementations are shown in Figures 18.8 and 18.9.

Figure 18.8.: *The flowchart for the standard Kalman Filter scheme.*

Figure 18.9.: *The flowchart for the alternative Kalman Filter scheme using equivalent expressions for the analysis error covariance and Kalman gain.*

18.4.4. Python Programming: Demonstration of DA using the Lorenz-63 Model

We will use the Lorenz-63 Model to demonstrate the implementation of Kalman Filter. First, borrow the codes from the last chapter to create a ground truth for the evolution of the Lorenz-63 system.

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.integrate import solve_ivp

def Lorenz63(t, y, sigma=10, beta=8/3, rho=28):
    X, Y, Z = y[0], y[1], y[2]
    dXdt = -sigma*X + sigma*Y
    dYdt = -X*Z + rho*X - Y
    dZdt = X*Y - beta*Z
    return([dXdt, dYdt, dZdt])

t_span = [0,5]
dt = 0.01
x_0 = [-4,10,8]

sol_t = solve_ivp(Lorenz63, t_span, x_0, t_eval=np.arange(
    t_span[0], t_span[1], dt))
x_t = sol_t.y.T
```

From this, we also generate a time series of synthetic observations by adding Gaussian noises with a constant error covariance of $R = 0.5I$.

```
R = np.diag([0.5]*3)
obs_errors = np.random.multivariate_normal([0,0,0], R, x_t.
    shape[0])
x_o = x_t + obs_errors
```

Since the Lorenz-63 system is non-linear, for the algorithm of Kalman Filter to work properly, we have to use the *Tangent Linear Model* in place of $M^{[i]}$ when

computing $P_f^{[i]}$:

$$M^{[i]} = I + (\frac{\partial F}{\partial \mathbf{x}})_a^{[i]} \Delta t \tag{18.160}$$

where $\partial F / \partial \mathbf{x}$ is the Jacobian of the Lorenz-63 system as in (17.38). We will integrate the model forward in time using the simple *forward Euler method* at the forecast step. On the other hand, at the analysis step, we have to compute the Kalman gain and analysis (as well as its error covariance). To initialize, we will make a guess of $P_f^{[0]} = 0.5I$, which is not a problem as the Kalman Filter procedure itself will cycle and tune the error covariance. Finally, we simply have $H = I$ as we will use the synthetic observations of all exact variables. The final program will look like this:

```
sigma, beta, rho = 10, 8/3, 28
P_f_i = np.diag([0.5]*3)
H = np.identity(3)
x_a = np.array([[-5,12,6]]) # Assume some initial error
x_f_i = x_a[0,:]

for ii in np.arange(x_t.shape[0]):

    Ts = ii*dt # Forecast start time
    Ta = (ii+1)*dt # Forecast end time (DA analysis time)

    #--------------
    # Analysis Step
    #--------------

    K_i = P_f_i @ H.T @ np.linalg.inv(H @ P_f_i @ H.T + R) #
        Compute Kalman gain

    d = x_o[ii, :] - H @ x_f_i # Innovation
    x_a_i = x_f_i + K_i @ d # Update analysis
    P_a_i = (np.identity(3) - K_i @ H) @ P_f_i # plus its
        error covariance

    if ii >= 1:
```

```
        x_a = np.vstack([x_a, [x_a_i]]) # Save the analysis
            output

    #--------------
    # Forecast Step
    #--------------

    x_f_i = x_a_i + np.array(Lorenz63(Ts, x_a_i))*dt #
        Forecast by Euler
    Jcb = np.array([[-sigma,        sigma,     0         ],
                    [rho-x_f_i[2], -1,        -x_f_i[0]],
                    [x_f_i[1],     x_f_i[0], -beta     ]]) #
                        Jacobian
    M_i = np.identity(3) + Jcb*dt # Tangent Linear Model
    P_f_i = M_i @ P_a_i @ M_i.T + np.diag([0.1]*3) # Assume Q
        = 0.1I
```

Let's visualize the performance of our DA system by looking at the errors:

```
err = np.sum(np.sqrt((x_a - x_t)**2), axis=1)

plt.plot(np.arange(t_span[0], t_span[1], dt), err, c="r",
    label="Error")
plt.hlines(0.5,0,5,linestyles="--",colors="k")
plt.legend()
plt.xlabel("Time")
plt.ylabel("Magnitude")
plt.title("Kalman Filter")
plt.text(4.5,0.55,"0.5",fontsize=16)

plt.savefig("Kalman Filter", dpi=300, bbox_inches="tight")
```

From Figure 18.10, we can see that the error quickly diminishes at just the beginning of DA and successfully remains at a low level that is slightly above the given observation error.

Figure 18.10.: *The evolution of the root-mean-squared error in the Lorenz-63 Model with a Kalman Filter DA system.*

18.5. Exercise

Exercise 18.1 Prove the following vector calculus identities using the Index Notation.

(a) $\nabla \cdot (\nabla \times \vec{v}) = 0$;

(b) $\nabla \times \nabla u = \mathbf{0}$;

(c) $\nabla \times (\nabla \times \vec{v}) = \nabla(\nabla \cdot \vec{v}) - \nabla^2 \vec{v}$

(d) $\nabla \cdot (\vec{u} \times \vec{v}) = \vec{v} \cdot (\nabla \times \vec{u}) - \vec{u} \cdot (\nabla \times \vec{v})$.

Exercise 18.2 Prove that for an $n \times n$ square matrix A, if its characteristic polynomial is

$$p(t) = (-1)^n (t^n + c_{n-1} t^{n-1} + \cdots + c_1 t + c_0)$$

then we have

$$\mathrm{tr}(A^n) + c_{n-1}\mathrm{tr}(A^{n-1}) + \cdots + c_1 \mathrm{tr}(A) = -nc_0$$

Hint: Use the Cayley-Hamilton Theorem with the subscripts for the powers of A written out, then contract the equation with a Kronecker delta.

Exercise 18.3 Confirm that if $\vec{u}$ and $\vec{v}$ are two vectors in the tensor sense, then the cross product

$$w_i = \epsilon_{ijk} u_j v_k$$

as in (18.12) must also be the components of a vector via Quotient Law.

Exercise 18.4 Given a two-dimensional plane stress with two measurements of normal/shear tractions $(7, 4)$ and $(9, -3)$ that are assumed to be accurate, find the corresponding Mohr's Circle and the angle between the two oriented surfaces on which the measurements are taken.

Exercise 18.5 A three-dimensional stress tensor is known to have principal stresses of $\sigma_1, \sigma_2, \sigma_3 = 3, 2, -1$. Determine if the following normal/shear traction pairs are possible:

(a) $(0, 1)$;

(b) $(1.5, 1.25)$;

(c) $(4, 1)$;

(d) $(2.4, -0.6)$.

Exercise 18.6 Derive the continuity equation for any extensive quantity, let's say specific humidity q:

$$\frac{\partial(\rho q)}{\partial t} + \nabla \cdot (\rho q \mathbf{v}) = 0$$

if there is no sink or source so that $dq/dt = 0$.

Exercise 18.7 Modify the DA system in Section 18.4.4 so that it only assimilates selected but not all variables (let's say, X and Z), and also try to change the values of R and Q.

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

(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) $\begin{bmatrix} 2 & 3 & 5 & 7 \\ 1 & 2 & 4 & 8 \\ 1 & 4 & 8 & 12 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & 2 & 4 & 8 \\ 5 & 12 & 23 & 37 \\ -1 & -1 & -2 & -6 \end{bmatrix}$

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 given 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 thus $x + y = 35$. Taking into account the total number 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 all diagonal entries $= (1)(6)(10)(13)(15) = 11700$. This is also true for any upper(lower)-triangular matrix.

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)

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

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 3 & 2 & 9 & 1 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 1 \end{array}\right] \qquad R_1 \leftrightarrow R_2$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & -4 & 0 & 1 & -3 & 0 \\ 0 & -8 & -8 & 0 & -4 & 1 \end{array}\right] \qquad \begin{array}{r} R_2 - 3R_1 \rightarrow R_2 \\ R_3 - 4R_1 \rightarrow R_3 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 1 & 1 & 0 & \frac{1}{2} & -\frac{1}{8} \end{array}\right] \qquad \begin{array}{r} -\frac{1}{4}R_2 \rightarrow R_2 \\ -\frac{1}{8}R_3 \rightarrow R_3 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{array}\right] \qquad R_3 - R_2 \rightarrow R_3$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{4} & \frac{1}{4} & \frac{3}{8} \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{array}\right] \qquad R_1 - 3R_3 - 2R_2 \rightarrow R_1$$

(b) $\det(A) = -32$ and

$$\begin{aligned} \operatorname{adj}(A) &= \begin{bmatrix} \begin{vmatrix} 2 & 3 \\ 0 & 4 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 4 & 4 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 4 & 0 \end{vmatrix} \\ -\begin{vmatrix} 2 & 9 \\ 0 & 4 \end{vmatrix} & \begin{vmatrix} 3 & 9 \\ 4 & 4 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ 4 & 0 \end{vmatrix} \\ \begin{vmatrix} 2 & 9 \\ 2 & 3 \end{vmatrix} & -\begin{vmatrix} 3 & 9 \\ 1 & 3 \end{vmatrix} & \begin{vmatrix} 3 & 2 \\ 1 & 2 \end{vmatrix} \end{bmatrix}^T \\ &= \begin{bmatrix} 8 & 8 & -8 \\ -8 & -24 & 8 \\ -12 & 0 & 4 \end{bmatrix}^T \\ &= \begin{bmatrix} 8 & -8 & -12 \\ 8 & -24 & 0 \\ -8 & 8 & 4 \end{bmatrix} \end{aligned}$$

Hence

$$\begin{aligned} A^{-1} &= \frac{1}{\det(A)} \operatorname{adj}(A) \\ &= -\frac{1}{32} \begin{bmatrix} 8 & -8 & -12 \\ 8 & -24 & 0 \\ -8 & 8 & 4 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{3}{8} \\ -\frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{8} \end{bmatrix} \end{aligned}$$

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

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

not equal to the identity.

Exercise 2.6

$$\det(A) = -42$$

$$\det\left(A^{-1}\right) = -\frac{1}{42}$$

$$A^{-1} = \begin{bmatrix} \frac{9}{7} & -\frac{3}{14} & -\frac{2}{7} & -\frac{6}{7} \\ -\frac{1}{21} & -\frac{1}{21} & \frac{1}{21} & \frac{1}{7} \\ -\frac{11}{7} & -\frac{15}{14} & \frac{11}{7} & \frac{5}{7} \\ \frac{3}{7} & \frac{3}{7} & -\frac{3}{7} & -\frac{2}{7} \end{bmatrix}$$

Exercise 2.7 By cofactor expansion along the first column, we can obtain the determinant of A as

$$\begin{aligned}|A| &= p[(2)(0) - (-2)(p)] + 4[(1)(p) - (2)(2)] \\ &= 2p^2 + 4p - 16\end{aligned}$$

which has two roots, $p = -4$ and $p = 2$, such that $|A| = 0$ and A becomes singular. All other values of p 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

$$\begin{aligned}A^{-1} &= \frac{1}{\det(A)}\operatorname{adj}(A) \\ \det\left(A^{-1}\right) &= \det\left(\frac{1}{\det(A)}\operatorname{adj}(A)\right) && \text{(Notice that } \frac{1}{\det(A)} \text{ is now a scalar)} \\ \frac{1}{\det(A)} &= \left(\frac{1}{\det(A)}\right)^n \det(\operatorname{adj}(A)) && \text{(Properties 2.3.9)} \\ \det(\operatorname{adj}(A)) &= (\det(A))^{n-1}\end{aligned}$$

Exercise 3.1

$$A^{-1} = \begin{bmatrix} \frac{1}{21} & \frac{5}{21} & \frac{2}{21} \\ \frac{11}{42} & -\frac{29}{42} & \frac{1}{42} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \end{bmatrix}$$

$$\vec{x} = A^{-1}\vec{h} = \begin{bmatrix} \frac{1}{21} & \frac{5}{21} & \frac{2}{21} \\ \frac{11}{42} & -\frac{29}{42} & \frac{1}{42} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \end{bmatrix} \begin{bmatrix} 6 \\ \frac{7}{2} \\ -\frac{13}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{3}{2} \end{bmatrix}$$

or

$$\left[\begin{array}{ccc|c} 5 & 1 & 3 & 6 \\ 2 & -1 & 1 & \frac{7}{2} \\ 3 & 2 & -4 & -\frac{13}{2} \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 2 & -1 & 1 & \frac{7}{2} \\ 3 & 2 & -4 & -\frac{13}{2} \end{array}\right] \qquad \frac{1}{5}R_1 \rightarrow R_1$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & -\frac{7}{5} & -\frac{1}{5} & \frac{11}{10} \\ 0 & \frac{7}{5} & -\frac{29}{5} & -\frac{101}{10} \end{array}\right] \qquad \begin{array}{r} R_2 - 2R_1 \rightarrow R_2 \\ R_3 - 3R_1 \rightarrow R_3 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & -\frac{7}{5} & -\frac{1}{5} & \frac{11}{10} \\ 0 & 0 & -6 & -9 \end{array}\right] \qquad R_3 + R_2 \rightarrow R_3$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & \frac{1}{5} & \frac{3}{5} & \frac{6}{5} \\ 0 & 1 & \frac{1}{7} & -\frac{11}{14} \\ 0 & 0 & 1 & \frac{3}{2} \end{array}\right] \qquad -\frac{5}{7}R_2 \rightarrow R_2, -\frac{1}{6}R_3 \rightarrow R_3$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & \frac{1}{5} & 0 & \frac{3}{10} \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & \frac{3}{2} \end{array}\right] \qquad R_1 - \frac{3}{5}R_3 \rightarrow R_1, R_2 - \frac{1}{7}R_3 \rightarrow R_2$$

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

Exercise 3.2

$$A^{-1} = \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix}$$

$$\begin{aligned} \vec{x}^{(1)} &= A^{-1}\vec{h}^{(1)} \\ &= \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix} \begin{bmatrix} -1 \\ 5 \\ 1 \end{bmatrix} \\ \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \vec{x}^{(2)} &= A^{-1}\vec{h}^{(2)} \\ &= \begin{bmatrix} -\frac{1}{8} & 0 & \frac{7}{8} \\ \frac{3}{16} & -\frac{1}{2} & -\frac{5}{16} \\ \frac{1}{16} & \frac{1}{2} & -\frac{7}{16} \end{bmatrix} \begin{bmatrix} \frac{19}{4} \\ 1 \\ \frac{5}{4} \end{bmatrix} \\ \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{4} \end{bmatrix} \end{aligned}$$

Exercise 3.3

$$\left[\begin{array}{ccc|c} 3 & 0 & 4 & 2 \\ 1 & 1 & 2 & -1 \\ 1 & -2 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 1 & 1 & 2 & -1 \\ 3 & 0 & 4 & 2 \end{array}\right] \qquad R_1 \leftrightarrow R_3$$

$$\to \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 3 & 2 & -1 \\ 0 & 6 & 4 & 2 \end{array}\right] \qquad \begin{aligned} R_2 - R_1 &\to R_2 \\ R_3 - 3R_1 &\to R_3 \end{aligned}$$

$$\to \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 6 & 4 & 2 \end{array}\right] \qquad \frac{1}{3}R_2 \to R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 0 & 0 & 4 \end{array}\right] \qquad R_3 - 6R_2 \to R_3$$

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

$$\left[\begin{array}{cccc|c} 1 & 1 & -1 & -3 & 2 \\ 1 & 0 & 0 & -1 & 5 \\ 3 & 2 & -2 & -7 & 9 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cccc|c} 1 & 0 & 0 & -1 & 5 \\ 1 & 1 & -1 & -3 & 2 \\ 3 & 2 & -2 & -7 & 9 \end{array}\right] \qquad R_1 \leftrightarrow R_2$$

$$\rightarrow \left[\begin{array}{cccc|c} 1 & 0 & 0 & -1 & 5 \\ 0 & 1 & -1 & -2 & -3 \\ 0 & 2 & -2 & -4 & -6 \end{array}\right] \qquad \begin{array}{r} R_2 - R_1 \rightarrow R_2 \\ R_3 - 3R_1 \rightarrow R_3 \end{array}$$

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

Let $p = s$ and $q = t$ be 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

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

The system will have no solution or infinitely many of them only when the determinant is equal to zero, which gives us three possible values of $\alpha = -1, 0, 1$. When $\alpha = -1$, the system is

$$\left[\begin{array}{ccc|c} 1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 1 & -1 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -2 \end{array}\right] \qquad R_3 + R_1 \to R_3$$

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

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

Finally, the last case, $\alpha = 1$, will have 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.

$$\begin{cases} -4I_1 + 6I_2 = 6 \\ -6I_2 + 9I_3 = -12 \\ I_1 + I_2 + I_3 = 0 \end{cases}$$

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

$$\begin{aligned}
\begin{vmatrix}
\omega & -kH & -lH\\
-kg & \omega & 0\\
-lg & 0 & \omega
\end{vmatrix} &= 0\\
\omega^3 - gHk^2\omega - gHl^2\omega &= 0\\
\omega^2 - gH(k^2+l^2) &= 0
\end{aligned}$$

as the dispersion relation of gravity wave.

Exercise 3.8 $T_{cd} \approx 15.9\,^\circ\mathrm{C}$, $z_{cd} \approx 0.97\,\mathrm{km}$.

Exercise 3.9 $x = 23$, $y = 12$. For the extra part, the new system of equations becomes (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

$$\begin{aligned} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 48 \\ 2 & 4 & 3 & 122 \end{array}\right] &\to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 48 \\ 0 & 2 & 1 & 26 \end{array}\right] && R_2 - 2R_1 \to R_2 \\ &\to \left[\begin{array}{ccc|c} 1 & 1 & 1 & 48 \\ 0 & 1 & \frac{1}{2} & 13 \end{array}\right] && \frac{1}{2}R_2 \to R_2 \\ &\to \left[\begin{array}{ccc|c} 1 & 0 & \frac{1}{2} & 35 \\ 0 & 1 & \frac{1}{2} & 13 \end{array}\right] && R_1 - R_2 \to R_1 \end{aligned}$$

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 number 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 $0, 1, \ldots, 13$ (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) Same as above

(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)

$$\vec{u} \times \vec{v} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 7 & 4 & 1 \\ 8 & 1 & 1 \end{vmatrix} = 3\hat{\imath} + \hat{\jmath} - 25\hat{k} = (3, 1, -25)^T$$

$$\vec{v} \times \vec{u} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 8 & 1 & 1 \\ 7 & 4 & 1 \end{vmatrix} = -3\hat{\imath} - \hat{\jmath} + 25\hat{k} = (-3, -1, 25)^T = -\vec{u} \times \vec{v}$$

(b)

$$A\vec{v} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 8 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 2 \\ 1 \end{bmatrix}$$

$$\begin{aligned} \vec{u} \cdot (\vec{Av}) &= (7, 4, 1)^T \cdot (10, 2, 1)^T \\ &= (7)(10) + (4)(2) + (1)(1) \\ &= 79 \end{aligned}$$

$$A^T\vec{u} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 11 \\ 12 \end{bmatrix}$$

$$\begin{aligned}(A^T\vec{u})\cdot\vec{v} &= (7,11,12)^T\cdot(8,1,1)^T\\ &= (7)(8)+(11)(1)+(12)(1)\\ &= 79\end{aligned}$$

(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 α and β.

Exercise 4.3

(a)

$$\begin{aligned}\|\vec{u}\| &= \sqrt{1^2+(-3)^2+9^2} = \sqrt{91}\\ \hat{u} &= \left(\frac{1}{\sqrt{91}}, -\frac{3}{\sqrt{91}}, \frac{9}{\sqrt{91}}\right)^T\\ \|\vec{v}\| &= \sqrt{1^2+(-2)^2+4^2} = \sqrt{21}\\ \hat{v} &= \left(\frac{1}{\sqrt{21}}, -\frac{2}{\sqrt{21}}, \frac{4}{\sqrt{21}}\right)^T\end{aligned}$$

(b)

$$\begin{aligned}\vec{u}\cdot\vec{v} &= (1)(1)+(-3)(-2)+(9)(4) = 43\\ \cos\theta &= \frac{43}{\sqrt{21}\sqrt{91}} \approx 0.9836\\ \theta &\approx 0.181\,\text{rad}\end{aligned}$$

(c) $\vec{u} \times \vec{v} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 1 & -3 & 9 \\ 1 & -2 & 4 \end{vmatrix} = 6\hat{\imath} + 5\hat{\jmath} + \hat{k} = (6, 5, 1)^T$

(d) $\vec{u} \cdot (\vec{u} \times \vec{v}) = (1, -3, 9)^T \cdot (6, 5, 1)^T = (1)(6) + (-3)(5) + (9)(1) = 0,$
$\vec{v} \cdot (\vec{u} \times \vec{v}) = (1, -2, 4)^T \cdot (6, 5, 1)^T = (1)(6) + (-2)(5) + (4)(1) = 0$

Exercise 4.4

Typhoon Name	Time	Speed	Direction	Vector Form
Nuri	2008/08/22, 08:00	$13\,\mathrm{km\,h^{-1}}$	315°	$(-9.192, 9.192)$
Vicente	2012/07/24, 02:00	$18\,\mathrm{km\,h^{-1}}$	299°	$(-15.743, 8.727)$
Linfa	2015/07/09, 23:00	$15\,\mathrm{km\,h^{-1}}$	245°	$(-13.595, -6.339)$
Mangkhut	2018/09/16, 22:00	$25\,\mathrm{km\,h^{-1}}$	288°	$(-23.776, 7.725)$

Exercise 4.5

$$\begin{aligned} \|\vec{u} + \vec{v}\|^2 &= (\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) \\ &= \|\vec{u}\|^2 + 2(\vec{u} \cdot \vec{v}) + \|\vec{v}\|^2 \\ &\leq \|\vec{u}\|^2 + 2\|\vec{u}\|\|\vec{v}\| + \|\vec{v}\|^2 && \text{(Theorem 4.2.6)} \\ &= (\|\vec{u}\| + \|\vec{v}\|)^2 \end{aligned}$$

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

$$\overrightarrow{F_{\text{cor}}} = (2\Omega(v\sin\varphi - w\cos\varphi))\hat{\imath} + (-2\Omega u\sin\varphi)\hat{\jmath} + (2\Omega u\cos\varphi)\hat{k}$$

and hence the rate of work done is

$$\begin{aligned}
&\overrightarrow{F_{\text{cor}}} \cdot \vec{v} \\
&= [(2\Omega(v\sin\varphi - w\cos\varphi))\hat{\imath} + (-2\Omega u\sin\varphi)\hat{\jmath} + (2\Omega u\cos\varphi)\hat{k}] \\
&\quad \cdot (u\hat{\imath} + v\hat{\jmath} + w\hat{k}) \\
&= (2\Omega(v\sin\varphi - w\cos\varphi))u + (-2\Omega u\sin\varphi)v + (2\Omega u\cos\varphi)w \\
&= 2\Omega uv\sin\varphi - 2\Omega uw\sin\varphi - 2\Omega uv\sin\varphi + 2\Omega uw\sin\varphi = 0
\end{aligned}$$

Alternatively, note that $\overrightarrow{F_{\text{cor}}} = -2\overrightarrow{\Omega} \times \vec{v}$ and $(\overrightarrow{\Omega} \times \vec{v}) \cdot \vec{v} = 0$ 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) Normal vector to the line is $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

Equation: $\begin{bmatrix} 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 9 \end{bmatrix} \to x - y = -7$

(b) Normal vector to the plane is $\begin{bmatrix} 7 \\ 4 \\ 1 \end{bmatrix} \times \begin{bmatrix} 8 \\ 0 \\ 5 \end{bmatrix} = \begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix}$.

Equation: $\begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 20 \\ -27 \\ -32 \end{bmatrix} \cdot \begin{bmatrix} 6 \\ 3 \\ 2 \end{bmatrix} \to 20x - 27y - 32z = -25$

Exercise 5.3 Part 1: Choose $(0, 0, 2)^T$ as a reference point on the plane. Projection of the displacement $(3, 2, 9)^T - (0, 0, 2)^T = 3\hat{\imath} + 2\hat{\jmath} + 7\hat{k}$ onto the normal vector $\hat{\imath} + 2\hat{\jmath} + 5\hat{k}$ 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 vector projection of $(3, 2, 9)^T - (0, 1, 2)^T = 3\hat{\imath} + 1\hat{\jmath} + 7\hat{k}$ onto the direction vector $\hat{\jmath} + 2\hat{k}$, which is

$$\left(\frac{(3)(0) + (1)(1) + (7)(2)}{0^2 + 1^2 + 2^2}\right)(\hat{\jmath} + 2\hat{k}) = 3(\hat{\jmath} + 2\hat{k}) = 3\hat{\jmath} + 6\hat{k}$$

The displacement vector between the point and line (which is orthogonal to the line) is then $(3\hat{\imath} + 1\hat{\jmath} + 7\hat{k}) - (3\hat{\jmath} + 6\hat{k}) = 3\hat{\imath} - 2\hat{\jmath} + \hat{k}$ and the required distance is equal to its norm $\sqrt{3^2 + (-2)^2 + 1^2} = \sqrt{14}$.

Exercise 5.4 Using the hints, we have the distance as

$$\begin{aligned}\frac{(\vec{v}-\vec{u})\cdot(\hat{l}\times\hat{m})}{\left\|\hat{l}\times\hat{m}\right\|} &= \frac{[(\vec{b}+\hat{m}t)-(\vec{a}+\hat{l}s)]\cdot(\hat{l}\times\hat{m})}{\left\|\hat{l}\times\hat{m}\right\|} \\ &= \frac{(\vec{b}-\vec{a})\cdot(\hat{l}\times\hat{m})+[\hat{m}\cdot(\hat{l}\times\hat{m})]t-[\hat{l}\cdot(\hat{l}\times\hat{m})]s}{\left\|\hat{l}\times\hat{m}\right\|}\end{aligned}$$

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

$$\frac{(\vec{b}-\vec{a})\cdot(\hat{l}\times\hat{m})}{\left\|\hat{l}\times\hat{m}\right\|}$$

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 value 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 $\hat{m}\times\hat{m} = \mathbf{0}$, and hence

$$\begin{aligned}(\vec{a}+\hat{l}s)\times\hat{m} &= (\vec{b}+\hat{m}s)\times\hat{m} \\ \vec{a}\times\hat{m}+s(\hat{l}\times\hat{m}) &= \vec{b}\times\hat{m}+s(\hat{m}\times\hat{m}) = \vec{b}\times\hat{m}+s\mathbf{0} \\ s(\hat{l}\times\hat{m}) &= (\vec{b}-\vec{a})\times\hat{m}\end{aligned}$$

s is then inferred from the 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) $\vec{u} \times \vec{v} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ 1 & 2 & 3 \\ 2 & 1 & 5 \end{vmatrix} = 7\hat{\imath} + \hat{\jmath} - 3\hat{k}$

Area $= \sqrt{7^2 + 1^2 + (-3)^2} = \sqrt{59}$

(b) Volume is the absolute value of $|\vec{u} \times \vec{v}| \cdot \vec{w}$: $\left|(7\hat{\imath} + \hat{\jmath} - 3\hat{k}) \cdot (\hat{\imath} + 4\hat{\jmath})\right| = |(7)(1) + (1)(4) + (-3)(0)| = 11$

(c) Volume $= \text{abs}\left(\begin{vmatrix} 1 & 2 & 3 \\ 2 & 1 & 5 \\ 1 & 5 & 4 \end{vmatrix}\right) = 0$, 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 $-\frac{1}{3}\hat{\imath} - \frac{5}{3}\hat{\jmath} + \hat{k}$ and passing through the point $(\frac{2}{3}, -\frac{5}{3}, 0)^T$.

Exercise 6.1

(a) It is just equivalent to solving

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 9 \end{bmatrix}$$

and the solution is $(a, b, c)^T = (-2, 4, 5)^T$.

(b) Similarly, $a = \frac{5}{2}$, $b = -\frac{3}{2}$, $c = \frac{13}{2}$.

Exercise 6.2

(a) Clearly $\vec{v} = -2\vec{u}$ so they are linearly dependent.

(b) The determinant of $[\vec{u}|\vec{v}|\vec{w}]$ is

$$\begin{vmatrix} 1 & 6 & 4 \\ 2 & 7 & 8 \\ 3 & 9 & 5 \end{vmatrix} = 35 \neq 0$$

so they are linearly independent by the equivalent statements.

(c) Meanwhile

$$\begin{vmatrix} 1 & 3 & 1 \\ 3 & 2 & -4 \\ 3 & 9 & 3 \end{vmatrix} = 0$$

so they are linearly dependent by the same logic.

Exercise 6.3 We can first derive the CR Factorization over the four $\vec{v}^{(j)}$ to obtain a basis:

$$\begin{bmatrix} 1 & 1 & -1 & 3 \\ 3 & -1 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 1 & -1 & 2 & -2 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & -1 & 3 \\ 0 & -4 & 5 & -9 \\ 0 & 2 & 1 & 1 \\ 0 & -2 & 3 & -5 \end{bmatrix} \qquad \begin{aligned} R_2 - 3R_1 &\to R_2 \\ R_4 - R_1 &\to R_4 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 1 & -1 & 3 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & -4 & 5 & -9 \\ 0 & -2 & 3 & -5 \end{bmatrix} \qquad \begin{aligned} R_2 &\leftrightarrow R_3 \\ \frac{1}{2}R_2 &\rightarrow R_2 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 1 & -1 & 3 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 7 & -7 \\ 0 & 0 & 4 & -4 \end{bmatrix} \qquad \begin{aligned} R_3 + 4R_2 &\rightarrow R_3 \\ R_4 + 2R_2 &\rightarrow R_4 \end{aligned}$$

$$\rightarrow \begin{bmatrix} 1 & 1 & -1 & 3 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{aligned} \frac{1}{7}R_3 &\rightarrow R_3 \\ R_4 - 4R_3 &\rightarrow R_3 \end{aligned}$$

So the first three vectors $\{(1,3,0,1)^T, (1,-1,2,-1)^T, (-1,2,1,2)^T\}$ form a basis of the subspace. For (a), we solve the system

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

and by repeating the same steps of Gaussian Elimination, we will obtain

$$\left[\begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

which is consistent and shows that $(1,4,3,2)^T$ is in the subspace (particularly $(1,4,3,2)^T = (1,3,0,1)^T + (1,-1,2,-1)^T + (-1,2,1,2)^T$ if we do the backward

phase too). Similarly for (b), we consider

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

and the same reduction steps give

$$\left[\begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

which is inconsistent and $(1, 2, -3, 1)^T$ is not in the subspace.

Exercise 6.4

(a) It is to solve

$$P_\beta^S[\vec{v}]_\beta = \begin{bmatrix} 6 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} [v_1]_\beta \\ [v_2]_\beta \\ [v_3]_\beta \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}$$

which yields $[v_1]_\beta = \frac{8}{13}$, $[v_2]_\beta = \frac{5}{13}$, $[v_3]_\beta = \frac{6}{13}$, and thus $[\vec{v}]_\beta = (\frac{8}{13}, \frac{5}{13}, \frac{6}{13})_\beta^T$ as required.

(b) It is simply given by

$$[\vec{v}]_S = P_\beta^S(1, -1, 1)^T = \begin{bmatrix} 6 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 4 \\ 4 \end{bmatrix}$$

Exercise 6.5 We directly apply Gaussian Elimination on

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

$$\to \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & -1 & -2 \\ 0 & 3 & 3 & 0 \\ 0 & -2 & 0 & 2 \end{array}\right] \qquad R_2 \leftrightarrow R_3$$

$$\to \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 6 & 6 \\ 0 & 0 & -2 & -2 \end{array}\right] \qquad \begin{array}{r} R_3 - 3R_2 \to R_3 \\ R_4 + 2R_2 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} \frac{1}{6}R_3 \to R_3 \\ R_4 + 2R_3 \to R_4 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_2 + R_3 \to R_2 \\ R_1 - R_3 \to R_1 \end{array}$$

$$\to \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad R_1 - 2R_2 \to R_1$$

So the given $\beta = \{\vec{w}^{(j)}\}$, $j = 1, 2, 3$ forms a basis and $(1, -1, -2, 3)^T = (2, -1, 1)^T_\beta$.

Exercise 6.6 Denote $\mathcal{W}_\cap = \mathcal{W}_1 \cap \mathcal{W}_2$. For $\vec{w}_\cap^{(1)}, \vec{w}_\cap^{(2)} \in \mathcal{W}_1, \mathcal{W}_2$ both, we have $\vec{w}_\cap^{(1)}, \vec{w}_\cap^{(2)} \in \mathcal{W}_\cap$, and thus $a\vec{w}_\cap^{(1)} + b\vec{w}_\cap^{(2)} \in \mathcal{W}_1, \mathcal{W}_2$ by the definition of subspaces, and $a\vec{w}_\cap^{(1)} + b\vec{w}_\cap^{(2)} \in \mathcal{W}_\cap$ so the intersection is indeed a subspace. However, unions in general are not subspaces: take the union of x/y-axes as two subspaces in $\mathbb{R}^2$ as a counter-example, then $(a, 0)^T \in \mathcal{W}_1$, $(0, b)^T \in \mathcal{W}_2$ but $(a, 0)T + (0, b)^T = (a, b)^T \notin \mathcal{W}_1 \cup \mathcal{W}_2$ as long as $a, b \neq 0$.

Exercise 6.7 We simply note that the three vectors $\{(1, 0, 0, 1)^T, (0, 1, -1, 1)^T, (1, 0, 1, -1)^T\}$ given are linearly independent and can serve as the bases β_1, β_2, and $\beta = \beta_1 \cup \beta_2$ themselves, seen by doing Gaussian Elimination. Moreover, $(2, 0, 1, 0)^T = (1, 0, 0, 1)^T + (1, 0, 1, -1)^T = (1, 0, 1)_\beta^T = (1, 0)_{\beta_1}^T + (1)_{\beta_2}^T$.

Exercise 6.8 We note that the results of Gaussian Elimination on A is

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

and thus the column space $\mathcal{C}(A)$ has a basis of $\{(1, 0, 1, 1)^T, (0, 1, 2, 0)^T, (1, 1, 0, 0)^T\}$ while the null space $\mathcal{N}(A)$ has a basis of $\{(-1, 1, 1, 0)^T\}$. Similarly, for A^T it produces

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

so the row space $\mathcal{C}(A^T)$ has a basis of $\{(1, 0, 1, 1)^T, (0, 1, -1, 1)^T, (1, 2, -1, 0)^T\}$ (the first three row vectors of A), and the left null space $\mathcal{N}(A^T)$ can be seen to possess a basis of $\{(-2, 2, -1, 3)\}^T$. Here we simply check Theorem 6.3.9 and Properties 6.3.13 for row/null space and leave the remaining to the readers:

$$\dim(\mathcal{C}(A^T)) + \dim(\mathcal{N}(A)) = 3 + 1 = 4$$

and

$$
\begin{aligned}
(1,0,1,1)^T \cdot (-1,1,1,0)^T &= 0 \\
(0,1,-1,1)^T \cdot (-1,1,1,0)^T &= 0 \\
(1,2,-1,0)^T \cdot (-1,1,1,0)^T &= 0
\end{aligned}
$$

Exercise 7.1 We compute the transformation on each of the β basis vectors:

$$
\begin{aligned}
&T((1,1,1)^T) \\
&= ((1)+(1)+(1), 2(1)+(1), (1)-(1)-3(1))^T = (3,3,-3)^T \\
&= \frac{1}{2}(1,2,3)^T - \frac{13}{2}(1,-1,0)^T + \frac{9}{2}(2,-1,-1)^T = \left(\frac{1}{2}, -\frac{13}{2}, \frac{9}{2}\right)_\eta^T
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
T((1,1,0)^T) &= (2,3,0)^T = \left(\frac{5}{6}, -\frac{23}{6}, \frac{5}{2}\right)_\eta^T \\
T((1,0,0)^T) &= (1,2,1)^T = \left(\frac{2}{3}, -\frac{5}{3}, 1\right)_\eta^T
\end{aligned}
$$

So

$$
[T]_\beta^\eta = \begin{bmatrix} \frac{1}{2} & \frac{5}{6} & \frac{2}{3} \\ -\frac{13}{2} & -\frac{23}{6} & -\frac{5}{3} \\ \frac{9}{2} & \frac{5}{2} & 1 \end{bmatrix}
$$

We can simply calculate its determinant $\det([T]_\beta^\eta) = 0$ which yields zero, implying that it is not full-rank, and by Properties 7.1.10 it is neither one-to-one, onto, nor bijective.

Exercise 7.2 It is elementary Calculus to compute

$$\begin{aligned} T(\cos x) &= (\cos x)' = -\sin x \\ T(\sin x) &= (\sin x)' = \cos x \\ T(x\cos x) &= (x\cos x)' = -x\sin x + \cos x \\ T(x\sin x) &= (x\sin x)' = x\cos x + \sin x \end{aligned}$$

and hence the required matrix representation with respect to β is

$$[T]_\beta = \begin{bmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

The transformed outputs are clearly linearly independent, so by Theorem 7.1.6, T is injective. Moreover, we can observe that

$$\begin{aligned} T(\sin x) &= \cos x \\ T(-\cos x) &= \sin x \\ T(x\sin x + \cos x) &= x\cos x \\ T(-x\cos x + \sin x) &= x\sin x \end{aligned}$$

so by Theorem 7.1.7, T is subjective. Therefore, it is bijective, and the inverse (matrix) is simply

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

the integral operator.

Exercise 7.3 Again, it is elementary Calculus:

$$\begin{aligned} T(1) &= \int_1^x dx = x - 1 \\ T(x) &= \int_1^x x dx = \frac{x^2}{2} - \frac{1}{2} \\ T(x^2) &= \int_1^x x^2 dx = \frac{x^2}{3} - \frac{1}{3} \end{aligned}$$

So the required matrix representation is

$$[T]_S = \begin{bmatrix} -1 & -\frac{1}{2} & -\frac{1}{3} \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

T is easily seen to be not isomorphic as the dimensions of $\mathcal{P}^2$ and $\mathcal{P}^3$ clearly differ.

Exercise 7.4 $\mathcal{U}$ $[\mathcal{V}]$ is isomorphic to an $m[n]$-dimensional real vector space means that it has m $[n]$ basis vectors and a dimension of m $[n]$ by Properties 7.1.16. Hence by definition, the direct sum $\mathcal{U} \oplus \mathcal{V}$ has $m + n$ basis vectors and a dimension of $m + n$. Using Properties 7.1.16 again in a reverse direction shows that $\mathcal{U} \oplus \mathcal{V}$ must be isomorphic to an $m + n$-dimensional real vector space.

Exercise 7.5 (7.8) with $\beta = \beta'$ implies that

$$\begin{aligned} [T]_\beta &= \left[[\vec{v}_\beta^{(1)}]_\beta | [\vec{v}_\beta^{(2)}]_\beta | \cdots | [\vec{v}_\beta^{(n)}]_\beta \right] \\ &= [e^{(1)} | e^{(2)} | \cdots | e^{(n)}] = I \end{aligned}$$

as each of $[\vec{v}_\beta^{(j)}]_\beta$ simply returns $e^{(j)}$.

Exercise 7.6

$$\begin{aligned} T(1) &= 1 \\ T(x) &= x - 1 \\ T(x^2) &= (x-1)^2 = x^2 - 2x + 1 \end{aligned}$$

so

$$[T]_S = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

If we do a variable substitution $x' = x + 2$, the change of coordinates matrix can be inferred from

$$\begin{aligned} 1 &= 1 \\ x' &= x + 2 \\ x'^2 &= (x+2)^2 = x^2 + 4x + 4 \end{aligned}$$

hence

$$P^{\beta}_{\beta'} = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus by Properties 7.2.2, we have

$$\begin{aligned} [T]_{\beta'} &= (P^{\beta}_{\beta'})^{-1} [T]_{\beta} P^{\beta}_{\beta'} \\ &= \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} = [T]_{\beta} \end{aligned}$$

This means that the transformation in question will retain the same matrix representation because the change of variables is essentially just a uniform translation of the x-axis by two units.

Exercise 7.7 Using Properties 7.2.1 twice:

$$\begin{aligned}
[S \circ T]^{\kappa'}_{\beta'} &= [S]^{\kappa'}_{\eta'}[T]^{\eta'}_{\beta'} \\
&= (R^{\kappa}_{\kappa'})^{-1}[S]^{\kappa}_{\eta}Q^{\eta}_{\eta'}(Q^{\eta}_{\eta'})^{-1}[T]^{\eta}_{\beta}P^{\beta}_{\beta'} \\
&= (R^{\kappa}_{\kappa'})^{-1}[S]^{\kappa}_{\eta}[T]^{\eta}_{\beta}P^{\beta}_{\beta'} = (R^{\kappa}_{\kappa'})^{-1}[S \circ T]^{\kappa}_{\beta}P^{\beta}_{\beta'}
\end{aligned}$$

where $P^{\beta}_{\beta'}$, $Q^{\eta}_{\eta'}$, and $R^{\kappa}_{\kappa'}$ are the coordinate transformation matrices between the unprimed and primed bases.

Exercise 7.8

(a) $\hat{v}^{(1)} = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})^T$, $\hat{v}^{(2)} = (-\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})^T$,

$$QR = \begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \sqrt{5} & \frac{19}{\sqrt{5}} \\ 0 & \frac{2}{\sqrt{5}} \end{bmatrix}$$

(b) $\hat{v}^{(1)} = (\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}})^T$, $\hat{v}^{(2)} = (-\frac{7}{\sqrt{174}}, -\frac{2}{\sqrt{174}}, \frac{11}{\sqrt{174}})^T$,
$\hat{v}^{(3)} = (\frac{4}{\sqrt{29}}, -\frac{3}{\sqrt{29}}, \frac{2}{\sqrt{29}})^T$,

$$QR = \begin{bmatrix} \frac{1}{\sqrt{6}} & -\frac{7}{\sqrt{174}} & \frac{4}{\sqrt{29}} \\ \frac{2}{\sqrt{6}} & -\frac{2}{\sqrt{174}} & -\frac{3}{\sqrt{29}} \\ \frac{1}{\sqrt{6}} & \frac{11}{\sqrt{174}} & \frac{2}{\sqrt{29}} \end{bmatrix} \begin{bmatrix} \sqrt{6} & \frac{13}{\sqrt{6}} & \frac{11}{\sqrt{6}} \\ 0 & \frac{29}{\sqrt{174}} & \frac{37}{\sqrt{174}} \\ 0 & 0 & \frac{12}{\sqrt{29}} \end{bmatrix}$$

(c) $\hat{v}^{(1)} = (\frac{1}{\sqrt{7}}, -\frac{1}{\sqrt{7}}, \frac{2}{\sqrt{7}}, \frac{1}{\sqrt{7}})^T$, $\hat{v}^{(2)} = (\frac{6}{\sqrt{434}}, \frac{15}{\sqrt{434}}, -\frac{2}{\sqrt{434}}, \frac{13}{\sqrt{434}})^T$,
$\hat{v}^{(3)} = (\frac{12}{\sqrt{186}}, -\frac{1}{\sqrt{186}}, -\frac{4}{\sqrt{186}}, -\frac{5}{\sqrt{186}})^T$,

$$QR = \begin{bmatrix} \frac{1}{\sqrt{7}} & \frac{6}{\sqrt{434}} & \frac{12}{\sqrt{186}} \\ -\frac{1}{\sqrt{7}} & \frac{15}{\sqrt{434}} & -\frac{1}{\sqrt{186}} \\ \frac{2}{\sqrt{7}} & -\frac{2}{\sqrt{434}} & -\frac{4}{\sqrt{186}} \\ \frac{1}{\sqrt{7}} & \frac{13}{\sqrt{434}} & -\frac{5}{\sqrt{186}} \end{bmatrix} \begin{bmatrix} \sqrt{7} & \frac{1}{\sqrt{7}} & -\frac{1}{\sqrt{7}} \\ 0 & \frac{62}{\sqrt{434}} & \frac{29}{\sqrt{434}} \\ 0 & 0 & \frac{27}{\sqrt{186}} \end{bmatrix}$$

Exercise 8.1 By De Moivre's Formula:

$$\begin{aligned}\cos(2\theta) + i\sin(2\theta) &= (\cos\theta + i\sin\theta)^2 \\ &= (\cos^2\theta - \sin^2\theta) + 2i\sin\theta\cos\theta\end{aligned}$$

Now simply compare the real and imaginary parts.

Exercise 8.2

(a) $5 + i$, $\sqrt{26}$, 0.1974;

(b) $\frac{7}{17} + \frac{6}{17}i$, $\sqrt{\frac{5}{17}}$, 0.7086;

(c) $\frac{22}{13} - \frac{20}{13}i$, $\frac{2\sqrt{17}}{\sqrt{13}}$, -0.7378. (All arguments are in radian.)

Exercise 8.3

(a) $7 - i$;

(b) $(3 + 2i, 3 - 3i, 3 + i)^T \cdot (1 + 2i, -1 - i, 2 + i)^T = 14 + i$;

(c) $4(2 + i, 1 - 2i, i)^T - \sqrt{11}(-i, 3, 1 - i)^T$
$= (8 + (4 + \sqrt{11})i, (4 - 3\sqrt{11}) - 8i, -\sqrt{11} + (4 + \sqrt{11})i)^T$

Exercise 8.4

$$\begin{aligned}AB^* &= \begin{bmatrix} 1+i & -i & 3 \\ 0 & 2-i & 1 \\ -1 & i & 2 \end{bmatrix} \begin{bmatrix} 1 & i & 0 \\ 2+i & 3-i & 1 \\ -i & 1+i & -2i \end{bmatrix} \\ &= \begin{bmatrix} 2-4i & 1+i & -7i \\ 5-i & 6-4i & 2-3i \\ -2 & 3+4i & -3i \end{bmatrix}\end{aligned}$$

and

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

Exercise 8.5

$$
\begin{aligned}
\det(A) &= i \\
A^{-1} &= \begin{bmatrix} -4 & 10-5i & -12i \\ 3i+3 & -11-3i & -8+9i \\ -3-2i & 10+i & 6-9i \end{bmatrix}
\end{aligned}
$$

Exercise 8.6 Notice that

$$
\begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \begin{bmatrix} A-BD^{-1}C & BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} = \begin{bmatrix} A-BD^{-1}C & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix}
$$

then

$$
\begin{aligned}
&\begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} \\
&= \begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \begin{bmatrix} A-BD^{-1}C & BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} = \begin{bmatrix} A-BD^{-1}C & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix}
\end{aligned}
$$

The determinant formula can easily be derived from this equation. Moreover,

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1}
$$

$$
\begin{aligned}
&= \left(\begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix}^{-1} \begin{bmatrix} A - BD^{-1}C & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -D^{-1}C & D^{-1} \end{bmatrix}^{-1} \right)^{-1} \\
&= \begin{bmatrix} I_p & [\mathbf{0}]_{p\times q} \\ -D^{-1}C & D^{-1} \end{bmatrix} \begin{bmatrix} (A - BD^{-1}C)^{-1} & [\mathbf{0}]_{p\times q} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \\
&= \begin{bmatrix} (A - BD^{-1}C)^{-1} & [\mathbf{0}]_{p\times q} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ [\mathbf{0}]_{q\times p} & I_q \end{bmatrix} \\
&= \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \\
&= \begin{bmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{bmatrix}
\end{aligned}
$$

Exercise 8.7

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

gi

Exercise 8.8 It is simply

$$
\begin{aligned}
[T]_\beta^\eta &= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \oplus \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}
\end{aligned}
$$

Exercise 9.1 By Theorem 9.1.3, the geometric multiplicities of A are all 1 and will be strictly equal to the algebraic multiplicities, and so by Properties 9.2.2, A is diagonalizable.

Exercise 9.2 For A, the characteristic polynomial is

$$p(\lambda) = \begin{vmatrix} 1-\lambda & 3 & 3 \\ 4 & -\lambda & 4 \\ 0 & 1 & 2-\lambda \end{vmatrix} = -\lambda^3 + 3\lambda^2 + 14\lambda - 16$$
$$= (1-\lambda)(-16 - 2\lambda + \lambda^2)$$

and the eigenvalues are $\lambda = 1,\ 1+\sqrt{17},\ 1-\sqrt{17}$. Meanwhile, B is an upper-triangular matrix and its characteristic polynomial is $p(\lambda) = (1-\lambda)(5-\lambda)(8-\lambda)(10-\lambda)$, seen by repeated cofactor expansion along rows starting from the left, and hence its eigenvalues are just the diagonal entries $\lambda = 1, 5, 8, 10$.

Exercise 9.3 The characteristic equation is given by

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

so the eigenvalues are 1, 2 and 4. For $\lambda = 1$, we set out to solve the system $A - I = \mathbf{0}$:

$$\left[\begin{array}{ccc|c} 2 & 1 & 4 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

This can be done by Gaussian elimination, which yields

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

We can let $z = t$ be the free variable, giving $y = \frac{4}{3}t$ and $x = -\frac{8}{3}t$. The eigenvector for $\lambda = 1$ is then $(-\frac{8}{3}, \frac{4}{3}, 1)^T$. Similarly, the eigenvectors for $\lambda = 2, 4$ are $(-1, 1, 0)^T$ and $(1, 1, 0)^T$ respectively. The calculation for its inverse should return the new eigenvalues as the reciprocals of the old ones while the eigenvectors remain the same, as suggested in the main text.

Exercise 9.4 The eigenvalues are $\lambda = -1, 1, 3$ and the corresponding eigenvectors are $(-2, -5, 6)^T$, $(0, 1, 0)^T$, $(2, 3, 2)^T$. After Gram-Schmidt orthogonalization, they become $(-\frac{2}{\sqrt{65}}, -\frac{5}{\sqrt{65}}, \frac{6}{\sqrt{65}})^T$, $(-\frac{1}{\sqrt{26}}, \frac{4}{\sqrt{26}}, \frac{3}{\sqrt{26}})^T$, $(\frac{3}{\sqrt{10}}, 0, \frac{1}{\sqrt{10}})^T$. (The numbers will be different if the ordering is different.)

Exercise 9.5 The eigenvalues and eigenvectors for the matrix A are $\lambda = 1, 2-\sqrt{3}, 2+\sqrt{3}$ and $(0, 0, 1)^T$, $(-1-\sqrt{3}, 1, 0)^T$, $(-1+\sqrt{3}, 1, 0)^T$. On the other hand, the eigenvalues for its transpose A^T should still be the same, but the eigenvectors are now $(0, 0, 1)^T$, $(-1-\sqrt{3}, 2, 0)^T$, $(-1+\sqrt{3}, 2, 0)^T$, where the first eigenvector is the same but the remaining two eigenvectors are different.

Exercise 9.6 Assume the matrix A can be diagonalized into

$$D = P^{-1}AP = \begin{bmatrix} D_{11} & 0 & \cdots & 0 \\ 0 & D_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_{nn} \end{bmatrix}$$

then its characteristic polynomial is clearly

$$\begin{aligned} p(\lambda) = \det(D - \lambda I) &= (D_{11} - \lambda)(D_{22} - \lambda) \cdots (D_{nn} - \lambda) \\ &= \cdots + (-1)^{n-1}(D_{11} + D_{22} + \cdots + D_{nn})\lambda^{n-1} + (-1)^n \lambda^n \\ &= \cdots + (-1)^n(-\mathrm{tr}(D))\lambda^{n-1} + (-1)^n \lambda^n \end{aligned}$$

so $\mathrm{tr}(D) = -c_{n-1}$ and by trace/characteristic polynomial invariance (Properties 9.2.3), we also have $\mathrm{tr}(A) = -c_{n-1}$. The determinant part is straightforward:

just set $\lambda = 0$ on both sides to get

$$\begin{aligned}\det(A - (0)I) &= (-1)^n((0)^n + c_{n-1}(0)^{n-1} + \cdots + c_0)\\ \det(A) &= (-1)^n c_0\end{aligned}$$

Exercise 9.7

(a) The characteristic polynomial is $p(\lambda) = \lambda^3 - 3\lambda^2 + 2\lambda$. So by Cayley-Hamilton Theorem, $A^3 - 3A^2 + 2A = 0I$ and thus

$$\begin{aligned}A^3 &= 3A^2 - 2A\\ &= 3\begin{bmatrix}1 & -3 & 3\\ 1 & 1 & -1\\ 1 & -3 & 3\end{bmatrix} - 2\begin{bmatrix}1 & -1 & 1\\ 1 & 1 & -1\\ 1 & -1 & 1\end{bmatrix}\\ &= \begin{bmatrix}1 & -7 & 7\\ 1 & 1 & -1\\ 1 & -7 & 7\end{bmatrix}\end{aligned}$$

(b) The generating formula is $A^{n+1} = 3A^n - 2A^{n-1}$. By mathematical induction, at A_{12}, A_{13}, A_{32}, A_{33}, substituting the proposed expressions into the equation yields $3(2^n - 1) - 2(2^{n-1} - 1) = 2^{n+1} - 1$ which is balanced.

Exercise 9.8

(a) Diagonalization is not possible, as the repeated eigenvalue $\lambda = 2$ (with an algebraic multiplicity of 2) has only one eigenvector, which is $(1, 1, 0)^T$.

(b) The eigenvectors are $(-1, 1, 0)^T$ and $(-1, 0, 1)^T$ for $\lambda = 1$:

$$\left[\begin{array}{ccc|c}1-1 & 0 & 0 & 0\\ 2 & 3-1 & 2 & 0\\ 0 & 0 & 1-1 & 0\end{array}\right] = \left[\begin{array}{ccc|c}0 & 0 & 0 & 0\\ 2 & 2 & 2 & 0\\ 0 & 0 & 0 & 0\end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

and $(0, 1, 0)^T$ for $\lambda = 3$. Hence

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

(c) The eigenvectors are $(0, -1, 4)^T$, $(4, -3, 2)^T$, $(0, 1, 0)^T$ for $\lambda = 1, 3, 5$, so we have

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

Exercise 9.9 The system is $\mathbf{y}' = A\mathbf{y}$ where

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

The eigenvectors of A are $(2, 1, 1)^T, (1, 0, 1)^T, (1, 1, 1)^T$ for the eigenvalues $\lambda = 1, 2, 3$. By substituting $\mathbf{y} = P\mathbf{z}$ where

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

the system becomes $\mathbf{z}' = (P^{-1}AP)\mathbf{z} = D\mathbf{z}$ with

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

The corresponding solutions are $z_1 = C_1e^x$, $z_2 = C_2e^{2x}$ and $z_3 = C_3e^{3x}$. Computing $\mathbf{y} = P\mathbf{z}$ recovers the equations

$$\begin{cases} y_1 = 2C_1e^x + C_2e^{2x} + C_3e^{3x} \\ y_2 = C_1e^x + C_3e^{3x} \\ y_3 = C_1e^x + C_2e^{2x} + C_3e^{3x} \end{cases}$$

The initial condition implies that

$$\begin{bmatrix} 3 \\ 2 \\ 9 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix}$$

whose solutions are $C_1 = -6$, $C_2 = 7$, $C_3 = 8$. Hence the complete solution is

$$\begin{cases} y_1 = -12e^x + 7e^{2x} + 8e^{3x} \\ y_2 = -6e^x + 8e^{3x} \\ y_3 = -6e^x + 7e^{2x} + 8e^{3x} \end{cases}$$

Exercise 9.10 The coefficient matrix for the system is

$$\begin{bmatrix} -k_{12} & k_{12} & 0 \\ 0 & -k_{23} & k_{23} \\ k_{31} & 0 & -k_{31} \end{bmatrix}$$

and the corresponding characteristic polynomial is

$$\begin{aligned} &\begin{vmatrix} -k_{12} - \lambda & k_{12} & 0 \\ 0 & -k_{23} - \lambda & k_{23} \\ k_{31} & 0 & -k_{31} - \lambda \end{vmatrix} \\ &= (-k_{12} - \lambda)(-k_{23} - \lambda)(-k_{31} - \lambda) + k_{12}k_{23}k_{31} \\ &= -k_{12}k_{23}k_{31} - (k_{12}k_{23} + k_{23}k_{31} + k_{31}k_{12})\lambda \end{aligned}$$

$$
\begin{aligned}
&-(k_{12}+k_{23}+k_{31})\lambda^2-\lambda^3+k_{12}k_{23}k_{31}\\
&=-[(k_{12}k_{23}+k_{23}k_{31}+k_{31}k_{12})\lambda+(k_{12}+k_{23}+k_{31})\lambda^2+\lambda^3]\\
&=-\lambda[(k_{12}k_{23}+k_{23}k_{31}+k_{31}k_{12})+(k_{12}+k_{23}+k_{31})\lambda+\lambda^2]
\end{aligned}
$$

Clearly the first eigenvalue is $\lambda = 0$ and the eigenvector can easily verified to be simply $(1, 1, 1)^T$. The remaining two eigenvalues come as a conjugate pair with a real part of $-\frac{1}{2}(k_{12} + k_{23} + k_{31}) < 0$, so the evolution induced by them will eventually diminish. What remains will be the part with an eigenvalue of $\lambda = 0$ which does not change in time, so after a long time the concentrations of the chemical tracers will tend to equalize: $[P] = [Q] = [R]$.

Exercise 10.1 For a matrix A, it is orthogonal if and only if $A^T A = I$.

(a) Yes, because

$$
\begin{bmatrix}
\frac{1}{\sqrt{2}} & \frac{3}{2\sqrt{6}} & -\frac{1}{2\sqrt{2}}\\
0 & \frac{1}{2} & \frac{\sqrt{3}}{2}\\
\frac{1}{\sqrt{2}} & -\frac{3}{2\sqrt{6}} & \frac{1}{2\sqrt{2}}
\end{bmatrix}
\begin{bmatrix}
\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}}\\
\frac{3}{2\sqrt{6}} & \frac{1}{2} & -\frac{3}{2\sqrt{6}}\\
-\frac{1}{2\sqrt{2}} & \frac{\sqrt{3}}{2} & \frac{1}{2\sqrt{2}}
\end{bmatrix}
=
\begin{bmatrix}
1 & 0 & 0\\
0 & 1 & 0\\
0 & 0 & 1
\end{bmatrix}
= I
$$

Its determinant is 1 so it is a rotation.

(b) Yes, because

$$
\begin{bmatrix}
-\frac{\sqrt{3}}{2} & \frac{1}{4} & \frac{\sqrt{3}}{4}\\
\frac{1}{2} & \frac{\sqrt{3}}{4} & \frac{3}{4}\\
0 & -\frac{\sqrt{3}}{2} & \frac{1}{2}
\end{bmatrix}
\begin{bmatrix}
-\frac{\sqrt{3}}{2} & \frac{1}{2} & 0\\
\frac{1}{4} & \frac{\sqrt{3}}{4} & -\frac{\sqrt{3}}{2}\\
\frac{\sqrt{3}}{4} & \frac{3}{4} & \frac{1}{2}
\end{bmatrix}
=
\begin{bmatrix}
1 & 0 & 0\\
0 & 1 & 0\\
0 & 0 & 1
\end{bmatrix}
= I
$$

Its determinant is -1 so it is a reflection.

(c) No, as

$$
\begin{bmatrix}
\frac{1}{2} & 1 & 1\\
-\frac{1}{2} & -1 & 1\\
\frac{1}{4} & -\frac{1}{2} & 2
\end{bmatrix}
\begin{bmatrix}
\frac{1}{2} & -\frac{1}{2} & \frac{1}{4}\\
1 & -1 & -\frac{1}{2}\\
1 & 1 & 2
\end{bmatrix}
=
\begin{bmatrix}
\frac{9}{4} & -\frac{1}{4} & \frac{13}{8}\\
-\frac{1}{4} & \frac{9}{4} & \frac{19}{8}\\
\frac{13}{8} & \frac{19}{8} & \frac{69}{16}
\end{bmatrix}
\neq I
$$

(d) No, as

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{2}{3} \\ 0 & -\frac{2}{3} & \frac{2}{3} \end{bmatrix} \neq I$$

Exercise 10.2

(a)

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(b)

$$P = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{bmatrix}$$

(c)

$$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & -\frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{\sqrt{3}}{2\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Exercise 10.3 We check that

$$P^T P = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2\theta + \sin^2\theta & \cos\theta\sin\theta - \sin\theta\cos\theta \\ \sin\theta\cos\theta - \cos\theta\sin\theta & \sin^2\theta + \cos^2\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

It is also not hard to compute the determinant, which is $-\cos^2\theta - \sin^2\theta = -1$, and hence it is a reflection. Finally, note that the unit x-direction vector $\hat{e}^{(1)}$ [$\hat{e}^{(2)}$] is transformed into $(\cos(2(\frac{\theta}{2})), \sin(2(\frac{\theta}{2})))^T = (\cos\theta, \sin\theta)^T$ [$(\sin(2(\frac{\theta}{2})), -\cos(2(\frac{\theta}{2})))^T = (\sin\theta, -\cos\theta)^T$] as indicated by the first [second] column so they are reflected about the infinitely long straight line with an angle of $\frac{\theta}{2}$ passing through the origin.

Exercise 10.4

(a) $\lambda_1\lambda_2 = 18$ and from the trace $\lambda_1 + \lambda_2 = 1 + 5 + 5 = 11$. Solving the equations gives $\lambda_1 = 2$ and $\lambda_2 = 9$.

(b) By determinant invariance,

$$\begin{vmatrix} 1 & a & a \\ a & 5 & 3 \\ a & 3 & 5 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 9 \end{vmatrix} = 0$$

Expanding the determinant gives $16 - 4a^2 = 0$, hence $a = -2$ or 2.

(c) If $a = -2$, the eigenvector corresponding to λ_0 is $(4, 1, 1)^T$, while if $a = 2$, it is $(-4, 1, 1)^T$. The other two eigenvectors are $(0, -1, 1)^T$ and $(-1, 2, 2)^T$ for $a = -2$ [$(0, -1, 1)^T$, $(1, 2, 2)^T$ for $a = 2$]. So after normalizing the eigenvectors, the required orthogonally diagonalized forms are

$$\begin{bmatrix} \frac{2\sqrt{2}}{3} & \frac{1}{3\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 1 & -2 & -2 \\ -2 & 5 & 3 \\ -2 & 3 & 5 \end{bmatrix} \begin{bmatrix} \frac{2\sqrt{2}}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{2}{3} \\ \frac{1}{3\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{2}{3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

and

$$\begin{bmatrix} -\frac{2\sqrt{2}}{3} & \frac{1}{3\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix} \begin{bmatrix} -\frac{2\sqrt{2}}{3} & 0 & \frac{1}{3} \\ \frac{1}{3\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{2}{3} \\ \frac{1}{3\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{2}{3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

Exercise 10.5

(a) Not possible since it is not symmetric.

(b) The eigenvectors are $(-1,-1,2)^T$, $(-1,1,0)^T$, $(1,1,1)^T$ for $\lambda = 0$, 2, 3 respectively. After normalization, they become $(-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}})^T$, $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)^T$, and $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})^T$. Orthogonal diagonalization gives

$$\begin{bmatrix} -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

(c) The eigenvectors are $(-1,1,0)^T, (-1,0,1)^T$ for $\lambda = 2$ and $(1,1,1)^T$ for $\lambda = 5$. It is required to carry out Gram-Schmidt orthogonalization on the two eigenvectors for $\lambda = 2$ since they are not guaranteed to be orthogonal to each other, making the second eigenvector become $(-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}})^T$. Subsequently, the orthogonal diagonalization reads

$$\begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Exercise 10.6 Complete an orthonormal basis with the orthogonal complement of span($\vec{u}$), denoted by $\mathcal{U}^{\perp}$. By part (d) of Spectral Theorem 10.3.4, we have

$$\begin{aligned} A &= I - \mathbf{u}\mathbf{u}^T \\ &= (\mathbf{u}\mathbf{u}^T + T_{\perp}) - \mathbf{u}\mathbf{u}^T = T_{\perp} \end{aligned}$$

where $T_{\perp}$ is the orthogonal projector onto the complement $\mathcal{U}^{\perp}$. So the eigenvalues of A will be 1s (where there will be $n-1$ of them) within the orthogonal complement and a single 0 along the $\vec{u}$ direction.

Exercise 10.7 We will find an orthonormal basis for the kernel and range. The kernel is generated by the two normal vectors to the hyperplanes $(1, 2, -2, 4)^T$ and $(3, -1, 1, -2)^T$. We carry out the Gram-Schmidt process on them and any two other linearly independent vectors that represent the range. One of the possible choices is the set $(\frac{1}{5}, \frac{2}{5}, -\frac{2}{5}, \frac{4}{5})^T$, $(\frac{2\sqrt{6}}{5}, -\frac{1}{5\sqrt{6}}, \frac{1}{5\sqrt{6}}, -\frac{\sqrt{2}}{5\sqrt{3}})^T$, $(0, \frac{1}{\sqrt{30}}, \sqrt{\frac{5}{6}}, \sqrt{\frac{2}{15}})^T$, $(0, -\frac{2}{\sqrt{5}}, 0, \frac{1}{\sqrt{5}})^T$. (The final answer should not depend on this.) Then, we can construct the orthogonal projection operator in this basis

$$\begin{aligned} T &= \begin{bmatrix} \frac{1}{5} & \frac{2\sqrt{6}}{5} & 0 & 0 \\ \frac{2}{5} & -\frac{1}{5\sqrt{6}} & \frac{1}{\sqrt{30}} & -\frac{2}{\sqrt{5}} \\ -\frac{2}{5} & \frac{1}{5\sqrt{6}} & \sqrt{\frac{5}{6}} & 0 \\ \frac{4}{5} & -\frac{\sqrt{2}}{5\sqrt{3}} & \sqrt{\frac{2}{15}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{5} & \frac{2\sqrt{6}}{5} & 0 & 0 \\ \frac{2}{5} & -\frac{1}{5\sqrt{6}} & \frac{1}{\sqrt{30}} & -\frac{2}{\sqrt{5}} \\ -\frac{2}{5} & \frac{1}{5\sqrt{6}} & \sqrt{\frac{5}{6}} & 0 \\ \frac{4}{5} & -\frac{\sqrt{2}}{5\sqrt{3}} & \sqrt{\frac{2}{15}} & \frac{1}{\sqrt{5}} \end{bmatrix}^T \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{5}{6} & \frac{1}{6} & -\frac{1}{3} \\ 0 & \frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{aligned}$$

Exercise 10.8 Notice that a projection matrix will have an orthogonally diagonalized form of

$$D = \begin{bmatrix} I_r & [\mathbf{0}]_{r\times(n-r)} \\ [\mathbf{0}]_{(n-r)\times r} & [\mathbf{0}]_{(n-r)\times(n-r)} \end{bmatrix}$$

in the direct sum basis of the range and kernel combined. Clearly the trace and rank of D coincide. By rank/trace invariants for diagonalization, this also holds for A.

Exercise 10.9 We will only show part (e) of the theorem for short. The normalized orthogonal eigenvectors are $(0, 1, 0)^T$, $(-\sqrt{\frac{2}{3}}, 0, \frac{1}{\sqrt{3}})^T$ for the eigenvalue of $\lambda = 2$, and $(\frac{1}{\sqrt{3}}, 0, \sqrt{\frac{2}{3}})^T$ for $\lambda = 1$. Then the theorem statement reads

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

Exercise 10.10 The Hermiticity is rather obvious. For checking, we simply note that the eigenvectors (not yet normalized) are $(i, -1, 1-i)^T$, $(-i(5+\sqrt{17}), -1-\sqrt{17}, 2-2i)^T$, and $(i(\sqrt{17}-5), \sqrt{17}-1, 2-2i)^T$ (which may be differed to your answer by a complex multiplicative factor) for $\lambda = 2, \frac{5-\sqrt{17}}{2}, \frac{5+\sqrt{17}}{2}$. The final result should be just

$$D = \begin{bmatrix}2 & 0 & 0\\0 & \frac{5-\sqrt{17}}{2} & 0\\0 & 0 & \frac{5+\sqrt{17}}{2}\end{bmatrix}$$

The unitary form of P is skipped to save the effort.

Exercise 10.11 The equations of motion are

$$\begin{cases} ma_1 = -kx_1 + k(x_2 - x_1) \\ ma_2 = -k(x_2 - x_1) - kx_2 \end{cases}$$

Organizing gives

$$\begin{cases} x_1'' = -\dfrac{2k}{m}x_1 + \dfrac{k}{m}x_2 \\ x_2'' = \dfrac{k}{m}x_1 - \dfrac{2k}{m}x_2 \end{cases}$$

The corresponding matrix form is

$$\begin{bmatrix} x_1'' \\ x_2'' \end{bmatrix} = \frac{k}{m}\begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Applying the transformation by the orthogonal matrix

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

the matrix form can be diagonalized and reduced to

$$\begin{bmatrix} y_1'' \\ y_2'' \end{bmatrix} = \frac{k}{m}\begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

The solutions for y_1 and y_2 are

$$\begin{cases} y_1 = C_1 \cos\left(\sqrt{\dfrac{k}{m}}t - \theta_1\right) = C_1 \cos(\omega t - \theta_1) \\ y_2 = C_2 \cos\left(\sqrt{\dfrac{3k}{m}}t - \theta_2\right) = C_2 \cos\left(\sqrt{3}\omega t - \theta_2\right) \end{cases}$$

where $\omega = \sqrt{k/m}$ is the *intrinsic frequency*. By a reverse transformation, it is not difficult to retrieve the general solution for the original variables as

$$\begin{cases} x_1 = \dfrac{1}{\sqrt{2}}C_1 \cos(\omega t - \theta_1) - \dfrac{1}{\sqrt{2}}C_2 \cos\left(\sqrt{3}\omega t - \theta_2\right) \\ x_2 = \dfrac{1}{\sqrt{2}}C_1 \cos(\omega t - \theta_1) + \dfrac{1}{\sqrt{2}}C_2 \cos\left(\sqrt{3}\omega t - \theta_2\right) \end{cases}$$

Exercise 11.1

(a) The discriminant is $\Delta = 5^2 - 4(1)(3) = 13 > 0$ so it is a hyperbola. The required degree of rotation is found by $\cot(2\theta) = \frac{1-3}{5} = -\frac{2}{5}$. The eigenvalues of

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

are $\lambda = \frac{4\pm\sqrt{29}}{2}$, so the transformed equation will be $\frac{4-\sqrt{29}}{2}(x')^2 + \frac{4+\sqrt{29}}{2}(y')^2 = 1$.

(b) The discriminant is $\Delta = (-1)^2 - 4(1)(2) = -7 > 0$ so it is an ellipse. The required degree of rotation is found by $\cot(2\theta) = \frac{1-2}{-1} = 1$ ($\theta = \frac{\pi}{8}$). The eigenvalues (where we have to divide the equation by 4 to standardize the coefficient on R.H.S.) of

$$\begin{bmatrix} \frac{1}{4} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{1}{2} \end{bmatrix}$$

are $\lambda = \frac{3\pm\sqrt{2}}{8}$, so the new equation will be $\frac{3-\sqrt{2}}{8}(x')^2 + \frac{3+\sqrt{2}}{8}(y')^2 = 1$.

(c) The discriminant is $\Delta = (2)^2 - 4(1)(1) = 0$ so it will be a degenerate case. In fact, the graphs are a pair of parallel straight lines $y = -x \pm 1$.

Exercise 11.2 It is to compute

$$\begin{bmatrix} \cos(60^\circ) & \sin(60^\circ) \\ -\sin(60^\circ) & \cos(60^\circ) \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(60^\circ) & -\sin(60^\circ) \\ \sin(60^\circ) & \cos(60^\circ) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}$$

so the new hyperbola equation is $\frac{1}{2}(x')^2 + \sqrt{3}x'y' - \frac{1}{2}(y')^2 = 1$. Reflection about the x-axis should leave the graph unchanged due to symmetry.

Exercise 11.3 The corresponding quadratic form is

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

and its eigenvalues are $\lambda \approx 0.3170, 1.1803$ and $\lambda = \frac{3}{2}$. The length of the three main axes will be the square roots of their reciprocals: $\sqrt{1/0.3170} = 1.776$, $\sqrt{1/1.1803} = 0.920$, and $\sqrt{\frac{2}{3}}$. The requirement for a three-dimensional quadratic form to represent an ellipsoid is that the three eigenvalues of that quadratic form are all positive.

Exercise 11.4 If $B\vec{x} = \mathbf{0}$, then $\mathbf{x}^T B\mathbf{x} = 0$, and also

$$\begin{aligned} \mathbf{x}^T B\mathbf{x} &= \mathbf{x}^T [I + (A - A^T)]\mathbf{x} \\ &= \mathbf{x}^T I\mathbf{x} + \mathbf{x}^T (A - A^T)\mathbf{x} \\ &= \mathbf{x}^T \mathbf{x} + 0 = \|\vec{x}\|^2 \end{aligned}$$

where from the second to last line we have used the fact that $A - A^T$ is skew-symmetric and do not contribute to the quadratic form. This implies that $\|\vec{x}\|^2 = 0$ so by positivity $\vec{x} = 0$ has to be the zero vector, and so the only solution to $B\vec{x} = \mathbf{0}$ is the trivial solution and by the equivalence statements $B = I + (A - A^T)$ is invertible.

Exercise 11.5 A positive-definite symmetric matrix has all positive eigenvalues and thus a canonical quadratic form of an identity matrix I according to Sylvester's Law of Inertia. By congruence, it must be $B = P^T IP = P^T P$ for some (invertible) P.

Exercise 11.6 The numerical values of the covariance matrix should be

$$\begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) & \mathrm{Cov}(Y,Z) \\ \mathrm{Cov}(Z,X) & \mathrm{Cov}(Z,Y) & \mathrm{Cov}(Z,Z) \end{bmatrix} = \begin{bmatrix} 15.264 & 12.3984 & 9.6756 \\ 12.3984 & 11.664 & 7.4333 \\ 9.6756 & 7.4333 & 11.3644 \end{bmatrix}$$

and the required variance of W is

$$\begin{bmatrix} 1 & -0.5 & -0.5 \end{bmatrix} \begin{bmatrix} 15.264 & 12.3984 & 9.6756 \\ 12.3984 & 11.664 & 7.4333 \\ 9.6756 & 7.4333 & 11.3644 \end{bmatrix} \begin{bmatrix} 1 \\ -0.5 \\ -0.5 \end{bmatrix} = 2.664$$

Exercise 11.7 The three principal directions should be approximately $(0.665, 0.561, 0.493)^T$, $(0.259, 0.445, -0.857)^T$, and $(-0.700, 0.698, 0.151)^T$, accounting for 85.9%/11.9%/2.2% of the explained variance. For checking, the reconstruction of the first data point should give $(X_1, Y_1, Z_1) = (17.92, 18.43, 21.78)$.

Exercise 12.1 We will only argue for (1), (4), (5), and (6) in Definition 6.1.1. Given $A, B \in \mathcal{M}_{n\times n}(\mathbb{C})$ are square complex matrices of extent n, for (1), we know that the usual matrix addition $A+B$ just gives another matrix (of $\mathcal{M}_{n\times n}(\mathbb{C})$), so it is closed under vector addition. In (4) and (5), the zero vector is simply the zero matrix $[\mathbf{0}]$ and the additive inverse of any $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ is just $-A$. Finally for (6), given $c \in \mathbb{C}$, it is obvious that $cA \in \mathcal{M}_{n\times n}(\mathbb{C})$ as well, so it is closed under scalar multiplication. Meanwhile for the Frobenius inner product, we will justify all axioms in Definition 12.1.1. Note that

$$\langle A, B\rangle = \operatorname{tr}(AB^*) = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{B^T_{ki}}\right) = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{B_{ik}}\right)$$

then, for (1), $\overline{\langle B, A\rangle} = \overline{\sum_{i=1}^{n}\left(\sum_{k=1}^{n} B_{ik}\overline{A_{ik}}\right)} = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{B_{ik}}\right) = \langle A, B\rangle$. For (2), $\langle A + C, B\rangle = \sum_{i=1}^{n}\left(\sum_{k=1}^{n}(A_{ik} + C_{ik})\overline{B_{ik}}\right) = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{B_{ik}}\right) + \sum_{i=1}^{n}\left(\sum_{k=1}^{n} C_{ik}\overline{B_{ik}}\right) = \langle A, B\rangle + \langle C, B\rangle$. For (3), it is simply $\langle aA, B\rangle = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} aA_{ik}\overline{B_{ik}}\right) = a\sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{B_{ik}}\right) = a\langle A, B\rangle$. The argument for (4) is the most important one: $\langle A, A\rangle = \operatorname{tr}(AA^*) = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} A_{ik}\overline{A_{ik}}\right) = \sum_{i=1}^{n}\sum_{k=1}^{n} |A_{ik}|^2$ which will be the sum of modulus of all entries in A. This will always be positive unless all entries are zero, i.e. A is the zero matrix.

Exercise 12.2 The inner product of $\vec{u}$ and $\vec{v}$ is now

$$\mathbf{u}^T B\mathbf{v} = \begin{bmatrix} -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 0$$

so they are orthogonal with respect to this inner product. The corresponding norm $\|\vec{u}\|$ is found from

$$\|\vec{u}\|^2 = \mathbf{u}^T B\mathbf{u} = \begin{bmatrix} -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix} = 2$$

so it is $\|\vec{u}\| = \sqrt{2}$. Similarly, $\|\vec{v}\| = \sqrt{3}$.

Exercise 12.3 For the first part, we check the eigenvalues of B which are $\lambda \approx 0.5188, 2.3111, 4.1701$ all positive. We can apply the Gram-Schmidt process on $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \vec{u}^{(3)}\} = \{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T\}$, which will yield

$$\begin{aligned}
\hat{v}^{(1)} &= \frac{\vec{u}^{(1)}}{\|\vec{u}^{(1)}\|} = \left(\frac{1}{\sqrt{3}}, 0, 0\right)^T \\
\vec{v}^{(2)} &= \vec{u}^{(2)} - \langle \vec{u}^{(2)}, \hat{v}^{(1)} \rangle \hat{v}^{(1)} \\
&= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \left(\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ 1 \\ 0 \end{bmatrix} \\
\hat{v}^{(2)} &= \frac{\vec{v}^{(2)}}{\|\vec{v}^{(2)}\|} = \frac{\vec{v}^{(2)}}{\mathbf{v}^{(2)T} B \mathbf{v}^{(2)}} \\
&= \frac{\sqrt{3}}{2\sqrt{2}} \begin{bmatrix} -\frac{1}{3} \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2\sqrt{6}} \\ \frac{\sqrt{3}}{2\sqrt{2}} \\ 0 \end{bmatrix} \\
\vec{v}^{(3)} &= \vec{u}^{(3)} - \langle \vec{u}^{(3)}, \hat{v}^{(1)} \rangle \hat{v}^{(1)} - \langle \vec{u}^{(3)}, \hat{v}^{(2)} \rangle \hat{v}^{(2)}
\end{aligned}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \left(\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ 0 \end{bmatrix}$$

$$- \left(\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{2\sqrt{6}} \\ \frac{\sqrt{3}}{2\sqrt{2}} \\ 0 \end{bmatrix} \right) \begin{bmatrix} -\frac{1}{2\sqrt{6}} \\ \frac{\sqrt{3}}{2\sqrt{2}} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ -\frac{1}{8} \\ 1 \end{bmatrix}$$

$$\hat{v}^{(3)} = \frac{\vec{v}^{(3)}}{\|\vec{v}^{(3)}\|} = \frac{2\sqrt{2}}{\sqrt{5}} \begin{bmatrix} \frac{3}{8} \\ -\frac{1}{8} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{3}{2\sqrt{10}} \\ -\frac{1}{2\sqrt{10}} \\ \frac{2\sqrt{2}}{\sqrt{5}} \end{bmatrix}$$

(Other answers are acceptable as long as the derived vectors are orthonormal with respect to B)

Exercise 12.4

$$\begin{aligned} \|\vec{u} + \vec{v}\|^2 &= \langle (\vec{u} + \vec{v}), (\vec{u} + \vec{v}) \rangle \\ &= \|\vec{u}\|^2 + 2\langle \vec{u} \cdot \vec{v} \rangle + \|\vec{v}\|^2 \\ &\leq \|\vec{u}\|^2 + 2\|\vec{u}\|\|\vec{v}\| + \|\vec{v}\|^2 && \text{(Theorem 12.1.6)} \\ &= (\|\vec{u}\| + \|\vec{v}\|)^2 \end{aligned}$$

Exercise 12.5 It is simply to note that x^3 is odd but $\cos(2x)$ is even, so the inner product integral

$$\int_{-\pi}^{\pi} x^3 \cos(2x) dx = 0$$

is zero as the integrand is overall odd and the integration limits are symmetric.

Exercise 12.6

(a)

$$
\begin{aligned}
&\int_a^b \mathcal{L}[f(x)]\overline{g(x)}dx \\
&= \int_a^b \frac{d}{dx}\left(x\frac{d}{dx}[f(x)]\right)\overline{g(x)}dx \\
&= \left[x\frac{df(x)}{dx}\overline{g(x)}\right]_a^b - \int_a^b x\frac{df(x)}{dx}\frac{d\overline{g(x)}}{dx}dx \\
&= \left[x\frac{df(x)}{dx}\overline{g(x)}\right]_a^b - [xf(x)\frac{d\overline{g(x)}}{dx}]_a^b + \int_a^b f(x)\frac{d}{dx}\left(x\frac{d\overline{g(x)}}{dx}\right)dx \\
&= \text{Boundary Terms} + \int_a^b f(x)\overline{\frac{d}{dx}\left(x\frac{d}{dx}[g(x)]\right)}dx
\end{aligned}
$$

so $\mathcal{L}^\dagger[g] = \frac{d}{dx}\left(x\frac{d}{dx}[g]\right) = \mathcal{L}[g]$ is self-adjoint.

(b) In a similar vein, we note that the new adjoint is
$\mathcal{L}^\dagger[f] = e^x\frac{d}{dx}\left(x\frac{d}{dx}[e^{-x}f]\right)$.

Exercise 12.7 According to (12.7) (adapted for real matrices), we simply check

$$
\begin{aligned}
T^\dagger &\equiv B^{-1}[T]_\beta^T B \\
&= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \equiv T
\end{aligned}
$$

Exercise 12.8

(a) Note that x^2 is even, so only the cosine components remain. Then we simply compute a_m as in (12.17a) with integration by parts:

$$\begin{aligned}
a_m &= \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\cos(mx)dx \\
&= \left[\frac{1}{m\pi}x^2\sin(mx)\right]_{-\pi}^{\pi} - \frac{2}{m\pi}\int_{-\pi}^{\pi} x\sin(mx)dx \\
&= (0) - \frac{2}{m\pi}\left[-\frac{1}{m}x\cos(mx)\right]_{-\pi}^{\pi} + \frac{2}{m^2\pi}\int_{-\pi}^{\pi}(-\cos(mx))dx \\
&= \frac{2}{m^2\pi}[\pi(-1)^m - (-\pi)(-1)^m] - (0) = \frac{4(-1)^m}{m^2}
\end{aligned}$$

Meanwhile, the constant term is just

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} x^2dx &= \frac{1}{2\pi}\left[\frac{x^3}{3}\right]_{-\pi}^{\pi} \\
&= \frac{1}{2\pi}\left[\frac{\pi^3}{3} - \frac{(-\pi)^3}{3}\right] = \frac{\pi^2}{3}
\end{aligned}$$

Therefore, the required Fourier series is

$$x^2 = \frac{\pi^2}{3} + \sum_{m=1}^{\infty}\frac{4(-1)^m}{m^2}\cos(mx)$$

(b) Similarly, x^3 is odd and we only need to calculate b_n as in (12.17b):

$$\begin{aligned}
b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} x^3\sin(nx)dx \\
&= \left[-\frac{1}{n\pi}x^3\cos(nx)\right]_{-\pi}^{\pi} + \frac{3}{n\pi}\int_{-\pi}^{\pi} x^2\cos(nx)dx \\
&= -\frac{1}{n\pi}\left[\pi^3(-1)^n - (-\pi)^3(-1)^n\right] + \frac{3}{n^2\pi}[x^2\sin(nx)]_{-\pi}^{\pi} \\
&\quad - \frac{6}{n^2\pi}\int_{-\pi}^{\pi} x\sin(nx)dx
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{2\pi^2(-1)^n}{n} + (0) + \frac{6}{n^3\pi}[x\cos(nx)]_{-\pi}^{\pi} - \frac{6}{n^3\pi}\int_{-\pi}^{\pi}\cos(nx)dx \\
&= -\frac{2\pi^2(-1)^n}{n} + \frac{6}{n^3\pi}[\pi(-1)^n - (-\pi)(-1)^n] - (0) \\
&= -\frac{2\pi^2(-1)^n}{n} + \frac{12(-1)^n}{n^3}
\end{aligned}
$$

So the required Fourier series is

$$
x^3 = \sum_{m=1}^{\infty}\left(-\frac{2\pi^2(-1)^n}{n} + \frac{12(-1)^n}{n^3}\right)\sin(nx)
$$

(c) This is a trick question where the given form is already a Fourier series so the answer is just $4 + \sin(3x) + 2\cos(8x)$ itself.

Exercise 12.9 The integrating factor needed is

$$
\begin{aligned}
F(x) &= \exp\left(\int \frac{R(x) - P'(x)}{P(x)}dx\right) \\
&= \exp\left(\int \frac{-2x - (1-x^2)'}{(1-x^2)}dx\right) \\
&= \exp\left(\int \frac{-2x - (-2x)}{1-x^2}dx\right) = \exp\left(\int 0dx\right) = 1
\end{aligned}
$$

So the original equation can already be transformed into the Sturm-Liouville form as

$$
((1-x^2)y')' + \lambda y = 0
$$

The $(1-x^2)$ factor inside the derivative term indicates that the natural interval is $[-1, 1]$ so that the operator $\mathcal{L} = \frac{d}{dx}((1-x^2)\frac{d}{dx})$ is Hermitian. By the same method of series solution as introduced in the chapter, we let

$$
y = a_0 + a_1x + a_2x^2 + \cdots = \sum_{n=0}^{\infty} a_nx^n
$$

$$y' = \sum_{n=1}^{\infty} n a_n x^{n-1}$$
$$y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2}$$

and substitution into the Legendre's Equation gives

$$(1-x^2)\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} - 2x\sum_{n=1}^{\infty} n a_n x^{n-1} + \lambda \sum_{n=0}^{\infty} a_n x^n = 0$$
$$\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} - \sum_{n=2}^{\infty} n(n-1)a_n x^n - 2\sum_{n=1}^{\infty} n a_n x^n + \lambda \sum_{n=0}^{\infty} a_n x^n = 0$$
$$\sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2} x^n - \sum_{n=2}^{\infty} n(n-1)a_n x^n - 2\sum_{n=1}^{\infty} n a_n x^n + \lambda \sum_{n=0}^{\infty} a_n x^n = 0$$

So the recurrence relation is

$$(n+2)(n+1)a_{n+2} - n(n-1)a_n - 2na_n + \lambda a_n = 0$$
$$(n+2)(n+1)a_{n+2} - (n(n+1) - \lambda)a_n = 0$$
$$a_{n+2} = \frac{n(n+1) - \lambda}{(n+2)(n+1)}$$

which implies that there are two polynomials solutions for the odd/even indices respectively when $\lambda = n(n+1)$. We only explicitly show the reasoning to construct the fifth Legendre polynomial: let $\lambda = 4 \cdot 5$ that corresponds to $n = 4$, then

$$a_2 = \frac{0(0+1) - (4)(5)}{(0+2)(0+1)} a_0 = -10a_0$$
$$a_4 = \frac{2(2+1) - (4)(5)}{(2+2)(2+1)} a_0 = -\frac{7}{6} a_2 = \frac{70}{6} a_0$$

So we have

$$a_0 \left(\frac{70}{6} x^4 - 10x^2 + 1 \right)$$

which after rescaling (the sum of coefficients has to be 1), gives the fifth Legendre polynomial at $n = 4$ as

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

The first four Legendre polynomials are given below for reference.

$$\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x \\
P_2(x) &= \frac{1}{2}(3x^2 - 1) \\
P_3(x) &= \frac{1}{2}(5x^3 - 3x)
\end{aligned}$$

To use the Gram-Schmidt method, consider the inner product below, applied on the standard polynomials

$$\langle \varphi_1, \varphi_2 \rangle = \int_{-1}^{1} \varphi_1 \varphi_2 dx$$

where the integration range is the natural interval $[-1, 1]$ and the weighting is just $w = 1$.

Exercise 13.1 If A (and hence $A^\dagger$) is invertible, then

$$\begin{aligned}
\vec{x}_f &= (A^\dagger A)^{-1} A^\dagger \vec{h} \\
&= A^{-1}(A^\dagger)^{-1} A^\dagger \vec{h} = A^{-1}\vec{h}
\end{aligned}$$

Exercise 13.2

$$\vec{x}_f = (A^T A)^{-1} A^T \vec{h}$$

$$
\begin{aligned}
&= \left(\begin{bmatrix} 1 & 2 & -3 \\ 1 & -1 & 3 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 2 & -3 \\ 1 & -1 & 3 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & -3 \\ 1 & -1 & 3 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}^T \begin{bmatrix} 4 \\ 5 \\ 1 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 7 & 4 & -1 \\ 4 & 7 & -9 \\ -1 & -9 & 20 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 2 & 1 \\ 2 & -1 & 1 & 1 \\ -3 & 3 & -1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 1 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} \frac{59}{158} & -\frac{71}{158} & -\frac{29}{158} \\ -\frac{71}{158} & \frac{139}{158} & \frac{59}{158} \\ -\frac{29}{158} & \frac{59}{158} & \frac{33}{158} \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 & 1 \\ 2 & -1 & 1 & 1 \\ -3 & 3 & -1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{225}{158} \\ \frac{147}{158} \\ \frac{109}{158} \end{bmatrix}
\end{aligned}
$$

Exercise 13.3 We have $m = 6$, $\sum X = 80$, $\sum(X^2) = 1102$, $\sum(XY) = 81368.9$, $\sum Y = 6105.9$. So

$$
\begin{aligned}
\alpha_f &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{m(\sum X^2) - (\sum X)^2} \\
&= \frac{(6105.9)(1102) - (80)(81368.9)}{6(1102) - (80)^2} \approx 1033.9 \\
\beta_f &= \frac{m(\sum XY) - (\sum X)(\sum Y)}{m(\sum X^2) - (\sum X)^2} \\
&= \frac{6(81368.9) - (80)(6105.9)}{6(1102) - (80)^2} \approx -1.220
\end{aligned}
$$

Therefore the regression formula is $P = 1033.9 - 1.220T$.

Exercise 13.4

(a) We have $m = 12$, $\sum X = 330$, $\sum(X^2) = 12650$, $\sum(XY) = 121977.0$, $\sum Y = 4235.2$. So

$$\begin{aligned}
\alpha_f &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{m(\sum X^2) - (\sum X)^2} \\
&= \frac{(4235.2)(12650) - (330)(121977.0)}{12(12650) - (330)^2} \approx 310.6 \\
\beta_f &= \frac{m(\sum XY) - (\sum X)(\sum Y)}{m(\sum X^2) - (\sum X)^2} \\
&= \frac{12(121977.0) - (330)(4235.2)}{12(12650) - (330)^2} \approx 1.541
\end{aligned}$$

So the linear fit is $[CO_2] = 310.6 + 1.541t$, RMSE is $\sqrt{\frac{126.5}{12}} \approx 3.25$.

(b) For a quadratic fit $[CO_2] = a + bt + ct^2$, we compute

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & \cdots \\ 0 & 5 & 10 & \cdots \\ 0 & 5^2 & 10^2 & \cdots \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 5 & 5^2 \\ 1 & 10 & 10^2 \\ \vdots & \vdots & \vdots \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 0 & 5 & 10 & \cdots \\ 0 & 5^2 & 10^2 & \cdots \end{bmatrix} \begin{bmatrix} 316.9 \\ 320.0 \\ 325.7 \\ \vdots \end{bmatrix}$$

which yields $a = 316.1$, $b = 0.880$, $c = 0.0120$, $[CO_2] = 316.1 + 0.880t + 0.0121t^2$. RMSE is $\sqrt{\frac{5.2}{12}} \approx 0.66$.

Exercise 13.5

t	6	8	12	24	36	48
$n = \ln N$	4.3281	4.2356	4.0483	3.4965	2.9339	2.3795

The parameters are $m = 6$, $\sum X = 134$, $\sum(X^2) = 4420$, $\sum(XY) = 412.1854$, $\sum Y = 21.4219$, and thus

$$\alpha_f = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{m(\sum X^2) - (\sum X)^2}$$

$$
\begin{aligned}
&= \frac{(21.4219)(4420) - (134)(412.1854)}{6(4420) - (134)^2} \approx 4.607 \\
\beta_f &= \frac{m(\sum XY) - (\sum X)(\sum Y)}{m(\sum X^2) - (\sum X)^2} \\
&= \frac{6(412.1854) - (134)(21.4219)}{6(4420) - (134)^2} \approx -0.0464
\end{aligned}
$$

As a result, $\ln N = 4.607 - 0.0464t$, hence $N = e^{4.607-0.0464t} = 100.18e^{-0.0464t}$, roughly corresponding to a half-life of $\frac{\ln 2}{0.0464} \approx 14.94$ hrs.

Exercise 13.6 $M = 94.96 + 6.24D + 12.25W + 2.22A$, $R^2 = 0.94$.

Exercise 14.1 Its DFT is $(43.9, 7.350 - 0.095i, -1.386 - 6.082i, -2.600 + 0.059i, -3.064+8.990i, 0.5, -3.064-8.990i, -2.600-0.059i, -1.386+6.082i, 7.350 + 0.095i)$. The required power and phase for the third frequency are

$$
\begin{aligned}
|C_3|^2 &= C_3C_{-3} = (-2.600 + 0.059i)(-2.600 - 0.059i) = 6.764 \\
\phi_3 &= \arctan\left(\frac{\mathrm{Im}(C_3)}{\mathrm{Re}(C_3)}\right) = \arctan\left(\frac{0.059}{-2.600}\right) = -0.022\,\mathrm{rad}
\end{aligned}
$$

Exercise 14.4 The desired circular convolution is

$$f \circledast g = (27, 28, 41, 14, 13, 4, 4, 19)$$

and the respective DFTs are

$$
\begin{aligned}
&F[f \circledast g](k) \\
&= (150, 34.5 - 50.4i, -5 + i, -6.5 + 23.6i, \\
&\quad 20, -6.5 - 23.6i, -5 - i, 34.5 + 50.4i) \\
&= (15, -0.59 - 7.24i, 3 + 2i, -3.41 - 1.24i, \\
&\quad -5, -3.41 + 1.24i, 3 - 2i, -0.59 + 7.24i)
\end{aligned}
$$

$$\odot\,(10, 6.54 + 5.29i, -1 + i, -0.54 - 6.71i,$$
$$-4, -0.54 + 6.71i, -1 - i, 6.54 - 5.29i) = \hat{f}(k)\hat{g}(k)$$

Exercise 15.1 $P = \begin{bmatrix} 0.75 & 0.6 & 0.3 \\ 0.1 & 0.25 & 0.2 \\ 0.15 & 0.15 & 0.5 \end{bmatrix}$

(a)

$$\vec{q} = P^3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0.639375 & 0.636 & 0.57675 \\ 0.13975 & 0.143125 & 0.1595 \\ 0.220875 & 0.220875 & 0.26375 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.57675 \\ 0.1595 \\ 0.26375 \end{bmatrix}$$

So the required probability is 26.375%.

(b) $I - P = \begin{bmatrix} 0.25 & -0.6 & -0.3 \\ -0.1 & 0.75 & -0.2 \\ -0.15 & -0.15 & 0.5 \end{bmatrix}$,
and the eigenvector (normalized to the steady-state vector) for $\lambda = 1$ can be found to be $(0.6244, 0.1448, 0.2308)^T$.

(c)

The required probability
$= P(\text{today is sunny}) + P(\text{today is not sunny and tomorrow is sunny})$
$= 0.75 + (0.1 \times 0.6 + 0.15 \times 0.3)$
$= 0.855$

Exercise 15.3 Node 1/2/3/4: Biosphere/Atmosphere/Ocean/Sediment. $k_{21} = 0.03$, $k_{12} = 0.1$, $k_{32} = 0.1$, $k_{23} = 0.0015$, $k_{43} = 0.000005$.

(a)

$$P = \begin{bmatrix} 0.97 & 0.1 & 0 & 0 \\ 0.03 & 0.8 & 0.0015 & 0 \\ 0 & 0.1 & 0.998495 & 0 \\ 0 & 0 & 0.000005 & 1 \end{bmatrix}$$

$$P_0 = \begin{bmatrix} 0.97 & 0.1 & 0 \\ 0.03 & 0.8 & 0.0015 \\ 0 & 0.1 & 0.998495 \end{bmatrix}$$

$$I - P_0 = \begin{bmatrix} 0.03 & -0.1 & 0 \\ -0.03 & 0.2 & -0.0015 \\ 0 & -0.1 & 0.001505 \end{bmatrix}$$

$$(I - P_0)^{-1} = \begin{bmatrix} 10066.66 & 10033.33 & 10000 \\ 3010 & 3010 & 3000 \\ 200000 & 200000 & 200000 \end{bmatrix}$$

The sum of the third column is 213000 yrs.

(b) $\tau = \frac{40000+600+2000}{0.2} = 213000$ yrs.

Exercise 15.4 Index 1/2/3/4/5/6/7/8: Start/A/B/C/D/E/F/End.

$$P_0 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}$$

$$I - P_0 = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & -\frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 & -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 & -\frac{1}{3} & -\frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{2} & -\frac{1}{3} & 1 & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 \end{bmatrix}$$

The first column of the inverse $(I - P_0)^{-1}$ can be seen to be $(4, 3, 3, 3, 3, \frac{3}{2}, \frac{3}{2})^T$ whose sum is 19, which is the required number of moves.

Exercise 16.1 We only need to prove this for two square upper-triangular matrices A and B, and apply this recursively. We have $A_{ik} = 0$ if $i > k$ and $B_{kj} = 0$ if $k > j$ and need to show that $(AB)_{ij} = 0$ if $i > j$:

$$\begin{aligned} (AB)_{ij} &= \sum_{k=1}^{n} A_{ik} B_{kj} \\ &= \sum_{k=1}^{i-1} A_{ik} B_{kj} + \sum_{k=i}^{n} A_{ik} B_{kj} \\ &= \sum_{k=1}^{i-1} (0) B_{kj} + \sum_{k=i}^{n} A_{ik} (0) = 0 \end{aligned}$$

The lower-triangular case is similar.

Exercise 16.2

(a)

$$U^T U = \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} \\ 0 & 0 & \sqrt{3} \end{bmatrix}^T \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} \\ 0 & 0 & \sqrt{3} \end{bmatrix}$$

(b)

$$LDU = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Exercise 16.3 One possible LU Decomposition of A is

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

We then solve $U\vec{x} = \vec{y}$ and $L\vec{y} = \vec{h}$. For $\vec{h} = (1, -3, 4)^T$, $L\vec{y} = \vec{h}$, forward substitution gives $\vec{y} = (1, 5, 2)^T$. Then for $U\vec{x} = \vec{y}$, backward substitution yields the answer $\vec{x} = (1, -1, 2)^T$. Similarly, the answer for (b) can be seen to be $\vec{x} = (\frac{5}{3}, 0, \frac{1}{3})^T$.

Exercise 16.4 We note that

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

the non-zero singular values (the square roots of eigenvalues for $A^T A$) are $\sigma_1, \sigma_2, \sigma_3 = \sqrt{5}, \sqrt{3}, \sqrt{2}$, and the respective right singular vectors (the normalized eigenvectors for $A^T A$) are

$$\begin{bmatrix} \frac{2\sqrt{2}}{\sqrt{15}} \\ -\sqrt{\frac{2}{15}} \\ \sqrt{\frac{3}{10}} \\ \frac{1}{\sqrt{30}} \end{bmatrix}, \begin{bmatrix} 0 \\ \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ 0 \\ \sqrt{\frac{2}{3}} \end{bmatrix}$$

The corresponding left singular vectors are then

$$\begin{bmatrix} \frac{1}{\sqrt{6}} \\ \sqrt{\frac{2}{3}} \\ -\frac{1}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$$

So its (compact) SVD is

$$\begin{aligned} U\Sigma V^* &= \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \sqrt{\frac{2}{3}} & 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{2\sqrt{2}}{\sqrt{15}} & -\sqrt{\frac{2}{15}} & \sqrt{\frac{3}{10}} & \frac{1}{\sqrt{30}} \\ 0 & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 & \sqrt{\frac{2}{3}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & 0 & 1 \end{bmatrix} \end{aligned}$$

The SVD of A^T here is simply $(U\Sigma V^T)^T = V\Sigma^T U^T = V\Sigma U^T$.

Exercise 16.5 For the first part, the required pseudoinverse is

$$\begin{bmatrix} -\frac{1}{30} & \frac{11}{45} \\ \frac{4}{15} & -\frac{13}{45} \\ \frac{1}{6} & \frac{1}{9} \end{bmatrix}$$

and the minimal solution is then

$$\begin{bmatrix} -\frac{1}{30} & \frac{11}{45} \\ \frac{4}{15} & -\frac{13}{45} \\ \frac{1}{6} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} \\ \frac{1}{5} \\ 1 \end{bmatrix}$$

For the second part, the appropriate new adjoint is

$$A^\dagger = C^{-1} A^* B$$

$$
\begin{aligned}
&= \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 & 3 \\ 2 & -1 & 2 \end{bmatrix}^T I \\
&= \begin{bmatrix} -\frac{1}{2} & \frac{5}{4} \\ \frac{1}{2} & -\frac{7}{4} \\ \frac{3}{2} & \frac{5}{4} \end{bmatrix}
\end{aligned}
$$

Then

$$
A^\dagger A = \begin{bmatrix} -\frac{1}{2} & \frac{5}{4} \\ \frac{1}{2} & -\frac{7}{4} \\ \frac{3}{2} & \frac{5}{4} \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & -\frac{9}{4} & 1 \\ -3 & \frac{11}{4} & -2 \\ 4 & \frac{7}{4} & 7 \end{bmatrix}
$$

The two non-zero singular values are $\sigma_{1,2} = \sqrt{\frac{47 \pm \sqrt{193}}{8}} \approx 2.759, 2.034$. The right singular vectors relative to C will be, numerically

$$
\begin{bmatrix} 0.3026 \\ -0.4598 \\ 0.6637 \end{bmatrix}, \begin{bmatrix} -0.5192 \\ 0.6415 \\ 0.3333 \end{bmatrix}
$$

The left singular vectors will then be

$$
\begin{bmatrix} 0.498 \\ 0.867 \end{bmatrix}, \begin{bmatrix} 0.867 \\ -0.498 \end{bmatrix}
$$

So the new pseudoinverse will be

$$
\begin{aligned}
A^+ &= V\Sigma^+ U^* B \\
&= \begin{bmatrix} 0.3026 & -0.5192 \\ -0.4598 & 0.6415 \\ 0.6637 & 0.3333 \end{bmatrix} \begin{bmatrix} \frac{1}{2.759} & 0 \\ 0 & \frac{1}{2.034} \end{bmatrix} \begin{bmatrix} 0.498 & 0.867 \\ 0.867 & -0.498 \end{bmatrix} I \\
&= \begin{bmatrix} -0.167 & 0.222 \\ 0.190 & -0.302 \\ 0.262 & 0.127 \end{bmatrix}
\end{aligned}
$$

and thus the new minimal solution with respect to the alternative norm definition will be

$$\begin{bmatrix} -0.167 & 0.222 \\ 0.190 & -0.302 \\ 0.262 & 0.127 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \approx \begin{bmatrix} 0 \\ -0.146 \\ 1.43 \end{bmatrix}$$

Exercise 17.1

(a) Trace $= -3 < 0$, Determinant $= 2 > 0$, Discriminant $= (-3)^2 - 4(2) = 1 > 0$: Stable Node. The general solution is $\mathbf{y} = c_1(1,0)^T e^{-t} + c_2(1,1)^T e^{-2t}$;

(b) Trace $= 2 > 0$, Determinant $= 2 > 0$, Discriminant $= 2^2 - 4(2) = -4 < 0$: Unstable Spiral. The general solution is $\mathbf{y} = (c_1, c_2)^T e^t \cos(t) + (-c_1 - c_2, 2c_1 + c_2)^T e^t \sin(t)$;

(c) Trace $= 3 > 0$, Determinant $= 2 > 0$, Discriminant $= 3^2 - 4(2) = 1 > 0$: Unstable Node. The general solution is $\mathbf{y} = c_1(2,1)^T e^{2t} + c_2(1,1)^T e^t$;

(d) Trace $= 3 > 0$, Determinant $= 0$, Discriminant $= 3^2 - 4(0) = 9 > 0$: Degenerate case. The generals solution is $\mathbf{y} = c_1(-1,2)^T e^{3t} + c_2(1,1)^T$.

Exercise 17.2 The equilibrium point is found by solving $A\tilde{\mathbf{y}} = -G$ which gives $\tilde{\mathbf{y}} = (\frac{3}{2}, 0, -\frac{1}{2})^T$. The eigenvalues of A are $\lambda = 2, -2 \pm 2i$ so it is an inward spiral on a cross-section that extends outwards along a cut-through axis.

Exercise 17.3 There are four equilibrium points: $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^T$, $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$, $(-\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})^T$, and $(\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}})^T$. The Jacobian matrix can be found as

$$\begin{bmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} \\ \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial(x^2 - 2y^2 + xy)}{\partial x} & \dfrac{\partial(x^2 - 2y^2 + xy)}{\partial y} \\ \dfrac{\partial(x^2 + y^2 - 1)}{\partial x} & \dfrac{\partial(x^2 + y^2 - 1)}{\partial y} \end{bmatrix}$$

$$= \begin{bmatrix} 2x + y & -4y + x \\ 2x & 2y \end{bmatrix}$$

Take the first equilibrium point $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^T$ as an example, the values of the Jacobian matrix will be

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

Its trace is $-\frac{5}{\sqrt{2}} < 0$, the determinant is $6 > 0$ and the discriminant is $(-\frac{5}{\sqrt{2}})^2 - 4(6) = -\frac{23}{2} < 0$. Hence it is a stable spiral. Similarly, we can deduce that $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$ is an unstable spiral, while $(-\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})^T$ and $(\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}})^T$ are saddle points.

Exercise 18.1

(a)

$$\begin{aligned} \nabla \cdot (\nabla \times \vec{v}) &= \partial_i(\epsilon_{ijk}\partial_j v_k) \\ &= \epsilon_{ijk}\partial_i\partial_j v_k \\ &= \epsilon_{ijk}\partial_j\partial_i v_k \\ &= -\epsilon_{jik}\partial_j\partial_i v_k \quad \text{(Flipping the order of the epsilon symbol)} \\ &= -\epsilon_{ijk}\partial_i\partial_j v_k \quad \text{(Swapping the roles of } i \text{ and } j\text{)} \end{aligned}$$

Since $\nabla \cdot (\nabla \times \vec{v}) = \epsilon_{ijk}\partial_i\partial_j v_k = -\epsilon_{ijk}\partial_i\partial_j v_k$, it must be zero. (The two partial derivatives are symmetric but the epsilon symbol is anti-symmetric.)

(b)

$$\nabla \times \nabla u = \epsilon_{ijk}\partial_j\partial_k u$$

By the same logic as the part above, it has to give the zero vector.

(c)

$$\begin{aligned}
\nabla \times (\nabla \times \vec{v}) &= \epsilon_{ijk}\partial_j(\epsilon_{klm}\partial_l v_m) \\
&= \epsilon_{kij}\epsilon_{klm}(\partial_j\partial_l v_m) \\
&= (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl})(\partial_j\partial_l v_m) \\
&= \partial_j\partial_i v_j - \partial_j\partial_j v_i = \partial_i\partial_j v_j - \partial_j\partial_j v_i \\
&= \nabla(\nabla \cdot \vec{v}) - \nabla^2\vec{v}
\end{aligned}$$

(d)

$$\begin{aligned}
\nabla \cdot (\vec{u} \times \vec{v}) &= \partial_i(\epsilon_{ijk}u_j v_k) \\
&= \epsilon_{ijk}\partial_i(u_j v_k) \\
&= \epsilon_{ijk}\partial_i(u_j)v_k + \epsilon_{ijk}u_j\partial_i(v_k) \\
&= v_k\epsilon_{kij}\partial_i u_j - u_j\epsilon_{jik}\partial_i v_k \\
&= \vec{v} \cdot (\nabla \times \vec{u}) - \vec{u} \cdot (\nabla \times \vec{v})
\end{aligned}$$

Exercise 18.2

$$\begin{aligned}
(-1)^n[(A^n)_{ij} + c_{n-1}(A^{n-1})_{ij} + \cdots + c_1 A_{ij} + c_0\delta_{ij}] &= [\mathbf{0}]_{ij} \\
(A^n)_{ij} + c_{n-1}(A^{n-1})_{ij} + \cdots + c_1 A_{ij} &= -c_0\delta_{ij} \\
[(A^n)_{ij} + c_{n-1}(A^{n-1})_{ij} + \cdots + c_1 A_{ij}]\delta_{ij} &= -c_0\delta_{ij}\delta_{ij} \\
(A^n)_{jj} + c_{n-1}(A^{n-1})_{jj} + \cdots + c_1 A_{jj} &= -c_0\delta_{jj} \\
\mathrm{tr}(A^n) + c_{n-1}\mathrm{tr}(A^{n-1}) + \cdots + c_1\mathrm{tr}(A) &= -nc_0
\end{aligned}$$

Exercise 18.3 Contracting the expression with the vector $\vec{u}$:

$$\begin{aligned}
u_i w_i &= u_i\epsilon_{ijk}u_j v_k \\
&= \epsilon_{ijk}u_i u_j v_k = 0
\end{aligned}$$

which gives the zero scalar due to the antisymmetric nature of the epsilon symbol and two symmetric $u_i u_j$ factors. Then by Quotient Law, the R.H.S. is a trivially contracted rank-0 tensor and $\vec{u}$ on L.H.S. is a rank-1 tensor, so $\vec{w}$ has to be a rank-1 tensor (vector) as well.

Exercise 18.4 The Mohr's Circle will lie on the x-axis and must touch the two points. Let the center of the Mohr's Circle be at $(x_m, 0)^T$ and its radius be r_m, then

$$\begin{cases} (7 - x_m)^2 + 4^2 = r_m^2 \\ (9 - x_m)^2 + (-3)^2 = r_m^2 \end{cases}$$

Solving these two simultaneous equations yield $x = \frac{25}{4}$, $r_m = \frac{\sqrt{265}}{4}$. The required angle can be found as half of the subtended arc by these two points on the Mohr's Circle:

$$\theta = \frac{1}{2}\left[\arctan\left(\frac{4}{7 - \frac{25}{4}}\right) - \arctan\left(\frac{-3}{9 - \frac{25}{4}}\right)\right] \approx 63.4^\circ$$

Exercise 18.5 No, Yes, No, Yes.

Exercise 18.6 The continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0$$

Multiplying this by q gives $q\frac{\partial \rho}{\partial t} + q\nabla \cdot (\rho \mathbf{v}) = 0$. On the other hand, $\frac{dq}{dt} = 0$ leads to

$$\frac{\partial q}{\partial t} + \mathbf{v} \cdot \nabla q = 0$$
$$\rho\frac{\partial q}{\partial t} + \rho\mathbf{v} \cdot \nabla q = 0$$

Combining these two equations then gives

$$\frac{\partial \rho}{\partial t} q + \rho \frac{\partial q}{\partial t} + q \nabla \cdot (\rho \mathbf{v}) + \rho \mathbf{v} \cdot \nabla q = 0$$
$$\frac{\partial (\rho q)}{\partial t} + \nabla \cdot (\rho q \mathbf{v}) = 0$$

(Refer to Footnote 13 with q replacing ρ and $\rho \mathbf{v}$ replacing $\mathbf{v}$.)

Appendix A

Supplementary Information for the Main Text

A.1. Chapter 2

Properties 2.3.4 Consider an $n \times n$ square matrix A that has two identical and adjacent rows with indices i_1 and i_2, where $i_2 = i_1 + 1$ (hence one of the indices is odd and another is even), then cofactor expansion along the odd row (let's say i_1) will give

$$\begin{aligned} |A| &= \sum_{k=1}^{n} A_{i_1 k} C_{i_1 k} \\ &= \sum_{k=1}^{n} (-1)^{i_1+k} A_{i_1 k} \det\left(M_{i_1 k}\right) \end{aligned} \tag{A.1a}$$

by Properties 2.3.3 and Definition 2.3.2. Similarly, by considering the even row, we have

$$|A| = \sum_{k=1}^{n} A_{i_2 k} C_{i_2 k}$$

$$= \sum_{k=1}^{n} (-1)^{i_2+k} A_{i_2 k} \det\left(M_{i_2 k}\right) \tag{A.1b}$$

But since the i_1-th and i_2-th row are identical, $A_{i_1 k} = A_{i_2 k}$. Furthermore, as these two identical rows are also adjacent, the minors $M_{i_1 k} = M_{i_2 k}$ are also equal. The only difference between the two expressions for $|A|$ above is the $(-1)^{i+j}$ factor. And because i_1 is odd and i_2 is even, they are differed by a negative sign only. Explicitly, we have

$$\begin{aligned}
|A| &= \sum_{k=1}^{n} (-1)^{i_1+k} A_{i_1 k} \det\left(M_{i_1 k}\right) \\
&= \sum_{k=1}^{n} (-1)^{(i_2-1)+k} A_{i_2 k} \det\left(M_{i_2 k}\right) \\
&= (-1) \sum_{k=1}^{n} (-1)^{i_2+k} A_{i_2 k} \det\left(M_{i_2 k}\right) \\
&= -|A|
\end{aligned} \tag{A.2}$$

Therefore $|A| = -|A|$ and $|A| = 0$ must equal to zero. Now we can generalize the results to non-adjacent, proportional rows by doing the first and third kind (multiplication and swapping) of elementary row operations when appropriate with the aid of Properties 2.3.5, and the second case in Properties 2.3.4 is completed. Subsequently, for the addition/subtraction type of elementary row operations, let's say $R_p + cR_q \to R_p$ is applied on some matrix A (this A is not the same one as in the first part) to produce A', then

$$A' = \begin{bmatrix} \vdots & \vdots & & \vdots \\ A_{p1} + cA_{q1} & A_{p2} + cA_{q2} & \cdots & A_{pn} + cA_{qn} \\ \vdots & \vdots & & \vdots \\ A_{q1} & A_{q2} & \cdots & A_{qn} \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

where we have only written out the rows R_p and R_q. By applying cofactor expansion along R_p following Properties 2.3.3, we have

$$\begin{aligned} |A'| &= \sum_{k=1}^{n} [(A_{pk} + cA_{qk})C_{pk}] \\ &= \sum_{k=1}^{n} A_{pk}C_{pk} + c\sum_{k=1}^{n} A_{qk}C_{pk} \end{aligned} \tag{A.3}$$

We identify the first term with $|A|$ that is computed using cofactor expansion on the row R_p of A. The second term can be thought of as the determinant of a matrix $\tilde{A}$ that is formed by replacing the p-th row R_p by the q-th row R_q in A and subsequently expanded along that row. So $\tilde{A}$ practically has two identical rows $R_p = R_q$, and by the previous result, the value of $\left|\tilde{A}\right|$ is zero. Therefore

$$|A'| = |A| + c\left|\tilde{A}\right| = |A| + c(0) = |A| \tag{A.4}$$

implying that the addition/subtraction type of elementary row operations does not affect the value of the determinant.

A.2. Chapter 6

Properties 6.1.6 $\mathcal{W}$ as a subspace of $\mathbb{R}^n$ contains at most n linearly independent vectors. Assume $\mathcal{W}$ is non-empty and take any non-zero vector in $\mathcal{W}$, denote it by $\vec{v}^{(1)}$. The span of $\beta = \{\vec{v}^{(1)}\}$ is itself a subspace $\mathcal{B} = \text{span}(\beta)$ of the subspace $\mathcal{W}$ by noting that it is a subset of $\mathcal{W}$ and using the first part of Properties 6.1.6. If this subspace $\mathcal{B}$ of the subspace $\mathcal{W}$ is exactly $\mathcal{W}$, in the sense that no other $\vec{v} \in \mathcal{W}$ is not included by $\mathcal{B} = \text{span}(\beta)$, then we are done because this subspace is a span by construction. Otherwise, take another non-zero vector $\vec{v}^{(2)} \in \mathcal{W}$ which is linearly independent of $\vec{v}^{(1)}$ and add it to β. Now $\mathcal{B} = \text{span}(\beta) = \text{span}(\{\vec{v}^{(1)}, \vec{v}^{(2)}\})$ is enlarged by one dimension but still a subspace of $\mathcal{W}$ and we can check if it coincides with $\mathcal{W}$, otherwise, repeat the

procedure with $\vec{v}^{(3)}, \vec{v}^{(4)}, \ldots$ until $\text{span}(\mathcal{B}) = \mathcal{W}$ then we can stop. Remember we can at most add up to $\vec{v}^{(n)}$ since $\mathcal{W}$ contains at most n linearly independent vectors, so in the middle of somewhere we would done with $\vec{v}^{(p)}$, where $p \leq n$, and hence we know that $\mathcal{W}$ is some span (and a p-dimensional subspace).

Dependence relations 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 $A = [\vec{v}^{(1)}|\vec{v}^{(2)}|\cdots|\vec{v}^{(r)}|\vec{v}^{(r+1)}]$, i.e.

$$\begin{bmatrix} \vdots \\ \vec{v}_p^{(r+1)} \\ \vdots \\ \vec{v}_q^{(r+1)} \\ \vdots \end{bmatrix} = c_1 \begin{bmatrix} \vdots \\ \vec{v}_p^{(1)} \\ \vdots \\ \vec{v}_q^{(1)} \\ \vdots \end{bmatrix} + c_2 \begin{bmatrix} \vdots \\ \vec{v}_p^{(2)} \\ \vdots \\ \vec{v}_q^{(2)} \\ \vdots \end{bmatrix} + \cdots + c_r \begin{bmatrix} \vdots \\ \vec{v}_p^{(r)} \\ \vdots \\ \vec{v}_q^{(r)} \\ \vdots \end{bmatrix}$$

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

$$\vec{v}^{(j)\prime} = \begin{bmatrix} \vdots \\ \vec{v}_p^{(j)} + c_q\vec{v}_q^{(j)} \\ \vdots \\ \vec{v}_q^{(j)} \\ \vdots \end{bmatrix}$$

for all j. Therefore,

$$\vec{v}^{(r+1)'} = \begin{bmatrix} \vdots \\ \vec{v}_p^{(r+1)} + c_q \vec{v}_q^{(r+1)} \\ \vdots \\ \vec{v}_q^{(r+1)} \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \vdots \\ (c_1\vec{v}_p^{(1)} + c_2\vec{v}_p^{(2)} + \cdots + c_r\vec{v}_p^{(r)}) + c_q(c_1\vec{v}_q^{(1)} + c_2\vec{v}_q^{(2)} + \cdots + c_r\vec{v}_q^{(r)}) \\ \vdots \\ (c_1\vec{v}_q^{(1)} + c_2\vec{v}_q^{(2)} + \cdots + c_r\vec{v}_q^{(r)}) \\ \vdots \end{bmatrix}$$

by the original dependence relation, which is then equal to

$$\begin{bmatrix} \vdots \\ c_1(\vec{v}_p^{(1)} + c_q\vec{v}_q^{(1)}) \\ \vdots \\ c_1\vec{v}_q^{(1)} \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ c_2(\vec{v}_p^{(2)} + c_q\vec{v}_q^{(2)}) \\ \vdots \\ c_2\vec{v}_q^{(2)} \\ \vdots \end{bmatrix} + \cdots + \begin{bmatrix} \vdots \\ c_r(\vec{v}_p^{(r)} + c_q\vec{v}_q^{(r)}) \\ \vdots \\ c_r\vec{v}_q^{(r)} \\ \vdots \end{bmatrix}$$

$$= c_1\vec{v}^{(1)'} + c_2\vec{v}^{(2)'} + \cdots + c_r\vec{v}^{(r)'}$$

This shows that the same dependence relation holds between the new column vectors $\vec{v}^{(j)'}$, $j = 1, 2, \ldots, r+1$.

Steinitz Replacement Theorem

Theorem A.2.1 (Steinitz Replacement Theorem)**.** Given a vector space $\mathcal{V}$ generated by a set γ consisting of n vectors (not necessarily linearly independent), and another set η containing m linearly independent vectors from $\mathcal{V}$. Then

$m \leq n$ and there exists a subset γ' of γ with exactly $n - m$ vectors such that $\eta \cup \gamma'$ spans $\mathcal{V}$.

Proof. We proceed with mathematical induction on m. The base case $m = 0$ is trivial as $\eta = \varnothing$ and $\gamma' = \gamma$. Now assume that the statement is true for some integer $m = k \geq 0$ and we have to prove it for $m = k + 1$. Let $\eta = \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}, \vec{v}^{(k+1)}\}$ be a subset of $\mathcal{V}$ with $m = k+1$ linearly independent vectors. It is apparent that after removing $\vec{v}^{(k+1)}$ from η, $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}\}$ is still linearly independent. Then we can use the induction hypothesis to obtain $k \leq n$ and a subset $\{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}$ of γ so that $\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}\} \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}$ generates $\mathcal{V}$. So $\vec{v}^{(k+1)} \in \mathcal{V}$ can be written as the linear combination of

$$\vec{v}^{(k+1)} = a_1\vec{v}^{(1)} + a_2\vec{v}^{(2)} + \cdots + a_k\vec{v}^{(k)} + b_1\vec{u}^{(1)} + b_2\vec{u}^{(2)} + \cdots + b_{n-k}\vec{u}^{(n-k)}$$

It must be true that $n - k \geq 1$, so that some b_j are present for otherwise $\vec{v}^{(k+1)}$ will be reduced to a linear combination of $\vec{v}^{(j)}$ and contradicts the assumption that η is linearly independent. Hence $m = k + 1 \leq n$. By part (b) of Theorem 6.1.11, $\operatorname{span}(\eta \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) = \operatorname{span}(\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}, \vec{v}^{(k+1)}\} \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) = \operatorname{span}(\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}\} \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) = \mathcal{V}$ (the last equality follows from the induction hypothesis). For the same reason, in particular, some b_j has to be non-zero, let's say b_1. Therefore we can write

$$\vec{u}_1 = -\frac{a_1}{b_1}\vec{v}^{(1)} - \frac{a_2}{b_1}\vec{v}^{(2)} - \cdots - \frac{a_k}{b_1}\vec{v}^{(k)} + \frac{1}{b_1}\vec{v}^{(k+1)} - \frac{b_2}{b_1}\vec{u}^{(2)} - \cdots - \frac{b_{n-k}}{b_1}\vec{u}^{(n-k)}$$

as a linear combination of $\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}, \vec{v}^{(k+1)}$ and $\vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}$. Let $\gamma' = \{\vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}$. Again by part (b) of Theorem 6.1.11

$$\begin{aligned}
&\operatorname{span}(\eta \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) \\
&= \operatorname{span}(\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}, \vec{v}^{(k+1)}\} \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) \\
&= \operatorname{span}(\{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(k)}, \vec{v}^{(k+1)}\} \cup \{\vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\})
\end{aligned}$$

$$= \operatorname{span}(\eta \cup \gamma')$$

But we just have $\operatorname{span}(\eta \cup \{\vec{u}^{(1)}, \vec{u}^{(2)}, \ldots, \vec{u}^{(n-k)}\}) = \mathcal{V}$ from above, which readily shows that $\operatorname{span}(\eta \cup \gamma') = \mathcal{V}$. As γ' is a subset of γ with $n - k - 1 = n - (k + 1)$ vectors, the theorem is true for $m = k + 1$, and the induction is completed. □

The key point of the theorem is that some vectors in a generating set of a vector space can be replaced by the same number of linearly independent vectors from that vector space, hence comes the name of the replacement theorem.

A.3. Chapter 14

Bessel's Inequality By part (d) of Theorem 12.3.4, we can expand a function f as

$$f = \lim_{p \to \infty} \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \tag{A.5}$$

where $\varphi^{(j)}$ are the orthonormal basis vectors/functions and the partial sum $S_p[f]$ will be just $\sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}$. Particularly, the completeness of the orthonormal

basis in Properties 12.1.8 means that (see the footnote below)[1]

$$\left\| f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\|^2 = 0 \tag{A.6}$$

On the other hand,

$$\begin{aligned}
\left\| f - S_p[f] \right\|^2 &= \left\| f - \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\|^2 \\
&= \left\langle f - \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}, f - \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\rangle \\
&= \langle f, f \rangle - \left\langle f, \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\rangle - \left\langle \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}, f \right\rangle \\
&\quad + \left\langle \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}, \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\rangle
\end{aligned}$$

[1] Otherwise, if $\| f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \|^2 > 0$, then consider $\tilde{\varphi} = f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}$ which is now a non-zero vector. It will be orthogonal to any of the original basis vectors $\varphi^{(j')}$ as

$$\begin{aligned}
\langle \tilde{\varphi}, \varphi^{(j')} \rangle &= \langle f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}, \varphi^{(j')} \rangle \\
&= \langle f, \varphi^{(j')} \rangle - \langle \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}, \varphi^{(j')} \rangle \\
&= \langle f, \varphi^{(j')} \rangle - (\cdots + (0) + \langle f, \varphi^{(j')} \rangle (1) + (0) + \cdots) \\
&\quad \text{(Orthonormality of the basis: } \langle \varphi^{(j)}, \varphi^{(j')} \rangle = 1 \\
&\quad \text{only when } j = j' \text{ and } 0 \text{ if } j \neq j') \\
&= \langle f, \varphi^{(j')} \rangle - \langle f, \varphi^{(j')} \rangle = 0
\end{aligned}$$

which violates the premise of completeness as it will become another vector that is linearly independent of all the other basis vectors and can be added to the basis.

$$
\begin{aligned}
&= \|f\|^2 - \left\langle f, \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\rangle \\
&\quad - \overline{\left\langle f, \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right\rangle} \quad \text{((1) of Definition 12.1.1)} \\
&\quad + \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \quad \text{(Orthonormality of the basis)} \\
&= \|f\|^2 - 2\,\mathrm{Re}\left\{ \langle f, \sum_{j=1}^{p} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \rangle \right\} + \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \\
&= \|f\|^2 - 2\,\mathrm{Re}\left\{ \sum_{j=1}^{p} (\overline{\langle f, \varphi^{(j)} \rangle} \langle f, \varphi^{(j)} \rangle) \right\} \\
&\quad \text{((4) of Properties 12.1.2)} \\
&\quad + \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \\
&= \|f\|^2 - 2 \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 + \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \\
&= \|f\|^2 - \sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \qquad \text{(A.7)}
\end{aligned}
$$

and since $\left\| f - S_p[f] \right\|^2 \geq 0$ is always non-negative, we have the *Bessel's Inequality* as

$$
\sum_{j=1}^{p} \left| \langle f, \varphi^{(j)} \rangle \right|^2 \leq \|f\|^2 \qquad \text{(A.8)}
$$

From Exercise 12.4, we can apply the Triangle Inequality to get

$$
\left\| \left(f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} \right) + \left(\sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} - S_p[f] \right) \right\|^2
$$

$$= \left\|f - S_p[f]\right\|^2 \leq \left\|f - \sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}\right\|^2 + \left\|\sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} - S_p[f]\right\|^2 \tag{A.9}$$

By Equation (A.6), the first term on R.H.S. of the equality is zero. And the second term

$$\begin{aligned} &\left\|\sum_{j=1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)} - S_p[f]\right\|^2 \\ &= \left\|\sum_{j=p+1}^{\infty} \langle f, \varphi^{(j)} \rangle \varphi^{(j)}\right\|^2 = \sum_{j=p+1}^{\infty} \left|\langle f, \varphi^{(j)} \rangle\right|^2 \end{aligned} \tag{A.10}$$

is a remainder term and when $p \longrightarrow \infty$, must tend to zero, because $\sum_{j=1}^{p} \left|\langle f, \varphi^{(j)} \rangle\right|^2$ is a convergent sequence, seen by applying the *Monotone Convergence Theorem* from elementary Analysis on Bessel's Inequality (A.8). Therefore, the L.H.S.

$$\lim_{p \to \infty} \left\|f - S_p[f]\right\|^2 = 0 \tag{A.11}$$

also tends to zero as we take the same limit, which implies the L^2 convergence for a complete orthonormal basis.

A.4. Chapter 15

Properties 15.2.4 *Reference Materials: [6]*

Let $\vec{q}$ be any eigenvector of A^T that has an eigenvalue of modulus $|\lambda| = 1$. Then $A^T \vec{q} = \lambda \vec{q}$ by the definition of an eigenvalue-eigenvector problem, i.e.

$$\sum_{k=1}^{n} a_{ki} \vec{q}_k = \lambda \vec{q}_i \tag{A.12}$$

where a_{ji} is the (j,i) $[(i,j)]$ entry of the matrix A $[A^T]$. Then

$$|\lambda\vec{q}_i| = |\lambda||\vec{q}_i| = \left|\sum_{k=1}^{n} a_{ki}\vec{q}_k\right|$$

$$|\vec{q}_i| \leq \sum_{k=1}^{n} |a_{ki}||\vec{q}_k| = \sum_{k=1}^{n} a_{ki}|\vec{q}_k| \quad (|\lambda| = 1) \tag{A.13}$$

(Triangle Inequality for complex numbers and a_{ij} are real)

Now assume that for some fixed i, $|\vec{q}_i| \geq |\vec{q}_k|$ for $1 \leq k \leq n$. Then the inequality above becomes

$$|\vec{q}_i| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_k| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_i| = |\vec{q}_i| \qquad (\sum_{k=1}^{n} a_{ki} \text{ sums to } 1) \tag{A.14}$$

By squeezing using $|\vec{q}_i|$ at the both ends, the part $\sum_{k=1}^{n} a_{ki}|\vec{q}_k| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_i|$ forces that $|\vec{q}_k| = |\vec{q}_i|$ for all k and it also holds for any i. This is where the positiveness of A is needed, otherwise a_{ki} can be 0 and the squeeze becomes vacuous ($0|\vec{q}_k| = 0|\vec{q}_i| = 0$). Moreover, as $|\vec{q}_i| = \left|\sum_{k=1}^{n} a_{ki}\vec{q}_k\right|$, incorporating it into the inequality

$$|\vec{q}_i| = \left|\sum_{k=1}^{n} a_{ki}\vec{q}_k\right| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_k| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_i| = |\vec{q}_i| \tag{A.15}$$

and apply squeezing again, the part of Triangle Inequality $\left|\sum_{k=1}^{n} a_{ki}\vec{q}_k\right| \leq \sum_{k=1}^{n} a_{ki}|\vec{q}_k|$ becomes an equality $\left|\sum_{k=1}^{n} a_{ki}\vec{q}_k\right| = \sum_{k=1}^{n} a_{ki}|\vec{q}_k|$, which means that all the components $\vec{q}_k$ have to lie along the same direction in the complex plane.[2] Therefore, $\vec{q}$ must be be in the form of $c(1,1,1,\ldots,1)^T$ where c will be

[2]For any two complex number z_1 and z_2, if $|z_1 + z_2| = |z_1| + |z_2|$, then

$$\begin{aligned} |z_1 + z_2|^2 &= (z_1 + z_2)\overline{(z_1 + z_2)} \\ &= z_1\overline{z_1} + z_1\overline{z_2} + z_2\overline{z_1} + z_2\overline{z_2} \\ &= |z_1|^2 + z_1\overline{z_2} + z_2\overline{z_1} + |z_2|^2 \end{aligned}$$

any complex constant and this shows that over $\mathbb{C}$, the only linearly independent eigenvector for A^T corresponding to $|\lambda| = 1$ will be $(1, 1, 1, \ldots, 1)^T$, matching the derivation in Properties 15.2.2 where the eigenvalue is exactly $\lambda = 1$ and the geometric multiplicity is hence 1 as well. Again, by Properties 9.1.5, this result will then also hold for the original matrix A (but the form of the eigenvector will be different).

A.5. Chapter 18

Uniqueness of rank-2 and 3 Isotropic Tensors The uniqueness of the Kronecker delta as the only isotropic rank-2 tensor can be shown as follows. By a 90° positive rotation about the 3-axis, we have

$$\mathbf{e}_1' = \mathbf{e}_2 \tag{A.16a}$$
$$\mathbf{e}_2' = -\mathbf{e}_1 \tag{A.16b}$$
$$\mathbf{e}_3' = \mathbf{e}_3 \tag{A.16c}$$

---

but also $(|z_1| + |z_2|)^2 = |z_1|^2 + 2|z_1||z_2| + |z_2|^2$. Hence we have

$$z_1\overline{z_2} + z_2\overline{z_1} = 2|z_1||z_2|$$

Assume that z_1 and z_2 points in some directions so that $z_1 = |z_1|e^{i\theta_1}$ and $z_2 = |z_2|e^{i\theta_2}$, then

$$\begin{aligned} z_1\overline{z_2} + z_2\overline{z_1} &= |z_1|e^{i\theta_1}|z_2|e^{-i\theta_2} + |z_2|e^{i\theta_2}|z_1|e^{-i\theta_1} \\ &= |z_1||z_2|e^{i(\theta_1-\theta_2)} + |z_1||z_2|e^{-i(\theta_1-\theta_2)} \\ &= 2|z_1||z_2|\left(\frac{e^{i(\theta_1-\theta_2)} + e^{-i(\theta_1-\theta_2)}}{2}\right) \\ &= 2|z_1||z_2|\cos(\theta_1 - \theta_2) && \text{(Properties 8.1.7)} \end{aligned}$$

but this has to equal to $2|z_1||z_2|$, thus $\cos(\theta_1 - \theta_2) = 1$ and the arguments θ_1 and θ_2 will be the same. Now apply this logic repetitively to $\left|\sum_{k=1}^n a_{ki}\vec{q}_k\right| = \sum_{k=1}^n a_{ki}|\vec{q}_k|$.

and according to (18.21), the change of coordinates matrix is

$$a_{ij} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}'_1 & \mathbf{e}_1 \cdot \mathbf{e}'_2 & \mathbf{e}_1 \cdot \mathbf{e}'_3 \\ \mathbf{e}_2 \cdot \mathbf{e}'_1 & \mathbf{e}_2 \cdot \mathbf{e}'_2 & \mathbf{e}_2 \cdot \mathbf{e}'_3 \\ \mathbf{e}_3 \cdot \mathbf{e}'_1 & \mathbf{e}_3 \cdot \mathbf{e}'_2 & \mathbf{e}_3 \cdot \mathbf{e}'_3 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.17}$$

For a general rank-2 tensor F, by (18.22) its transformation will then be

$$\begin{aligned} \begin{bmatrix} F'_{11} & F'_{12} & F'_{13} \\ F'_{21} & F'_{22} & F'_{23} \\ F'_{31} & F'_{32} & F'_{33} \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} F_{22} & -F_{21} & F_{23} \\ -F_{12} & F_{11} & -F_{13} \\ F_{32} & -F_{31} & F_{33} \end{bmatrix} \end{aligned} \tag{A.18}$$

If the tensor is isotropic such that $F'_{ij} = F_{ij}$, then by comparing the entries on both sides, we must have

$$F_{11} = F_{22} \tag{A.19a}$$

$$F_{13} = F_{23} = -F_{13} \implies F_{13} = F_{23} = 0 \tag{A.19b}$$

$$F_{31} = F_{32} = -F_{31} \implies F_{31} = F_{32} = 0 \tag{A.19c}$$

By the same technique, we can carry out a 90° positive rotation about the 2-axis to obtain

$$F_{11} = F_{33} \tag{A.20a}$$

$$F_{12} = F_{32} = 0 \tag{A.20b}$$

$$F_{21} = F_{23} = 0 \tag{A.20c}$$

Therefore, $F_{11} = F_{22} = F_{33}$ and $F_{12} = F_{21} = F_{13} = F_{31} = F_{23} = F_{32} = 0$, the diagonal entries have the same value and all off-diagonal entries are zero, and any isotropic rank-2 tensor must be in the form of $\lambda\delta_{ij}$, a constant multiple of the Kronecker delta symbol. We can show the same for the rank-3 isotropic

tensor to be found. First, do a rotation about the $(1, 1, 1)^T$ axis so that the effect is a subscript permutation $1 \to 2 \to 3 \to 1$, then we have

$$T_{111} = T_{222} = T_{333} \tag{A.21a}$$
$$T_{112} = T_{223} = T_{331} \tag{A.21b}$$
$$T_{122} = T_{233} = T_{311} \tag{A.21c}$$
$$T_{212} = T_{323} = T_{131} \tag{A.21d}$$
$$T_{211} = T_{322} = T_{133} \tag{A.21e}$$
$$T_{121} = T_{232} = T_{313} \tag{A.21f}$$
$$T_{221} = T_{332} = T_{113} \tag{A.21g}$$
$$T_{123} = T_{231} = T_{312} \tag{A.21h}$$
$$T_{132} = T_{321} = T_{213} \tag{A.21i}$$

Next, apply a 90° positive rotation about the 3-axis as suggested by (A.17) before. After some tedious algebra, we can obtain the following relations:

$$T_{111} = -T_{222} \implies T_{111} = T_{222} = T_{333} = 0 \tag{A.22a}$$
$$T_{112} = T_{221} \quad \text{and} \quad T_{221} = -T_{112} \implies T_{112} = T_{221} = 0 \tag{A.22b}$$
$$T_{122} = T_{211} \quad \text{and} \quad T_{211} = -T_{122} \implies T_{122} = T_{211} = 0 \tag{A.22c}$$
$$T_{121} = T_{212} \quad \text{and} \quad T_{212} = -T_{121} \implies T_{121} = T_{212} = 0 \tag{A.22d}$$
$$T_{123} = -T_{213} \implies \begin{aligned} &T_{123} = T_{231} = T_{312} \\ &= -T_{213} = -T_{321} = -T_{132} \end{aligned} \tag{A.22e}$$

The first four of them show that the rank-3 isotropic tensor must take a value of zero when there are repeated subscripts. The last relation implies that it will have an opposite sign when the subscripts belong to an odd/even permutation respectively. This coincides with Definition 18.1.2 for the epsilon symbol and thus any isotropic rank-3 tensor must be in the form of $\lambda\epsilon_{ijk}$, a constant multiple of the epsilon symbol.

Appendix B

Cyclic Subspaces and Jordan Normal Forms

Note that this chapter is basically a rewritten summary of Friedberg's Section 5.4 and Chapter 7.

B.1. From Cyclic Subspaces to Cayley-Hamilton Theorem

A ***cyclic subspace*** of a vector space is generated by repeatedly applying the same linear operator on a fixed initial vector and taking the span of all the resulting vectors.

Definition B.1.1 (Cyclic Subspace)**.** Given a vector space $\mathcal{V}$ with a linear operator $T : \mathcal{V} \to \mathcal{V}$ on it, and choose a non-zero vector $\vec{v} \in \mathcal{V}$, then the subspace

$$\mathcal{W} = \operatorname{span}(\{\vec{v}, T(\vec{v}), T^2(\vec{v}), \ldots\}) \tag{B.1}$$

is known as the *T-cyclic subspace of $\mathcal{V}$ generated by $\vec{v}$*.

It is obviously T-invariant (Definition 9.1.7) because for any vector $\vec{w} \in \mathcal{W}$, if we write $\vec{w} = c_0\vec{v} + c_1T(\vec{v}) + c_2T^2(\vec{v}) + \cdots$ as a span (Definition 6.1.5, extended to an infinite one if needed), then $T(\vec{w}) = c_0T(\vec{v}) + c_1T^2(\vec{v}) + c_2T^3(\vec{v}) + \cdots \in \mathcal{W}$ lies in the T-cyclic subspace too. It is also the smallest T-invariant subspace that contains $\vec{v}$: For other T-invariant subspaces $\mathcal{W}'$ that also contain $\vec{v}$, by definition they contain $T(\vec{v}), T^2(\vec{v}), \ldots$ as well, and by Properties 6.1.8 they have to completely contain the corresponding T-cyclic subspace $\mathcal{W} = \text{span}(\{\vec{v}, T(\vec{v}), T^2\vec{v}, \ldots\})$ too, i.e. $\mathcal{W} \subseteq \mathcal{W}'$.

To proceed further, we need some more intermediate results. First,

Properties B.1.2. Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a *finite-dimensional* vector space $\mathcal{V}$, and $\mathcal{W}$ be a T-invariant subspace of $\mathcal{V}$. Then, the characteristic polynomial (Definition 9.1.2) of the restriction of T to $\mathcal{W}$, that is, $T|_W$ (Definition 8.3.5), divides the characteristic polynomial of T.

We refer the readers to Friedberg for the proof. It means that if the characteristic polynomial of $T|_W$ is $g(t)$ and that of T is $f(t)$, then there exists some other polynomial $q(t)$ such that $f(t) = q(t)g(t)$. Next,

Properties B.1.3. Given $T : \mathcal{V} \to \mathcal{V}$ as a linear operator on a *finite-dimensional* vector space $\mathcal{V}$ and let $\mathcal{W}$ be a T-cyclic subspace of $\mathcal{V}$ generated by a non-zero vector $\vec{v} \in \mathcal{V}$. If $\dim(\mathcal{W}) = k$, then

(a) $\{\vec{v}, T(\vec{v}), T^2(\vec{v}), \ldots, T^{k-1}(\vec{v})\}$ is a basis for $\mathcal{W}$, and

(b) If $a_0\vec{v} + a_1T(\vec{v}) + \cdots + a_{k-1}T^{k-1}(\vec{v}) + T^k(\vec{v}) = \mathbf{0}$, then the characteristic polynomial of $T|_W$ is $g(t) = (-1)^k(a_0 + a_1t + \cdots + a_{k-1}t^{k-1} + t^k)$.

Again, we refer the readers to Friedberg for the proof of part (a) as it is explained in detail there. Meanwhile, to show (b), denote the basis in part (a) as $\beta = \{\vec{v}, T(\vec{v}), T^2(\vec{v}), \ldots, T^{k-1}(\vec{v})\}$. Then we have $T(T^j(\vec{v})) = T^{j+1}(\vec{v})$ for $0 \le j < k-1$, and $T(T^{k-1}(\vec{v})) = T^k(\vec{v}) = -a_0\vec{v} - a_1T(\vec{v}) - \cdots - a_{k-1}T^{k-1}(\vec{v})$.

Subsequently, the linear operator $T|_W$ will have a matrix representation of

$$[T|_W]_\beta = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_{k-1} \end{bmatrix} \tag{B.2}$$

This is called a *companion matrix*. Its characteristic polynomial (in t) is then directly obtained by repeated cofactor expansion (Properties 2.3.3) along the rows from top to bottom:

$$\begin{aligned}
&\det\big([T|_W]_\beta - tI\big) \\
&= \begin{vmatrix} -t & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & -t & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & -t & \cdots & 0 & -a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_{k-1}-t \end{vmatrix} \\
&= -t \begin{vmatrix} -t & 0 & \cdots & 0 & -a_1 \\ 1 & -t & \cdots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{k-1}-t \end{vmatrix} + (-a_0)(-1)^{k+1} \begin{vmatrix} 1 & -t & 0 & \cdots & 0 \\ 0 & 1 & -t & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} \\
&= -t \begin{vmatrix} -t & 0 & \cdots & 0 & -a_1 \\ 1 & -t & \cdots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{k-1}-t \end{vmatrix} + a_0(-1)^k(1) \\
&\quad \text{(The determinant of an upper-triangular block is simply} \\
&\quad \text{the product of its diagonal entries, which are all 1 here)} \\
&= (-t)(-t) \begin{vmatrix} -t & \cdots & 0 & -a_2 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & -a_{k-1}-t \end{vmatrix} + (-t)(-a_1)(-1)^{(k+1)-1} \begin{vmatrix} 1 & -t & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix}
\end{aligned}$$

$$
\begin{aligned}
&+ a_0(-1)^k \\
&= t^2 \begin{vmatrix} -t & \cdots & 0 & -a_2 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & -a_{k-1} - t \end{vmatrix} + (a_0 + a_1 t)(-1)^k \\
&\vdots \\
&= (-t)^{k-1}(-a_{k-1} - t) + (a_0 + a_1 t + \cdots + a_{k-2}t^{k-2})(-1)^k \\
&= (a_0 + a_1 t + \cdots + a_{k-2}t^{k-2} + a_{k-1}t^{k-1} + t^k)(-1)^k \qquad \text{(B.3)}
\end{aligned}
$$

The Cayley-Hamilton Theorem then easily follows.

Theorem B.1.4 (Cayley-Hamilton Theorem for Linear Operators)**.** Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a finite-dimensional vector space $\mathcal{V}$ and $f(t)$ be its characteristic polynomial. Then $f(T) = T_0$ yields the zero transformation.

Proof. We have to show that for all $\vec{v} \in \mathcal{V}$, $[f(T)](\vec{v}) = \mathbf{0}$. Let $\mathcal{W}$ be the T-cyclic subspace generated by $\vec{v}$ and $\dim(\mathcal{W}) = k$. By part (a) of Properties B.1.3, $T^k(\vec{v})$ can be written as a linear combination of $\{\vec{v}, T(\vec{v}), T^2(\vec{v}), \ldots, T^{k-1}(\vec{v})\}$, and thus there exist scalars a_j such that $a_0\vec{v} + a_1 T(\vec{v}) + \cdots + a_{k-1}T^{k-1}(\vec{v}) + T^k(\vec{v}) = \mathbf{0}$. By part (b) of Properties B.1.3, the characteristic polynomial of $T|_W$ is then $g(t) = (-1)^k(a_0 + a_1 t + \cdots + a_{k-1}t^{k-1} + t^k)$. These two statements together imply that

$$
\begin{aligned}
[g(T)](\vec{v}) &= (-1)^k(a_0\vec{v} + a_1 T(\vec{v}) + \cdots + a_{k-1}T^{k-1}(\vec{v}) + T^k(\vec{v})) \\
&= (-1)^k[a_0 I + a_1 T + \cdots + a_{k-1}T^{k-1} + T^k](\vec{v}) = \mathbf{0}
\end{aligned}
$$

By Properties B.1.2, $g(t)$ divides the characteristic polynomial $f(t)$ of T so that $f(t) = q(t)g(t)$, and therefore

$$
[f(T)](\vec{v}) = [q(T)g(T)](\vec{v}) = [q(T)][g(T)(\vec{v})] = [q(T)](\mathbf{0}) = \mathbf{0}
$$

so we are done. The Cayley-Hamilton Theorem 9.1.9 for matrices is then an immediate corollary of this. □

B.2. Jordan Normal Forms

B.2.1. Generalized Eigenvectors, Existence of Jordan Normal Forms

We have seen that the matrix in Example 9.1.1

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

has only one eigenvalue $\lambda = 1$ and its geometric multiplicity = 1, is less than the algebraic multiplicity = 2: there is only one eigenvector $\vec{v}^{(1)} = (1, 0)^T$, so that $A\vec{v}^{(1)} = \lambda\vec{v}^{(1)}$, or equivalently $(A - \lambda I)\vec{v}^{(1)} = \mathbf{0}$. Now, let's iterate the procedure and consider $(A - \lambda I)\vec{v}^{(2)} = \vec{v}^{(1)}$, then we have

$$\begin{bmatrix} 1-\lambda & -1 \\ 0 & 1-\lambda \end{bmatrix} \vec{v}^{(2)} = \begin{bmatrix} 1-1 & -1 \\ 0 & 1-1 \end{bmatrix} \vec{v}^{(2)} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \vec{v}^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The solution to $\vec{v}^{(2)}$ is clearly $(0, -1)^T$. $\{\vec{v}^{(1)}, \vec{v}^{(2)}\}$ then forms a basis for $\mathbb{R}^2$ in this case. We have successfully salvaged the problem where an eigenvalue is deficient by finding a ***generalized eigenvector*** $\vec{v}^{(2)}$ that fulfills

$$\begin{aligned} (A - \lambda I)[(A - \lambda I)\vec{v}^{(2)}] &= (A - \lambda I)\vec{v}^{(1)} \\ (A - \lambda I)^2\vec{v}^{(2)} &= \mathbf{0} \end{aligned}$$

This motivates us to make the following definition.

Definition B.2.1. Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a vector space $\mathcal{V}$. A non-zero vector $\vec{v} \in \mathcal{V}$ is known as a generalized eigenvector of T corresponding to an eigenvalue of λ if

$$(T - \lambda I)^p(\vec{v}) = \mathbf{0} \tag{B.4}$$

for some integer p.

Clearly, generalized eigenvectors of $p = 1$ are just the ordinary eigenvectors. In parallel, we also have

Definition B.2.2. Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a vector space $\mathcal{V}$. The ***generalized eigenspace*** of T corresponding to an eigenvalue of λ is denoted by $\mathcal{K}_\lambda$, which is a subspace of $\mathcal{V}$ such that

$$\mathcal{K}_\lambda = \{\vec{v} \in \mathcal{V} \mid (T - \lambda I)^p(\vec{v}) = \mathbf{0} \text{ for some integer p.}\} \tag{B.5}$$

It is not hard to show that $\mathcal{K}_\lambda$ is T-invariant and contains the ordinary eigenspace $\mathcal{E}_\lambda$. We note two of the most relevant results of generalized eigenvectors/eigenspaces without proof.

Properties B.2.3. The generalized eigenvectors/eigenspaces corresponding to two distinct eigenvalues $\lambda \neq \mu$ are linearly independent.

Properties B.2.4. The generalized eigenspace $\mathcal{K}_\lambda$ for the linear operator T, corresponding to some fixed eigenvalue λ with an algebraic multiplicity of m, satisfies the following properties:

1. $\mathcal{K}_\lambda = \mathcal{N}((T - \lambda I)^m)$,
2. $\dim(\mathcal{K}_\lambda) = m$.

As a result, if T has a characteristic polynomial

$$p(\lambda) = (\lambda_1 - \lambda)^{m_1}(\lambda_2 - \lambda)^{m_2} \cdots (\lambda_k - \lambda)^{m_k} \tag{B.6}$$

then the respective generalized eigenspaces of all distinct eigenvalues form a direct sum that reproduces the full vector space:

$$\mathcal{V} = \mathcal{K}_{\lambda_1} \oplus \mathcal{K}_{\lambda_2} \oplus \cdots \oplus \mathcal{K}_{\lambda_k} \tag{B.7}$$

On the other hand, extending the above example and definitions of generalized eigenvectors/eigenspaces, we can make a ***cycle of generalized eigenvectors***.

Definition B.2.5. Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a vector space $\mathcal{V}$ and $\vec{v} \in \mathcal{V}$ be a generalized eigenvector of T corresponding to the eigenvalue λ. If p is the smallest positive integer such that $(T - \lambda I)^p(\vec{v}) = \mathbf{0}$, then

$$\begin{aligned} &\{(T - \lambda I)^{p-1}(\vec{v}), (T - \lambda I)^{p-2}(\vec{v}), \ldots, (T - \lambda I)(\vec{v}), \vec{v}\} \\ &:= \{\vec{v}^{(1)}, \vec{v}^{(2)}, \ldots, \vec{v}^{(p-1)}, \vec{v}^{(p)}\} \end{aligned} \tag{B.8}$$

is called a *cycle of generalized eigenvectors* of T corresponding to λ with a length of p. The vectors $\vec{v}^{(1)} := (T - \lambda I)^{p-1}(\vec{v})$ and $\vec{v}^{(p)} := \vec{v}$ are called the *initial vector* and *end vector* of that cycle respectively.

The initial vector of any cycle of generalized eigenvectors is then the only eigenvector in that cycle.

Now we are going to deduce the form of a *Jordan block* for a cycle of generalized eigenvectors of some linear operator T corresponding to an eigenvalue of λ. Consider the linear operator $T - \lambda I$ and the cycle described in (B.8). It is not hard to see that (B.8) is $(T - \lambda I)$-cyclic (as well as T-cyclic) and thus it forms a basis γ by part (a) of Properties B.1.3. Then the matrix representation of $T - \lambda I$ with respect to γ is readily seen to be

$$[T - \lambda I]_\gamma = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{B.9}$$

and thus the matrix representation of T, also with respect to γ, will be

$$[T]_\gamma = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix} \tag{B.10}$$

This $[T]_\gamma$ is then known as the ***Jordan block*** of size p corresponding to λ, which is "almost diagonal" with 1s filling above the main diagonal. Note that it has a characteristic polynomial of $g_\gamma(t) = (\lambda - t)^p$. Moreover, it is not hard to check that $([T - \lambda I]_\gamma)^p = [\mathbf{0}] = [T_0]_\gamma$ is the zero transformation and $T - \lambda I$ is known as ***nilpotent*** (in the generalized eigenspace $\mathcal{K}_\lambda$). To proceed, we first note some intermediate observations.

Theorem B.2.6. Given T as a linear operator on a vector space $\mathcal{V}$, if $\gamma_1, \gamma_2, \ldots, \gamma_q$ are cycles of generalized eigenvectors of T corresponding to the same eigenvalue λ such that the initial vectors of these γ_j are linearly independent, then these γ_j will be disjoint and their union $\gamma = \bigcup_{j=1}^{q} \gamma_j$ will also be linearly independent.

Theorem B.2.7. Given T as a linear operator on a *finite-dimensional* vector space $\mathcal{V}$ and λ as an eigenvalue of T, then the associated generalized eigenspace $\mathcal{K}_\lambda$ has a basis consisting of a union of disjoint cycles of generalized eigenvectors corresponding to λ.

These two results are derived in Friedberg and we refer the readers to it for the proof. Using our previous example, the matrix

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

has a basis made up of two generalized eigenvectors $\gamma = \{(1,0)^T, (0,-1)^T\}$ for the generalized eigenspace $\mathcal{K}_{\lambda=1}$, and thus we can express the matrix as a Jordan block using this ***Jordan canonical basis*** following (B.10) by a coordinate transformation to the γ system (refer to Properties 7.2.2):

$$J = Q^{-1} A Q$$

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

The final result is that we can derive the ***Jordan Normal Form*** of a matrix/linear operator by composing Jordan blocks as a direct sum.

Theorem B.2.8. Let $T : \mathcal{V} \to \mathcal{V}$ be a linear operator on a finite-dimensional vector space $\mathcal{V}$ and the characteristic polynomial of T *splits* over the chosen field (scalar: $\mathbb{R}$ or $\mathbb{C}$). Then T can be transformed into its Jordan Normal Form as the direct sum of smaller Jordan blocks. Using the matrix representation A, we have

$$J = Q^{-1}AQ \tag{B.11}$$

where Q consists of the union of Jordan canonical bases for all the Jordan blocks in columns.

By Theorem B.2.7, $\mathcal{K}_{\lambda_i}$ has a basis $\gamma_{\lambda_i} = \bigcup_{j=1}^{q} (\gamma_j)_{\lambda_i}$ formed by smaller cycles to the same eigenvalue λ_i and hence $[T|_{\mathcal{K}_{\lambda_i}}]_\gamma = (J_1)_{\lambda_i} \oplus (J_2)_{\lambda_i} \oplus \cdots (J_q)_{\lambda_i} = J_{\lambda_i}$, and by (B.7), we have

$$[T]_\beta = J_{\lambda_1} \oplus J_{\lambda_2} \oplus \cdots \oplus J_{\lambda_k} \tag{B.12}$$

as its Jordan Normal Form, where $\beta = \bigcup_{i=1}^{k} \gamma_{\lambda_i}$. The Jordan Normal Form of the last example is simply its only Jordan block:

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

To conclude, Jordan Normal Forms are the generalization and remedy of diagonalized forms where the constraint is relaxed to accommodate the case of deficient eigenvalues, so that it exists whenever the characteristic polynomial splits. But the price is that there will be some extra "1" above the main diagonal in the Jordan blocks of size ≥ 2. Conversely, if all the Jordan blocks are of size

1, then it will coincide with the usual diagonalization. For example,

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

consists of one Jordan block corresponding to $\lambda = 2$ of size 2 (red), one Jordan block corresponding to $\lambda = 3$ of size 3 (blue), and two Jordan blocks corresponding to $\lambda = 3$ of size 1 (green and purple). The positions of blocks can be swapped with each other. However, note that a characteristic polynomial (B.6) does not uniquely determine the Jordan Normal Form. In the above example, the characteristic polynomial is $p(\lambda) = (2-\lambda)^2(3-\lambda)^5$. But a different Jordan Normal Form like

$$J' = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

also shares the same characteristic polynomial. We note in passing that there is a closely related concept called a *minimal polynomial*, which the readers can search about if they are interested in exploring the deeper theoretical aspect.

B.2.2. Methods to Compute Jordan Normal Forms

In this subsection, we will investigate the techniques to compute the Jordan Normal Form of any given matrix practically.

Example B.2.1. Derive the Jordan Normal Form of

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

Solution. We leave it to the readers to check that the characteristic polynomial is $p(\lambda) = (1-\lambda)^2(2-\lambda)$. Hence, both the algebraic and geometric multiplicity of $\lambda = 2$ is 1, which is easy to deal with. The problem will be $\lambda = 1$ which has an algebraic multiplicity of 2, and there are two possibilities: if the geometric multiplicity is also 2, then the situation is the same as the usual diagonalization where we have two ordinary eigenvectors. Otherwise, if the geometric multiplicity is 1, then we will have a single Jordan block of length 2. Now consider the system $A - I = \mathbf{0}$:

$$\left[\begin{array}{ccc|c} 2 & 3 & -2 & 0 \\ -1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & \frac{3}{2} & -1 & 0 \\ -1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right] \qquad \frac{1}{2}R_1 \to R_1$$

$$\to \left[\begin{array}{ccc|c} 1 & \frac{3}{2} & -1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right] \qquad R_2 + R_1 \to R_2$$

$$\to \left[\begin{array}{ccc|c} 1 & \frac{3}{2} & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} 2R_2 \to R_2 \\ R_3 - R_2 \to R_3 \end{array}$$

The nullity is 1 and there is only one eigenvector. By letting the third entry be the free variable, it is not hard to see that the eigenvector will be $\vec{v}^{(1)} = (1, 0, 1)^T$. So there will be only one Jordan block, and due to this, we can obtain the next generalized eigenvector in a relatively simple reverse fashion by solving

$$(A - I)\vec{v}^{(2)} = \vec{v}^{(1)}$$

$$\begin{bmatrix} 2 & 3 & -2 \\ -1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \vec{v}^{(2)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

We can pick any particular solution for this, and again, we will leave it to the readers to confirm that $\vec{v}^{(2)} = (-1, 1, 0)^T$ is a valid choice. Finally, we simply note that the eigenvector for the other eigenvalue $\lambda = 2$ is $(-1, 1, 1)^T$. Thus, the required Jordan Normal Form is

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

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

□

Next, we will tackle a much harder case where there are multiple Jordan blocks corresponding to the same eigenvalue.

Example B.2.2. Convert the matrix

$$A = \begin{bmatrix} 5 & -1 & -1 & -2 & 0 & -1 \\ -1 & 19 & 9 & 24 & 6 & 17 \\ -1 & 3 & 5 & 4 & 2 & 1 \\ 2 & -10 & -6 & -12 & -4 & -10 \\ 1 & -5 & -3 & -8 & 2 & -5 \\ -1 & 1 & 1 & 2 & 0 & 5 \end{bmatrix}$$

to its Jordan Normal Form.

Solution. We will leave it to the readers to verify that the characteristic polynomial is $p(\lambda) = (4 - \lambda)^6$, so the only eigenvalue is $\lambda = 4$. There can be

many possible scenarios: the geometric multiplicity may be 6, so there will be 6 eigenvectors and 6 singleton Jordan blocks, which reduces to the usual diagonalization. The other extreme will have a single Jordan block of length 6. In between these two extremes, we may have 2 to 5 Jordan blocks of varying sizes. To begin, we compute the eigenvectors that are also the generalized eigenvectors of order 1, by solving the system $A - 4I = \mathbf{0}$:

$$\left[\begin{array}{cccccc|c} 1 & -1 & -1 & -2 & 0 & -1 & 0 \\ -1 & 15 & 9 & 24 & 6 & 17 & 0 \\ -1 & 3 & 1 & 4 & 2 & 1 & 0 \\ 2 & -10 & -6 & -16 & -4 & -10 & 0 \\ 1 & -5 & -3 & -8 & -2 & -5 & 0 \\ -1 & 1 & 1 & 2 & 0 & 1 & 0 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cccccc|c} 1 & -1 & -1 & -2 & 0 & -1 & 0 \\ 0 & 14 & 8 & 22 & 6 & 16 & 0 \\ 0 & 2 & 0 & 2 & 2 & 0 & 0 \\ 0 & -8 & -4 & -12 & -4 & -8 & 0 \\ 0 & -4 & -2 & -6 & -2 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{aligned} R_2 + R_1 &\rightarrow R_2 \\ R_3 + R_1 &\rightarrow R_3 \\ R_4 - 2R_1 &\rightarrow R_4 \\ R_5 - R_1 &\rightarrow R_5 \\ R_6 + R_1 &\rightarrow R_6 \end{aligned}$$

$$\rightarrow \left[\begin{array}{cccccc|c} 1 & -1 & -1 & -2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 14 & 8 & 22 & 6 & 16 & 0 \\ 0 & -8 & -4 & -12 & -4 & -8 & 0 \\ 0 & -4 & -2 & -6 & -2 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{aligned} R_2 &\leftrightarrow R_3 \\ \frac{1}{2}R_2 &\rightarrow R_2 \end{aligned}$$

$$\rightarrow \left[\begin{array}{cccccc|c} 1 & -1 & -1 & -2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 8 & 8 & -8 & 16 & 0 \\ 0 & 0 & -4 & -4 & 4 & -8 & 0 \\ 0 & 0 & -2 & -2 & 2 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{aligned} R_3 - 14R_2 &\rightarrow R_3 \\ R_4 + 8R_2 &\rightarrow R_4 \\ R_5 + 4R_2 &\rightarrow R_5 \end{aligned}$$

$$\rightarrow \left[\begin{array}{cccccc|c} 1 & -1 & -1 & -2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} \frac{1}{8}R_3 \rightarrow R_3 \\ R_4 + 4R_3 \rightarrow R_4 \\ R_5 + 2R_3 \rightarrow R_5 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_1 + R_3 \rightarrow R_1 \\ R_1 + R_2 \rightarrow R_1 \end{array}$$

The nullity is 3. We can assign the last three components as the free variables to obtain three eigenvectors: $(0, -1, -1, 1, 0, 0)^T$, $(0, -1, 1, 0, 1, 0)^T$, $(-1, 0, -2, 0, 0, 1)^T$ (the readers should do the checking), and there will be three Jordan blocks of unknown lengths. Now here is the trick. We create an augmented system where the left is $A - \lambda I$ and the right consists of the three eigenvectors we just obtained. We will derive its RREF but we can reuse the previous reduction step:

$$\left[\begin{array}{cccccc|ccc} 1 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 \\ -1 & 15 & 9 & 24 & 6 & 17 & -1 & -1 & 0 \\ -1 & 3 & 1 & 4 & 2 & 1 & -1 & 1 & -2 \\ 2 & -10 & -6 & -16 & -4 & -10 & 1 & 0 & 0 \\ 1 & -5 & -3 & -8 & -2 & -5 & 0 & 1 & 0 \\ -1 & 1 & 1 & 2 & 0 & 1 & 0 & 0 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 \\ 0 & 14 & 8 & 22 & 6 & 16 & -1 & -1 & -1 \\ 0 & 2 & 0 & 2 & 2 & 0 & -1 & 1 & -3 \\ 0 & -8 & -4 & -12 & -4 & -8 & 1 & 0 & 2 \\ 0 & -4 & -2 & -6 & -2 & -4 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_2 + R_1 \rightarrow R_2 \\ R_3 + R_1 \rightarrow R_3 \\ R_4 - 2R_1 \rightarrow R_4 \\ R_5 - R_1 \rightarrow R_5 \\ R_6 + R_1 \rightarrow R_6 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{3}{2} \\ 0 & 14 & 8 & 22 & 6 & 16 & -1 & -1 & -1 \\ 0 & -8 & -4 & -12 & -4 & -8 & 1 & 0 & 2 \\ 0 & -4 & -2 & -6 & -2 & -4 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_2 \leftrightarrow R_3 \\ \frac{1}{2}R_2 \rightarrow R_2 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 8 & 8 & -8 & 16 & 6 & -8 & 20 \\ 0 & 0 & -4 & -4 & 4 & -8 & -3 & 4 & -10 \\ 0 & 0 & -2 & -2 & 2 & -4 & -2 & 3 & -5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_3 - 14R_2 \rightarrow R_3 \\ R_4 + 8R_2 \rightarrow R_4 \\ R_5 + 4R_2 \rightarrow R_5 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & -1 & -1 & -2 & 0 & -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 1 & 1 & -1 & 2 & \frac{3}{4} & -1 & \frac{5}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} \frac{1}{8}R_3 \rightarrow R_3 \\ R_4 + 4R_3 \rightarrow R_4 \\ R_5 + 2R_3 \rightarrow R_5 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 1 & 1 & -1 & 2 & \frac{3}{4} & -1 & \frac{5}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_1 + R_3 \rightarrow R_1 \\ R_1 + R_2 \rightarrow R_1 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 1 & 1 & -1 & 2 & \frac{3}{4} & -1 & \frac{5}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{r} R_4 \leftrightarrow R_5 \\ -2R_4 \rightarrow R_4 \end{array}$$

$$\rightarrow \left[\begin{array}{cccccc|ccc} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & -\frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 1 & 1 & -1 & 2 & 0 & \frac{1}{2} & \frac{5}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{l} R_1 - \dfrac{1}{4}R_4 \rightarrow R_1 \\ R_2 + \dfrac{1}{2}R_4 \rightarrow R_1 \\ R_3 - \dfrac{3}{4}R_4 \rightarrow R_1 \end{array}$$

We only need to look at the top three rows where pivots exist in the left portion and discard the bottom three rows. The first column to the right is trivially all zeros, so it is a dead end. Meanwhile, by considering the second/third column to the right, we can simply take the particular solutions $(0, -\frac{1}{2}, \frac{1}{2}, 0, 0, 0)^T$ and $(0, -\frac{3}{2}, \frac{5}{2}, 0, 0, 0)^T$ as candidates for the generalized eigenvectors of order 2. Now there are 5 generalized eigenvectors, and the last generalized eigenvector will be found with an order of 3. We repeat the same procedure and the result is simply shown below:

$$\left[\begin{array}{cccccc|cc} 1 & 0 & 0 & 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 1 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

So the end vector of the Jordan block with length 3 is $(-\frac{1}{2}, \frac{1}{2}, -1, 0, 0, 0)^T$. There will be two other Jordan blocks with lengths 1 and 2 respectively. We will compute the cycle of generalized eigenvectors for all three Jordan blocks in the opposite direction. We begin from the largest Jordan block of length 3, where the end vector has been found to be $(-\frac{1}{2}, \frac{1}{2}, -1, 0, 0, 0)^T$, and then compute:

$$\begin{aligned} (A - 4I)\left(-\frac{1}{2}, \frac{1}{2}, -1, 0, 0, 0\right)^T &= (0, -1, 1, 0, 0, 0)^T \\ (A - 4I)(0, -1, 1, 0, 0, 0)^T &= (0, -6, -2, 4, 2, 0)^T \end{aligned}$$

For the second Jordan block of length 2, we note that at the third stage, the inconsistent 1 to the right appears in the second column, and hence we will choose the second candidate generalized eigenvector $(0, -\frac{3}{2}, \frac{5}{2}, 0, 0, 0)^T$ as the new end vector at the second stage, and calculate:

$$(A - 4I)\left(0, -\frac{3}{2}, \frac{5}{2}, 0, 0, 0\right)^T = (-1, 0, -2, 0, 0, 1)^T$$

By the same logic, we choose the first eigenvector $(0, -1, -1, 1, 0, 0)^T$ for the last singleton Jordan block. The final answer will then be

$$Q^{-1}AQ = \begin{bmatrix} 0 & 0 & -\frac{1}{2} & -1 & 0 & 0 \\ -6 & -1 & \frac{1}{2} & 0 & -\frac{3}{2} & -1 \\ -2 & 1 & -1 & -2 & \frac{5}{2} & -1 \\ 4 & 0 & 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 5 & -1 & -1 & -2 & 0 & -1 \\ -1 & 19 & 9 & 24 & 6 & 17 \\ -1 & 3 & 5 & 4 & 2 & 1 \\ 2 & -10 & -6 & -12 & -4 & -10 \\ 1 & -5 & -3 & -8 & 2 & -5 \\ -1 & 1 & 1 & 2 & 0 & 5 \end{bmatrix}$$

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

$$J = \begin{bmatrix} 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

□

B.3. Matrix Exponentials

In Section 9.3, we have seen how to solve a system of first-order ODEs $\mathbf{y}' = A\mathbf{y}$ by diagonalization: the new system $\mathbf{z}' = D\mathbf{z}$ where D is a diagonal matrix can be solved according to (9.17) so that $z_i = c_i e^{D_{ii}x}$. Since in the one-dimensional case, we simply have $z = ce^{\alpha x}$, we may like to ask if we can also write something along the lines of $\mathbf{z} = e^{Dx}\mathbf{c}$ (or even one step further, $\mathbf{y} = e^{Ax}\mathbf{c}$) without resorting to dealing with each variable separately. This motivates the definition of a ***matrix exponential***. Recall that from elementary Calculus, the usual exponential function can be expressed in terms of a Taylor series:

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

and thus in parallel, we demand that

Definition B.3.1. The exponential of a matrix X is computed by

$$e^X = I + X + \frac{X^2}{2!} + \cdots = \sum_{k=0}^{\infty} \frac{X^k}{k!} \tag{B.13}$$

When we set $X = Ax$ where A is a matrix and x is an independent variable, it becomes

$$e^{Ax} = I + Ax + \frac{A^2}{2!}x^2 + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!}x^k \tag{B.14}$$

Now, if we are given a diagonal matrix D such that

$$D = \begin{bmatrix} D_{11} & 0 & \cdots & 0 \\ 0 & D_{22} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & D_{nn} \end{bmatrix}$$

Then it is obvious that

$$D^k = \begin{bmatrix} D_{11}^k & 0 & \cdots & 0 \\ 0 & D_{22}^k & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & D_{nn}^k \end{bmatrix}$$

and thus

$$\begin{aligned} e^{Dx} &= I + Dx + \frac{D^2}{2!}x^2 + \cdots = \sum_{k=0}^{\infty} \frac{D^k}{k!}x^k \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{bmatrix} + \begin{bmatrix} D_{11}x & 0 & \cdots & 0 \\ 0 & D_{22}x & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & D_{nn}x \end{bmatrix} \\ &\quad + \begin{bmatrix} \frac{D_{11}^2}{2!}x^2 & 0 & \cdots & 0 \\ 0 & \frac{D_{22}^2}{2!}x^2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \frac{D_{nn}^2}{2!}x^2 \end{bmatrix} + \cdots \\ &= \begin{bmatrix} 1 + D_{11}x + \frac{D_{11}^2}{2!}x^2 + \cdots & 0 & \cdots & 0 \\ 0 & 1 + D_{22}x + \frac{D_{22}^2}{2!}x^2 + \cdots & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 + D_{nn}x + \frac{D_{nn}^2}{2!}x^2 + \cdots \end{bmatrix} \\ &= \begin{bmatrix} e^{D_{11}x} & 0 & \cdots & 0 \\ 0 & e^{D_{22}x} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & e^{D_{nn}x} \end{bmatrix} \end{aligned}$$

so the proposed solution form $\mathbf{z} = e^{Dx}\mathbf{c}$ will be consistent with what we have established in Section 9.3.

Since diagonalization is not always possible but we can compensate with a Jordan Normal Form, the next step is to derive e^{Jx} where J is a Jordan block with an eigenvalue of λ, and we will start from the easier case of length 3:

$$J = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

then

$$J^2 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} \lambda^2 & 2\lambda & 1 \\ 0 & \lambda^2 & 2\lambda \\ 0 & 0 & \lambda^2 \end{bmatrix}$$

$$J^3 = \begin{bmatrix} \lambda^2 & 2\lambda & 1 \\ 0 & \lambda^2 & 2\lambda \\ 0 & 0 & \lambda^2 \end{bmatrix} \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} \lambda^3 & 3\lambda^2 & 3\lambda \\ 0 & \lambda^3 & 3\lambda^2 \\ 0 & 0 & \lambda^3 \end{bmatrix}$$

Going on, we have

$$J^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} & \frac{k(k-1)}{2}\lambda^{k-2} \\ 0 & \lambda^k & k\lambda^{k-1} \\ 0 & 0 & \lambda^k \end{bmatrix}$$

and thus

$$\begin{aligned} e^{Jx} &= I + Jx + \frac{J^2}{2!}x^2 + \frac{J^3}{3!}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{J^k}{k!}x^k \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} \lambda x & x & 0 \\ 0 & \lambda x & x \\ 0 & 0 & \lambda x \end{bmatrix} + \frac{1}{2!} \begin{bmatrix} \lambda^2 x^2 & 2\lambda x^2 & x^2 \\ 0 & \lambda^2 x^2 & 2\lambda x^2 \\ 0 & 0 & \lambda^2 x^2 \end{bmatrix} \\ &\quad + \frac{1}{3!} \begin{bmatrix} \lambda^3 x^3 & 3\lambda^2 x^3 & 3\lambda x^3 \\ 0 & \lambda^3 x^3 & 3\lambda^2 x^3 \\ 0 & 0 & \lambda^3 x^3 \end{bmatrix} + \cdots \end{aligned}$$

$$= \begin{bmatrix} 1+\lambda x+\frac{\lambda^2}{2!}x^2+\frac{\lambda^3}{3!}x^3+\cdots & x(1+\lambda x+\frac{\lambda^2}{2!}x+\cdots) & \frac{x^2}{2!}(1+\lambda x+\cdots) \\ 0 & 1+\lambda x+\frac{\lambda^2}{2!}x^2+\frac{\lambda^3}{3!}x^3+\cdots & x(1+\lambda x+\frac{\lambda^2}{2!}x+\cdots) \\ 0 & 0 & 1+\lambda x+\frac{\lambda^2}{2!}x^2+\frac{\lambda^3}{3!}x^3+\cdots \end{bmatrix}$$

$$= \begin{bmatrix} e^{\lambda x} & xe^{\lambda x} & \frac{x^2}{2!}e^{\lambda x} \\ 0 & e^{\lambda x} & xe^{\lambda x} \\ 0 & 0 & e^{\lambda x} \end{bmatrix}$$

In general, we have the following formula.

Properties B.3.2. The matrix exponential e^{Jx} where J is a Jordan block of size n corresponding to an eigenvalue of λ is

$$e^{Jx} = \begin{bmatrix} e^{\lambda x} & xe^{\lambda x} & \frac{x^2}{2!}e^{\lambda x} & \cdots & \frac{x^{n-1}}{(n-1)!}e^{\lambda x} \\ 0 & e^{\lambda x} & xe^{\lambda x} & & \frac{x^{n-2}}{(n-2)!}e^{\lambda x} \\ 0 & 0 & e^{\lambda x} & & \frac{x^{n-3}}{(n-3)!}e^{\lambda x} \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda x} \end{bmatrix} \tag{B.15}$$

Given a Jordan Normal Form $J = J_1 \oplus J_2 \oplus \cdots$ as a direct sum composed of smaller Jordan blocks, its matrix exponential e^{Jx} will also be a direct sum of the respective smaller matrix exponential $e^{J_i x}$, i.e. $e^{Jx} = e^{J_1 x} \oplus e^{J_2 x} \oplus \cdots$. For example, if $J = J_1 \oplus J_2$ with

$$J_1 = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \qquad \text{and} \qquad J_2 = \begin{bmatrix} 5 & 1 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

then

$$e^{Jx} = \begin{bmatrix} e^{2x} & xe^{2x} & 0 & 0 & 0 & 0 \\ 0 & e^{2x} & 0 & 0 & 0 & 0 \\ 0 & 0 & e^{5x} & xe^{5x} & \frac{x^2}{2}e^{5x} & \frac{x^3}{6}e^{5x} \\ 0 & 0 & 0 & e^{5x} & xe^{5x} & \frac{x^2}{2}e^{5x} \\ 0 & 0 & 0 & 0 & e^{5x} & xe^{5x} \\ 0 & 0 & 0 & 0 & 0 & e^{5x} \end{bmatrix}$$

Extending the above results, if we have a square matrix A that has a Jordan Normal Form of J such that $J = Q^{-1}AQ$, where Q forms its Jordan canonical basis, then

$$\begin{aligned} e^{Ax} &= I + Ax + \frac{A^2}{2!}x^2 + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!}x^k \\ &= QIQ^{-1} + QJQ^{-1}x + \frac{1}{2!}(QJQ^{-1})(QJQ^{-1})x^2 + \cdots \\ &= QIQ^{-1} + QJQ^{-1}x + \frac{1}{2!}QJ^2Q^{-1}x^2 + \cdots \\ &= Q(I + Jx + \frac{J^2}{2!}x^2 + \cdots)Q^{-1} \\ &= Qe^{Jx}Q^{-1} \end{aligned} \tag{B.16}$$

Now let's apply these to solve Example 9.4.1 when $k_1 = k_2 = k$ and diagonalization is no longer possible. The system becomes

$$\mathbf{y}' = A\mathbf{y} = \begin{bmatrix} -k & 0 & 0 \\ k & -k & 0 \\ 0 & k & 0 \end{bmatrix} \mathbf{y}$$

It remains the same that the eigenvector for $\lambda = 0$ is $(0, 0, 1)^T$. For $\lambda = -k$, it will form a Jordan block of length 2: the initial eigenvector can be seen to be $(0, -1, 1)^T$ and the end generalized eigenvector is found by solving

$$\left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ k & 0 & 0 & -1 \\ 0 & k & k & 1 \end{array}\right]$$

where one of the possible choices is $(-\frac{1}{k}, 0, \frac{1}{k})^T$. The solution of the system is then

$$\mathbf{y} = e^{At}\mathbf{c} = Qe^{Jt}Q^{-1}\mathbf{c}$$

with

$$Q = \begin{bmatrix} 0 & 0 & -\frac{1}{k} \\ 0 & -1 & 0 \\ 1 & 1 & \frac{1}{k} \end{bmatrix} \qquad J = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -k & 1 \\ 0 & 0 & -k \end{bmatrix}$$

and

$$e^{Jt} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-kt} & te^{-kt} \\ 0 & 0 & e^{-kt} \end{bmatrix}$$

Therefore,

$$\begin{aligned}
\mathbf{y} &= Qe^{Jt}Q^{-1}\mathbf{c} \\
&= \begin{bmatrix} 0 & 0 & -\frac{1}{k} \\ 0 & -1 & 0 \\ 1 & 1 & \frac{1}{k} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-kt} & te^{-kt} \\ 0 & 0 & e^{-kt} \end{bmatrix} \begin{bmatrix} 0 & 0 & -\frac{1}{k} \\ 0 & -1 & 0 \\ 1 & 1 & \frac{1}{k} \end{bmatrix}^{-1} \mathbf{c} \\
&= \begin{bmatrix} 0 & 0 & -\frac{1}{k} \\ 0 & -1 & 0 \\ 1 & 1 & \frac{1}{k} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-kt} & te^{-kt} \\ 0 & 0 & e^{-kt} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ -k & 0 & 0 \end{bmatrix} \mathbf{c} \\
&= \begin{bmatrix} e^{-kt} & 0 & 0 \\ ke^{-kt}t & e^{-kt} & 0 \\ -ke^{-kt}t - e^{-kt} + 1 & -e^{-kt} + 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}
\end{aligned}$$

Plugging in the initial condition gives $c_1 = 1, c_2 = 0, c_3 = 0$, and thus the new solution will be

$$\begin{cases} [P] &= e^{-kt} \\ [Q] &= ke^{-kt}t \\ [R] &= 1 - ke^{-kt}t - e^{-kt} \end{cases}$$

Index

R-squared, 550
l^2-norm, 133
n-volume, 171

Absorbing Markov Chain, 620
Active Transformation, 274
Adiabatic Lapse Rate, 53
Adjoint, 492
Adjugate, 85
Advective Derivative, 755
Algebraic Multiplicity, 339
Amplitude, 584
Analysis Field, 769
Argand Plane, 297
Argument, 298
Asymptotically Stable, 694
Attractive, 694
Augmented Matrix, 39
Automorphism, 278

Backward Phase (of Gaussian Elimination), 66
Baroclinicity, 767
Bessel's Inequality, 561
Best Approximation Theorem, 560
Bijective, 266
Bilinear Form, 436
Block Matrix, 310
Boundary Terms, 494
Buckingham Pi Theorem, 702
Butterfly Effect, 714

Canonical Quadratic Form, 442
Cartesian Coordinate System, 129
Cartesian Tensor, 730
Cauchy Stress Tensor, 742
Cauchy's Formula, 742
Cauchy-Schwarz Inequality, 142
Cayley-Hamilton Theorem, 348
Center, 686
Central Conic, 447
Chaos, 714
Characteristic Equation/Polynomial, 339
Cholesky Decomposition/Factorization, 634

Circular Convolution, 591
Circulation, 765
Clairaut's Theorem, 758
Closure (under an Operation), 186
Co-planar, 170
Codomain, 256
Coefficient Matrix, 38
Coefficient of Determination, 550
Cofactor, 77
Cofactor Expansion, 78
Column Space, 197
Column Vector, 128
Column-row (CR) Factorization, 204
Commutative, 33
Compact SVD, 657
Companion Matrix, 877
Complement (Subspace), 222
Complementary Solution, 103
Complete, 489
Complex n-space, 309
Complex Conjugate, 294
Complex Dot Product, 301
Complex Number, 294
Complex Plane, 297
Components, 128
Conditional Probability, 610
Conformable, 32
Congruent, 440
Conic Section, 445
Conjugate Transpose, 304
Conservation of Mass, 757
Consistent, 96
Constrained Extremum Theorem, 469
Continuity Equation, 760
Contraction, 722
Control Volume, 762
Convolution, 588
Convolution Theorem (Circular), 592
Coordinate Basis (System), 212
Coordinate Derivative, 737
Coordinates, 212
Coriolis Effect/Force, 149
Coriolis Parameter, 150
Correlation, 460
Cosine Law, 143
Cosine Law for Spherical Trigonometry, 174
Covariance, 458
Covariance Matrix, 462
Cramer's Rule, 111
Cross Product, 144
Curl, 739
Cycle of Generalized Eigenvectors, 880
Cyclic Subspace, 875

Data Assimilation, 767
De Moivre's Formula, 301
Deficient (Eigenvalue), 401
Definiteness, 437

Dependence Relation, 201
Determinant, 75
Diagonal Matrix, 350
Diagonalization, 350
Dimension, 220
Dimensionality Reduction, 464
Direct Sum, 221
Direction, 134
Discrete Fourier Transform (DFT, complex), 577
Discrete Fourier Transform (DFT, real), 573
Displacement Field, 752
Divergence, 738
Divergence Theorem, 741
Domain, 256
Dot Product, 135
Dummy Indices, 722

Eigenspace, 337
Eigenvalue, 336
Eigenvector, 336
Einstein Convention, 722
El Niño-Southern Oscillation (ENSO), 479
Elementary Matrix, 72
Elementary Row Operations, 39
Emission Temperature, 117
Empirical Orthogonal Functions (EOFs), 465
Endomorphism, 278
Epsilon Symbol, 726
Epsilon-delta Identity, 728
Equation(s) of Motion, 764
Equilibrium Point, 684
Equivalence Statements, 88
Error Covariance Matrix, 773
Euclidean n-space, 485
Euclidean Norm, 133
Euler's Equations, 764
Euler's Formula, 298
Eulerian Description, 751
Expected Value, 455
Explained Variance, 469
Explanatory Variable, 542

Fast Fourier Transform (FFT), 597
Field Equations, 756
Field Lines, 685
Finite-dimensional, 220
First-order Linear ODE, 366
Forward Phase (of Gaussian Elimination), 66
Fourier Basis, 501
Fourier Coefficients, 513
Fourier Series, 513
Free Indices, 722
Free Variable/Parameter, 102
Frobenius Inner Product, 531
Full-rank, 237
Fundamental Frequency, 597
Fundamental Theorem of Algebra, 351

Gauss-Jordan Elimination, 66
Gaussian Elimination, 66
General Solution, 103
Generalized Eigenspace, 880
Generalized Eigenvector, 879
Generating Set, 197
Geomagnetic Dipole Model, 178
Geometric Multiplicity, 339
Geometric Sum (Matrix), 621
Gradient, 738
Gram-Schmidt Orthogonalization, 280

Hartman-Grobman Theorem, 698
Haversine (Formula), 177
Hermite polynomials, 521
Hermite's Equation, 519
Hermitian (Operator), 496
Hermitian Form, 453
Hermitian Matrix, 304
Hermitian Transpose, 304
Hilbert Space, 489
Homogeneous Linear Equation, 36
Homogeneous Linear System of Equations, 37
Hydrostatic Pressure, 764
Hyperbolic Equilibrium Point, 697
Hyperplane, 163

Idempotent, 406
Identity Matrix, 55
Identity Transformation/Mapping, 265
Image, 256
Imaginary Axis, 297
Imaginary Number, 294
Imaginary Part, 294
Improper Subspace, 190
Inconsistent, 96
Index Notation, 722
Inertial Oscillation, 380
Infinite-dimensional, 220
Injective, 256
Inner Product, 484
Inner Product Space, 484
Innovation, 774
Integration by Parts, 494
Invariant (Matrix), 357
Invariant Subspaces, 347
Inverse, 60
Invertible, 61
Inviscid, 764
Isomorphic, 266
Isomorphism, 271
Isotropic, 734

Jacobi Identity, 173
Jacobian (Deformation), 753
Jacobian (Matrix), 696
Jordan Block, 882
Jordan Canonical Basis, 882
Jordan Normal Form, 883

Kalman Filter, 780
Kalman Gain, 774
Kelvin's Circulation Theorem, 765
Kernel, 261
Kronecker Delta, 724

Lagged (Lag-k) Auto-correlation, 608
Lagrange Multiplier, 465
Lagrange's Identity, 174
Lagrangian Description, 751
Laplace Expansion, 78
Laplacian, 705
Law of Total Probability, 611
LDU Decomposition/Factorization, 640
Least-square Approximation, 536
Left Null Space, 240
Left Singular Vectors, 650
Legendre Polynomials, 533
Legendre's Equation, 533
Length, 133
Levi-Civita Symbol, 726
Linear Combination, 191
Linear Equation, 36
Linear Independence, 198
Linear Operator, 278
Linear Regression, 542
Linear System of Equations, 36
Linear Transformation/Mapping, 248
Linearity, 248
Linearization, 697
Linearization Theorem, 698
Linearized Shallow Water System, 525
Linearly Dependent, 198
Linearly Independent, 198
Lorenz-63 Model, 700
LU Decomposition/Factorization, 637

Magnitude, 133
Main Diagonal, 56
Mapping, 248
Markov Chain, 613
Material Derivative, 755
Material Description, 751
Material Displacement Gradient, 752
Material Volume, 757
Matrix, 28
Matrix Direct Sum, 325
Matrix Exponential, 892
Matrix Multiplication/Matrix Product, 30
Matrix Representation of a Linear Transformation, 249
Mean, 455
Memory, 607
Memoryless, 615
Minimal Generating Set, 204
Minimal Polynomial, 884

Minimal Solution, 665
Minor, 77
Modulus, 298
Mohr's Circle, 746
Moore-Penrose Pseudoinverse, 663
Multi-layer Model, 43

Natural Interval, 518
Negative-Definite, 437
Nilpotent, 882
Norm, 486
Normal, 418
Normal Equation, 536
Normal Vector, 160
Null Space, 235
Nullity, 235
Nyquist Frequency, 572
Nyquist Sampling Theorem, 572

One-to-one, 256
Onto, 256
Optimal Interpolation, 776
Orbit, 684
Ordered Tuple, 128
Ordinary Differential Equation (ODE), 366
Orthogonal, 142
Orthogonal Complement, 240
Orthogonal Diagonalization, 398
Orthogonal Matrix, 388
Orthogonal Projection, 407
Orthonormal, 282
Orthonormal Matrix, 388
Overdetermined, 104

Parallelogram Law, 155
Parameterization, 159
Parseval's Identity, 561
Parseval's Theorem, 585
Partial Differential Equation (PDE), 515
Particle, 684
Particular Solution, 103
Passive Transformation, 275
Permutation Matrix, 73
Permutation Symbol, 726
Phase, 584
Pivot/Pivoting, 66
Pivotal Column, 67
PLU Decomposition/Factorization, 641
Plus/Minus Theorem, 203
Population Covariance, 458
Population Variance, 455
Positive Stochastic Matrix, 618
Positive-Definite, 437
Power, 584
Power Spectrum, 585
Precision, 770
Preimage, 256
Principal Axes Theorem, 452
Principal Axes/Directions, 468
Principal Axes/Directions (Stress

Tensor), 744
Principal Component Analysis (PCA), 465
Principal Components (PCs), 469
Principal Stresses, 744
Projection (onto a Subspace), 405
Proper Subspace, 190
Pseudoinverse, 662
Pythagoras' Theorem, 133

QR Decomposition, 284
Quadratic Form, 433
Quotient Law, 733

Radix-2 Algorithm, 597
Random Variable, 455
Range, 261
Rank, 231
Rank (of a Tensor), 730
Rank-nullity Theorem, 236
Rank-nullity Theorem (for a Linear Transformation), 263
Rayleigh-Barnard Convection, 701
Real n-space, 128
Real Axis, 297
Real Dot Product, 135
Real Part, 294
Real Vector Space, 186
Reduced Row Echelon Form (RREF), 64
Regular Markov Chain, 618
Residual, 548
Resolution of the Identity, 417
Response Variable, 542
Restriction of a Linear Transformation, 320
Reynold's Transport Theorem, 761
Right Singular Vectors, 650
Right-hand Rule, 144
Row Echelon Form, 64
Row Equivalent, 69
Row Space, 230
Row Vector, 128

Saddle Point, 686
Sample Covariance, 458
Sample Mean, 458
Sarrus' Rule, 75
Scalar, 186
Scalar Multiplication, 131
Scalar Multiplication (for Matrices), 29
Scalar Product, 135
Scalar Projection, 164
Scalar Triple Product, 168
Schuader Basis, 490
Schur Complement, 316
Schur's Triangularization Theorem, 419
Self-adjoint, 496
Semidefiniteness, 437
Separable (Hilbert Space), 489

Separation of Variables, 515
Series Solution, 519
Shallow Water Equations, 124
Sherman-Morrison-Woodbury Formula, 779
Signature, 444
Similar, 279
Sine Law, 182
Singular, 61
Singular Value, 647
Singular Value Decomposition (SVD), 646
Sink, 686
Skew-symmetric Matrix, 59
Snell's Law, 41
Source, 686
Span, 195
Spanning Set, 197
Spatial Derivative, 755
Spatial Description, 751
Spatial Displacement Gradient, 752
Special Polynomials, 522
Spectral Decomposition, 417
Spectral Theorem, 415
Spectrum, 417
Split, 351
Square Matrix, 28
Stable, 694
Stable Node, 686
Stable Spiral, 686
Standard Basis, 212
Standard Inner Product, 485
Standard Position, 448
Standard Unit Vector, 129
State Vector, 613
Static Equilibrium, 742
Stationary, 613
Steady-state Vector, 615
Stefan-Boltzmann Law, 43
Steinitz Replacement Theorem, 865
Stochastic Matrix, 612
Stokes' Theorem, 741
Strange Attractor, 714
Stream Function, 704
Sturm-Liouville Equation/Operator, 515
Submatrices, 310
Subspace, 187
Subspace Sum, 221
Sum of Squares Error (SSE), 548
Sum of Squares Regression (SSR), 548
Sum of Squares Total (SST), 548
Summation Indices, 722
Surjective, 256
Sylvester's Law of Inertia, 444
Symmetric Bilinear Form, 436
Symmetric Matrix, 59
System of ODEs, 368

Tangent Linear Model, 785
Tensor Divergence Theorem, 740

Tensors, 730
The Four Fundamental Subspaces, 240
Time Series, 607
Transformation, 248
Transformation Law, 731
Transition Matrix, 393
Transpose, 57
Triangle Inequality, 155
Triangle Inequality (for a General Inner Product), 532
Truncated SVD, 657

Underdetermined, 105
Unit Vector, 133
Unitary Diagonalization, 418
Unitary Matrix, 417
Unstable, 694
Unstable Node, 686
Unstable Spiral, 686

Vandermonde Matrix, 548
Variance, 455
Vector, 186
Vector (Geometric), 127
Vector Product, 144
Vector Projection, 164
Vector Space, 186
Vector Triple Product, 172

Weak Inverse, 669

Zero Subspace, 190

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] J. Hateley. *The Lorenz System*. URL: https://web.math.ucsb.edu/~jhateley/paper/lorenz.pdf.

[4] George N. Kiladis et al. "Convectively coupled equatorial waves". In: *Reviews of Geophysics* 47.2 (2009). DOI: https://doi.org/10.1029/2008RG000266. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008RG000266. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008RG000266.

[5] G.-Y. Lian. *AtmSci7056 - Data Assimilation for Numerical Modeling (Lecture Materials)*.

[6] K. Matthews. *Markov Matrices*. URL: http://www.numbertheory.org/courses/MP274/markov.pdf.

[7] 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.

[8] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2022. ISBN: 9781733146678.

[9] K.-C. Tseng. *Chaos and Predictability – Week 3: The minimalist models for studying chaos and predictability I: Lorenz 63*. URL: https://kuiper2000.github.io/chaos_and_predictability/week3/week3.html.

[10] T.F. Wong. *ESSC3010 - Continuum Mechanics (Lecture Materials)*.