Foundations of Linear Algebra for Data Science

Wanmo Kang
KAIST
wanmo.kang@kaist.ac.kr

Kyunghyun Cho
New York University
Genentech
kyunghyun.cho@nyu.edu

July 15, 2025

Contents

Preface

TODO

Chapter 1

Introduction

In mathematics, most areas deal with various types of spaces and study functions defined between these spaces as well as how these functions preserve or do not preserve various properties of these spaces. Linear algebra in particular deals with spaces in which objects can be scaled by a scalar and be added together and studies functions that preserve such properties. We refer to these objects, spaces and functions as vectors, vector spaces and linear transformations, respectively. We use linearity to collectively refer to the scalability and additivity of vectors that are preserved under linear transformation across vector spaces. This linearity enables us to understand the universal structures of the vector spaces and linear transformation. For example, vector spaces of the same dimension are roughly equivalent. We also found that a linear transformation can be characterized by a rectangular array of numbers called a matrix. Then, to investigate and classify linear transformations, we work with matrices corresponding to linear transformations, and this allows us to successfully classify these linear transformations into Jordan forms. Such fundamental efforts for classification have led to a variety of by-products that have proven to be useful for many real-world applications. In this book, we do not make compromise between the fundamental results and useful by-products, by providing readers with gap-free derivations of useful by-products from the fundamental results.

Let us take a look in detail. Two most important objects in describing and studying linear algebra are vectors and matrices. We may refer to vectors as any mathematical objects for which vector addition and scalar multiplication can be well-introduced. For the vector addition, we denote the identity which is usually called a zero vector by $\mathbf{0}$. We also place a minus sign (-) in front of the

original vector to denote the inverse of the addition. A vector space is defined as the collection of vectors that satisfy the distributive laws between the vector addition and the scalar multiplication. For example, the distributive laws with a notational convention of $1\mathbf{v} = \mathbf{v}$ enable, for any vector $\mathbf{v}$ in a vector space,

$$\mathbf{v} + \mathbf{v} = 1\mathbf{v} + 1\mathbf{v} = (1 + 1)\mathbf{v} = 2\mathbf{v}.$$

We will dive deeper into the vector space in Chapter 3.

One of the most intuitive yet important examples of a vector is a finite array of numbers. We can vertically stack m real values $v_1, v_2, \ldots, v_m$ to form an m-dimensional vector $\mathbf{v}$;

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$

where we use v_i to refer to the i-th element of $\mathbf{v}$. We express that a vector $\mathbf{v}$ is an m-dimensional vector by $\mathbf{v} \in \mathbb{R}^m$ and often refer to it as either an m-vector or $\mathbb{R}^m$-vector. For $\mathbb{R}^m$-vectors, we define the vector addition and scalar multiplication by

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_m + w_m \end{bmatrix}, \quad c\mathbf{v} = \begin{bmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_m \end{bmatrix}$$

where $\mathbf{w} \in \mathbb{R}^m$ and $c \in \mathbb{R}$. For the vector addition, the zero vector $\mathbf{0} \in \mathbb{R}^m$, whose entries are all 0, serves as the additive identity. When the dimensionality matters, we use a subscript to emphasize it, such as in $\mathbf{0}_m$ for the m-dimensional zero vector.

In this book, we define a matrix as a rectangular array of numbers by horizontally concatenating $\mathbb{R}^m$-vectors. Given n $\mathbb{R}^m$-vectors

$$\mathbf{a}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}, \mathbf{a}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix}, \cdots, \mathbf{a}_n = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix},$$

we horizontally concatenate them to obtain a matrix A;

$$A = [\,\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\,] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

We call the rows and columns of a matrix row vectors and column vectors, respectively. When we are given the name of a matrix, such as A in this case, we use $(A)_{ij}$ or a_{ij} to denote the entry in the i-th row and j-column. We say that A is $m \times n$ matrix if the matrix A has m rows and n columns. We add two matrices of the same size, A and B, by adding each pair of corresponding components from these two matrices. That is,

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

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

We define scalar multiplication by multiplying each entry with a scalar;

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

We can regard any $\mathbb{R}^m$-vector as a matrix of m rows and 1 column. We can similarly consider each row of an $m \times n$ matrix as a $1 \times n$ matrix.

We often add various structures and operations to a vector space to use it in practice. For instance, it is natural to equip a vector space of matrices with matrix multiplication. We will delve deeper into matrix multiplication in Chapter 2, and here, we consider a simple case of multiplying an $m \times n$ matrix to an $n \times 1$ matrix, which is equivalent to an $\mathbb{R}^n$-vector. Such multiplication is defined as

$$A\mathbf{v} = v_1\mathbf{a}_1 + \cdots + v_n\mathbf{a}_n = \sum_{j=1}^{n} v_j\mathbf{a}_j. \tag{1.1}$$

This results in an $\mathbb{R}^m$-vector, i.e., $A\mathbf{v} \in \mathbb{R}^m$. This vector is a linear combination of the column vectors of the matrix A, and v_i's work as the weights/coefficients for this combination.

It is useful to use matrices and $\mathbb{R}^m$-vectors to model data. For instance, we can represent a group of n people by horizontally stacking $\mathbb{R}^m$-vectors corresponding to their characteristics to form an $m \times n$ matrix, A. In this matrix, $(A)_{ij}$ represents the i-th person's j-th property. We can compute the average of each of these properties as matrix-vector multiplication:

$$\frac{1}{n}\mathbf{a}_1 + \cdots + \frac{1}{n}\mathbf{a}_n = \frac{1}{n}\big(\mathbf{a}_1 + \cdots + \mathbf{a}_n\big) = \frac{1}{n}(A\mathbf{1}) = A\Big(\frac{1}{n}\mathbf{1}\Big) \quad \text{where } \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n.$$

As another example, consider the following model of economy.

- a_{ij}: the contribution of material j to the i-th product, $\mathbf{a}_j$: an $\mathbb{R}^m$-vector whose i-th entry is a_{ij};
- x_j: the amount of material j available, $\mathbf{x}$: an $\mathbb{R}^n$-vector whose j-th entry is x_j.

Under this model, we can interpret the linear combination of $\mathbf{a}_j$'s with coefficients of x_j's, i.e., $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \sum_{j=1}^n x_j\mathbf{a}_j$, as the amounts of m manufactured products produced from the given amounts of n raw materials. We can write this more concisely as $A\mathbf{x}$, with $(A)_{ij} = a_{ij}$. Given a target production vector $\mathbf{b}$, we can now write the problem of finding the amount of enough raw materials to satisfy the target production quantities, as solving for

$$A\mathbf{x} = \mathbf{b}.$$

These examples illustrate how simple it is to use matrices and vectors to reason about data.

We can naturally and seamlessly introduce and derive various concepts and results in linear algebra by solving $A\mathbf{x} = \mathbf{b}$ above. More specifically, we will learn the following concepts and results in this book:

- How to solve linear systems: Gaussian elimination;
- Abstraction and manipulation of data: a vector space and linear transformation;

- Approximation of data: orthogonality, projection and least squares;
- Factorization of data: SVD (singular value decomposition), PCA (principal components analysis) and pseudoinverse;
- Shapes of data: covariance, positive definiteness and convexity.
- Key features of matrices: determinant, eigenvalue and eigenvector;
- Advanced results for matrices: adjoint, positive definite cone and Perron-Frobenius theorem;
- Theorems by Cayley-Hamilton and Jordan.

In data science, it is usual to analyze complex data by projecting their high-dimensional vector representations in a lower-dimensional subspace and investigating the corresponding lower-dimensional vectors. SVD is one of the most representative approaches to determining the best subspace for approximating high-dimensional data. Additionally, positive definite matrices and their properties are frequently used to characterize the relationship within data, both in data science and engineering. In this book, we make a significant departure from existing textbooks and lecture notes in linear algebra and go directly into the concept of projection, SVD and positive definiteness without introducing eigenvalues nor eigenvectors in detail.

We support readers by providing them with real-world implementations of concepts and algorithms investigated in this book. These implementations use standard and modern tools, such as `numpy`, `scipy`, `sympy` and `scikit-learn`, and are openly available at

`https://github.com/kyunghyuncho/Foundations_of_LADS`

Chapter 2

Matrices and Gaussian Elimination

We say that we solve multiple linear equations, when we determine the values of unknown variables that satisfy multiple linear equations simultaneously. We can do so by progressively eliminating unknown variables from these equations by reading off the values of these unknown variables. This process, to which we refer as Gaussian elimination, progressively modifies linear equations without altering the solution of the original linear equations. By using matrices, we can describe this process of successive elimination of variables from linear equations without referring to variables, signs nor equalities. In other words, such Gaussian elimination can be described as a sequence of matrix operations. In addition to scalar multiplication and addition, we define matrix-matrix multiplication, in order to represent any modification of linear equations by Gaussian elimination as multiplying a matrix representing the linear equations with a specially-structured matrix. In doing so, we obtain a surprisingly rich set of concepts and mathematical results on matrices.

More specifically, we introduce matrix-matrix multiplication in this chapter. When we multiply a matrix with a vector from left, we get the linear combination of the columns of the matrix. When we multiple two matrices, then, the resulting matrix consists of columns resulting from multiplying the first matrix with the columns of the second matrix, respectively. The inverse of matrix-matrix multiplication is called an inverse matrix, and not every matrix has its inverse. We define the transpose of a matrix by swapping the column and row

indices and introduce a symmetric matrix as a matrix whose transpose is itself. Symmetric matrices show up in many places throughout this chapter and the rest of the book, as they exhibit mathematically favorable properties. We define lower and upper triangular matrices, as matrices whose entries above and below the diagonal are zeros, respectively. With these various types of matrices, we show that Gaussian elimination corresponds to multiplying a series of lower triangular matrices to the matrix of linear equations to arrive at an upper triangular matrix. We call this process LU factorization, and this connects to the process of inverting the matrix of the linear equations. We eventually describe this whole process in terms of block matrices.

2.1 Matrix Operations

We can extend matrix-vector multiplication (1.1) to matrix-matrix multiplication. Instead of $\mathbf{v} \in \mathbb{R}^n$, consider an $n \times \ell$ matrix $B = [\,\mathbf{b}_1|\mathbf{b}_2|\cdots|\,\mathbf{b}_\ell\,] = (b_{jk})$. Matrix-vector multiplication between A and the k-th column of B, $\mathbf{b}_k$, is then

$$A\mathbf{b}_k = b_{1k}\mathbf{a}_1 + b_{2k}\mathbf{a}_2 + \cdots + b_{nk}\mathbf{a}_n.$$

We define matrix-matrix multiplication of A and B by horizontally stacking the resulting vectors;

$$AB = [\,A\mathbf{b}_1|A\mathbf{b}_2|\cdots|\,A\mathbf{b}_\ell\,].$$

Matrix-matrix multiplication is well defined only when the number of the rows of the first matrix and the number of the columns of the second matrix coincide with each other. In other words, multiplying an $m \times n$ matrix and an $n \times \ell$ matrix results in an $m \times \ell$ matrix. We can express $(AB)_{ij}$ in multiple ways:

$$\begin{aligned}(AB)_{ij} &= b_{1j}a_{i1} + \cdots + b_{nj}a_{in} \\ &= \sum_{k=1}^{n} a_{ik}b_{kj} \\ &= \begin{bmatrix} a_{i1} & \cdots & a_{in} \end{bmatrix} \begin{bmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{bmatrix}.\end{aligned} \tag{2.1}$$

Instead of constructing AB by considering the columns of AB, we can instead focus on the rows of AB. The i-th row of AB can be expressed as

$$\Big[\sum_{k=1}^{n} a_{ik}b_{k1} \;\cdots\; \sum_{k=1}^{n} a_{ik}b_{k\ell}\Big] \;=\; \sum_{k=1}^{n} [a_{ik}b_{k1} \;\cdots\; a_{ik}b_{k\ell}]$$

$$= \sum_{k=1}^{n} a_{ik}[b_{k1} \cdots b_{k\ell}]$$
$$= a_{i1}\mathbf{b}'_1 + a_{i2}\mathbf{b}'_2 + \cdots + a_{in}\mathbf{b}'_n, \qquad (2.2)$$

where $\mathbf{b}'_k$ is the k-th row (a $1 \times \ell$ matrix) of B. In other words, the i-th row of AB is a weighted sum of the rows of B with the entries of the i-th row of A as their weights.

Matrix-matrix multiplication is associative, i.e. $(AB)C = A(BC)$. Matrix-matrix addition and multiplication satisfy distributivity, i.e. $A(B+C) = AB + AC$ and $(B+C)D = BC + CD$. Unlike the product of real numbers, however, matrix-matrix multiplication does not exhibit the commutative property. It is easy to find two matrices, E and F, such that $EF \neq FE$.

The identity for matrix addition is a matrix of all zeros, and we use $\mathbf{0}$ to denote it. Although it is often unnecessary to specify the size of such an all-zero matrix, if necessary, we use the subscript, i.e. $\mathbf{0}_{m,n}$ for an $m \times n$ matrix of all zeros. The identity for matrix multiplication is a square matrix, which has the same number of rows and columns, whose diagonal entries a_{ii} are 1 and off-diagonal ones are all zeros. We call this an identity matrix and use the following notation:

$$I = I_n = \begin{bmatrix} 1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1 \end{bmatrix}.$$

When necessary, we use a subscript to indicate the size of the identity matrix, as in

$$AI_n = A = I_m A,$$

where A is an $m \times n$ matrix. Another helpful, special matrix is a diagonal matrix whose off-diagonal entries are all zeros;

$$\begin{bmatrix} d_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & d_n \end{bmatrix} = \text{diag}(d_1, \ldots, d_n).$$

Diagonal entries may also be zero. When a matrix is multiplied by a diagonal matrix from right, its columns are scaled by the corresponding diagonal entries. When multiplied from left, the columns are scaled instead due to (2.2).

It is sometimes useful to transpose a matrix, which is defined as

Definition 2.1 *An $n \times m$ matrix $A^\top$ is the transpose of an $m \times n$ matrix A if $(A^\top)_{ij} = (A)_{ji}$ for all i and j.*

A simple example is

$$\begin{bmatrix} 2 & 1 & 4 \\ 0 & 0 & 3 \end{bmatrix}^\top = \begin{bmatrix} 2 & 0 \\ 1 & 0 \\ 4 & 3 \end{bmatrix}.$$

There are two useful properties of transpose in conjunction with matrix addition and multiplication:

- $(A+B)^\top = A^\top + B^\top$
  since $\big((A+B)^\top\big)_{ij} = (A+B)_{ji} = (A)_{ji} + (B)_{ji} = (A^\top)_{ij} + (B^\top)_{ij} = (A^\top + B^\top)_{ij}$;
- $(AB)^\top = B^\top A^\top$
  since $((AB)^\top)_{ij} = (AB)_{ji} = \sum_{k=1}^{\ell} (A)_{jk}(B)_{ki} = \sum_{k=1}^{\ell} (B^\top)_{ik}(A^\top)_{kj} = (B^\top A^\top)_{ij}$.

It is natural to extend matrix transpose to vector transpose. Since an $\mathbb{R}^n$-vector $\mathbf{a}$ can be thought of as an $n \times 1$ matrix, $\mathbf{a}^\top$ is correspondingly thought of as an $1 \times n$ matrix. If we use this in the context of matrix-vector multiplication (1.1) with an $1 \times n$ matrix, i.e., $A = \mathbf{a}^\top$, $A\mathbf{v} = \mathbf{a}^\top \mathbf{v}$ results in an 1×1 matrix which is a real-valued scalar. In such a case, we do not write it as a matrix but simply as a scalar;

$$\mathbf{a}^\top \mathbf{v} = \sum_{j=1}^{n} a_j v_j. \tag{2.3}$$

We call this (standard) inner product of $\mathbf{a}$ and $\mathbf{v}$ and will discuss it more in detail later when we introduce the notion of inner products in a vector space (Definition 4.3). With this definition of inner product, we can view matrix-vector multiplication as repeatedly computing the inner product between each row vector of the matrix and the vector.

Now that we know what the transpose of a matrix is, we can think of a matrix that is invariant to the transposition. We call such a matrix a symmetric matrix.

Definition 2.2 *A is symmetric if $A^\top = A$.*

Symmetric matrices possess many desirable properties and have been an important object of investigation in linear algebra. Some properties of symmetric matrices include;

- Any symmetric matrix is square.
- Every diagonal matrix is symmetric.
- For any matrix A, both $A^\top A$ and $AA^\top$ are symmetric.
- $ADA^\top$ and $A^\top DA$ are symmetric when D is a diagonal matrix.

We encourage you to think of how these properties hold. We will introduce you to a richer set of properties of symmetric matrices throughout the rest of the book.

2.2 Simultaneous Linear Equations

There is a close relationship between solving simultaneous linear equations and manipulating matrices. Consider the following system of two linear equations. We can represent the same system using matrix-vector multiplication or even more succinctly using a single matrix:[1]

$$\begin{cases} \text{(eqn 1)} & 1x & + & 1y & = & 5 \\ \text{(eqn 2)} & 2x & - & 1y & = & 1 \end{cases} \iff \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

$$\iff \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 2 & -1 & 1 \end{array}\right]$$

How do we solve these linear equations? We modify the second equation by multiplying both sides of the first equation by two and subtracting it from the second equation. This process, to which we often refer as Gaussian elimination, is equivalent to multiplying the 2×3 matrix on the right-hand side above with a special matrix, called an elementary matrix, from left. In this particular example, the elementary matrix is $\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}$ and captures the process of eliminating the first variable x from the second equation:

$$\text{(eqn 2)} - 2 \times \text{(eqn 1)} \rightsquigarrow \text{(eqn 2)}:$$

$$\begin{cases} 1x & + & 1y & = & 5 \\ & - & 3y & = & -9 \end{cases} \iff \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 2 & -1 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 1 & 5 \\ 0 & -3 & -9 \end{array}\right].$$

[1] We will often use 'eqn' in place of 'equation' for brevity.

After this step, we can determine the value of the second variable y by

$$-3y = -9 \implies y = 3.$$

After substituting y with 3 in the first equation, we can determine the value of the first variable x as

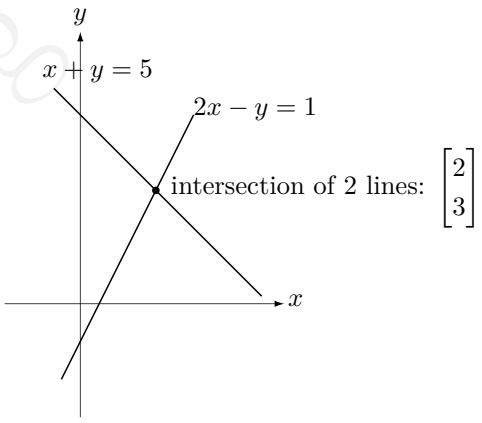
$$x + y = 5,\ y = 3 \implies x = 5 - 3 = 2.$$

The Geometry of Linear Equations

Unlike when there are three or more variables, it is possible for us to use visualization to investigate the geometry behind linear equations when there are only two variables. More specifically, we can interpret the geometry of linear equations from two perspectives.

Row-wise interpretation. We can plot the solution curve of each equation (i.e., a curve over which the equation is satisfied). A point where these two curves meet corresponds to the variable values that satisfy both equations. In the matrix-vector notation, this corresponds to comparing the multiplication of the row vector of the coefficient matrix and the variable vector against each element of the vector on the right-hand side:

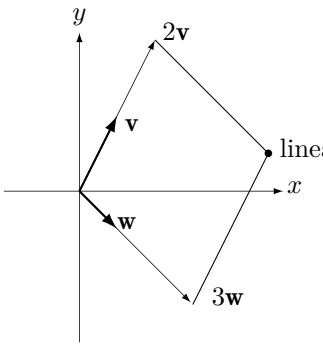
$$\begin{cases} x + y = 5 \\ 2x - y = 1 \end{cases} \iff \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$



Column-wise interpretation. According to (1.1), the left-hand side of the matrix-vector notation can be thought of as computing the linear combination of the column vectors of the coefficient matrix A with the variables serving as

linear weights. In this case, we can consider solving linear equations as finding the linear weights that produces the right-hand-side vector.

$$\begin{array}{rcrcl} x & + & y & = & 5 \\ 2x & - & y & = & 1 \end{array} \iff x\underbrace{\begin{bmatrix}1\\2\end{bmatrix}}_{\mathbf{v}} + y\underbrace{\begin{bmatrix}1\\-1\end{bmatrix}}_{\mathbf{w}} = \begin{bmatrix}5\\1\end{bmatrix}.$$



linear combination of 2 column vectors: $\begin{bmatrix}5\\1\end{bmatrix}$

With these in our mind, let us extend this two-variable example into a system of m equations with n variables. We now know that we can represent this system using an $m \times n$ coefficient matrix A, an n-dimensional variable vector $\mathbf{x}$ and an m-dimensional vector $\mathbf{b}$ as $A\mathbf{x} = \mathbf{b}$. Just like in the two-variable case above, we can interpret the solution to the system $\mathbf{x}$ as the intersection of m hyperplanes in $\mathbb{R}^n$, represented by the m rows of A, or as the combination of n column vectors of A in $\mathbb{R}^m$. Based on how those m hyperplanes are arranged relatively to each other, there may be either one unique solution, no solution or infinitely many solutions. From the column-wise interpretation, we need to define the concept of linear independence of vectors, which we will define later in Definition 3.4, in order to determine the existence of and the number of solutions. In the example above, the column vectors, $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, and therefore for any vector $\mathbf{b}$ on the right-hand side, there exists a unique solution. Later, we will show more generally that there exists a unique solution for any given $\mathbf{b}$ when there are at least m linearly independent column vectors in the coefficient matrix A.

2.3 An Example of Gaussian Elimination

Let us consider the following system of three equations and three variables:

$$\begin{array}{rrrcr} 2u & +v & +w & = & 5 \\ 4u & -6v & & = & -2 \\ -2u & +7v & +2w & = & 9. \end{array}$$

We can write this system more concisely as a matrix: $\left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{array}\right]$.
How would we solve these equations? We eliminate the first variable from the second equation and then eliminate the first and second variables from the third equation. We then determine the third variable from the third equation (because we have eliminated the first two variables already,) and plug it into the first and second equations, after which we can determine the rest of the variables. This whole process of progressive elimination can be expressed as a series of multiplication from left by so-called elementary matrices, where the elementary matrix is defined as an identity matrix with only one off-diagonal entry set to a non-zero number.[2] For instance, if we multiply A from left with an elementary matrix E that has $(E)_{ij} = b$ for some $i > j$, we end up with a matrix that satisfies

- All rows of EA are identical to those of A except for the i-th row;
- The i-th row of EA equals to the sum of the j-th row of A scaled by b and the i-th row of A.

Let us solve the linear system above by Gaussian elimination:

1. (eqn 2) -2 (eqn 1) $\rightsquigarrow$ (eqn 2) $\Leftrightarrow$ left multiplication of $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$$\begin{array}{rrrcr} 2u & +v & +w & = & 5 \\ & -8v & -2w & = & -12 \\ -2u & +7v & +2w & = & 9 \end{array}$$

$$\Longleftrightarrow \quad \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ -2 & 7 & 2 & 9 \end{array}\right]$$

2. (eqn 3) + (eqn 1) $\rightsquigarrow$ (eqn 3) $\Leftrightarrow$ left multiplication of $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

$$\begin{array}{rrrcr} 2u & +v & +w & = & 5 \\ & -8v & -2w & = & -12 \\ & 8v & +3w & = & 14 \end{array}$$

[2] We sometimes call a product of elementary matrices also an elementary matrix.

$$\iff \left[\begin{array}{ccc|c} 1 & 0 & 0 & \\ 0 & 1 & 0 & \\ 1 & 0 & 1 & \end{array}\right] \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ -2 & 7 & 2 & 9 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & 14 \end{array}\right]$$

3. (eqn 3) + (eqn 2) $\rightsquigarrow$ (eqn 3) $\Leftrightarrow$ left multiplication of $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$

$$\begin{array}{rrrcr} 2u & +v & +w & = & 5 \\ & -8v & -2w & = & -12 \\ & & w & = & 2 \end{array}$$

$$\iff \left[\begin{array}{ccc|c} 1 & 0 & 0 & \\ 0 & 1 & 0 & \\ 0 & 1 & 1 & \end{array}\right] \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & 14 \end{array}\right] = \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

More concisely, we can write the whole process above as successive multiplication of three elementary matrices from left (be conscious of the order of the elementary matrices):

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

$$= \left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

We call this process Gaussian elimination. After Gaussian elimination, the resulting i-th equation should have all the variables up to the $(i-1)$-th one eliminated. Equivalently, the resulting coefficient matrix $C = (c_{ij})$ satisfies $c_{i1} = \cdots = c_{i(i-1)} = 0$. Such a matrix is called an upper triangular matrix, because non-zero entries only exist in the upper triangular region of the matrix. We can similarly define a lower triangular matrix. Both upper and lower triangular matrices remain upper and lower triangular, respectively, even after matrix multiplication.

Fact 2.1 *The product of upper (lower) triangular matrices is again upper (lower) triangular. Furthermore, if the triangular matrices have unit diagonals, so does their product.*

Proof: Let $A = (a_{ij})$ and $B = (b_{ij})$ be upper triangular, that is, $a_{ij} = b_{ij} = 0$ if $i > j$. Assume that A has n columns and B has n rows. Then, for $i > j$,

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \sum_{k=1}^{i-1} \underbrace{a_{ik}}_{=0} b_{kj} + \sum_{k=i}^{n} a_{ik} \underbrace{b_{kj}}_{=0} = 0\,.$$

Therefore, AB is upper triangular. For lower triangular A and B, $A^\top$ and $B^\top$ are upper triangular, and $AB = \left(B^\top A^\top\right)^\top$ is lower triangular. The second statement holds since $(AB)_{ii} = a_{ii} b_{ii}$. ■

An alternative proof of Fact 2.1 can be done using Fact 2.2. We leave it for you as an exercise.

Once we have an upper triangular coefficient matrix, we can determine the solution readily by back-substitution. In the example above, we determine the values of the variables, starting from the final one to the first one by sweeping through the equations from bottom to top.

$$\begin{aligned} &\text{row } 3 : w = 2 \\ \Rightarrow\quad &\text{row } 2 : -8v = -12 + 2w = -8,\ v = 1 \\ \Rightarrow\quad &\text{row } 1 : 2u = 5 - v - w = 5 - 1 - 2 = 2,\ u = 1\,. \end{aligned}$$

Gaussian elimination may fail due to one of the following reasons.

- Non-singular case (fixable by row exchange): We may end up eliminating too many variables in the second row and cannot perform elimination in the third row. In this case, we simply exchange the second and third rows. This works because the order of equations in a linear system does not change the problem. See as an example:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 6 & 8 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 3 \\ 0 & 2 & 4 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 & 1 \\ 4 & 6 & 8 \\ 2 & 2 & 5 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix}.$$

- Singular case (not fixable): if a row is a scalar multiple of another row, Gaussian elimination results in a row with all zeros. In this case, there may be either infinitely many solutions or no solution, depending on the right-hand-side vector, and we cannot fix it to have a unique solution. For instance,

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 4 & 8 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 4 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

2.4 Block Matrices

We can view an entry of a matrix as encoding a relationship between the corresponding row and column. This view can be extended to a group of consecutive columns and a group of consecutive rows, in which case a submatrix corresponding to these column and row groups can be thought of as encoding the relationship between these groups. It is then natural to treat such submatrix as a block within the original matrix and define various operations in terms of these blocks rather than individual entries.

Let us write two vectors, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n_1+n_2}$ as $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$, where $\mathbf{u}_1, \mathbf{v}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{u}_2, \mathbf{v}_2 \in \mathbb{R}^{n_2}$. The inner product of these two vectors can then be written as

$$\mathbf{u}^\top \mathbf{v} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \mathbf{u}_1^\top \mathbf{v}_1 + \mathbf{u}_2^\top \mathbf{v}_2.$$

We can furthermore express it as matrix multiplication by treating $\mathbf{u}^\top$ and $\mathbf{v}$ as a $1 \times (n_1 + n_2)$ matrix and an $(n_1 + n_2) \times 1$ matrix, respectively:

$$\mathbf{u}^\top \mathbf{v} = \begin{bmatrix} \mathbf{u}_1^\top & \mathbf{u}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^\top \mathbf{v}_1 + \mathbf{u}_2^\top \mathbf{v}_2 \end{bmatrix}.$$

We can generalize this observation by considering a matrix $A = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix}$ instead of $\mathbf{u}$, where A_{11} and A_{12} are $m \times n_1$ matrix and $m \times n_2$ matrix, respectively. This matrix-vector multiplication, $A\mathbf{v}$, can then be understood as the sum of two vectors from matrix-vector multiplication, $A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2$:

$$A\mathbf{v} = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2.$$

We can further replace $\mathbf{v}$ with an $(n_1 + n_2) \times \ell$ matrix $B = \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}$, which results in the following expression for matrix multiplication between A and B:

$$AB = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix} = A_{11}B_{11} + A_{12}B_{21}. \tag{2.4}$$

This procedure applies equally well even when the order of A and B is swapped:[3]

$$\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} & A_{11}B_{12} \\ A_{21}B_{11} & A_{21}B_{12} \end{bmatrix} \tag{2.5}$$

[3]Recall that matrix multiplication is not commutative.

Example 2.1 When we encounter a matrix representing some data, such a matrix often exhibits a structure within it. For example, the following matrix is symmetric with diagonal blocks of zero entries, $A = \begin{bmatrix} \mathbf{0} & B \\ B^\top & \mathbf{0} \end{bmatrix}$. Such a structure can be used to facilitate the analysis of data behind the matrix. As another example, consider the following matrices which include blocks of heterogeneous data, and their product.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ a & b \end{bmatrix} \begin{bmatrix} -1 & c & e \\ -2 & d & f \end{bmatrix} = \begin{bmatrix} -5 & c+2d & e+2f \\ -11 & 3c+4d & 3e+4f \\ -a-2b & ac+bd & ae+bf \end{bmatrix}.$$

By grouping entries of the same type into a block and applying (2.5), we see that the diagonal blocks correspond to the products of the blocks of the same type, while the off-diagonal ones to the products of two blocks of two separate types, which provides us with a new perspective into the product of two original matrices.

$$\begin{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \\ \begin{bmatrix} a & b \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} & \begin{bmatrix} c & e \\ d & f \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} & \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} c & e \\ d & f \end{bmatrix} \\ \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} & \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} c & e \\ d & f \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} -5 \\ -11 \end{bmatrix} & \begin{bmatrix} c+2d & e+2f \\ 3c+4d & 3e+4f \end{bmatrix} \\ \begin{bmatrix} -a-2b \end{bmatrix} & \begin{bmatrix} ac+bd & ae+bf \end{bmatrix} \end{bmatrix}$$

∎

We can multiply two block matrices, each of which consists of more than two submatrices, by recursively applying (2.4) and (2.5) above. Let A be a block matrix consisting of smaller matrices A_{ij} as follows:

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

where the sizes of these sub-matrices are $A_{11} : m_1 \times n_1$, $A_{12} : m_1 \times n_2$, $A_{21} : m_2 \times n_1$, $A_{22} : m_2 \times n_2$, and $A : (m_1 + m_2) \times (n_1 + n_2)$, respectively, for

positive integers m_i's and n_i's. Similarly, let B be a block matrix consisting of appropriately sized sub-matrices B_{ij}:

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Then, AB equals

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}. \quad (2.6)$$

It must be satisfied that $A_{ij}B_{jk}$ is well-defined, for this matrix multiplication to hold.

To see the similarity between the block matrix multiplication and usual matrix multiplication, let us compare (2.4) with the multiplication of two 2×2 matrices:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

All indices of components in both cases coincides exactly. We have to however keep in mind that the order of blocks in each component of the resulting matrix must be strictly as it is in (2.4): neither $B_{11}A_{11} + A_{12}B_{21}$ nor $A_{11}B_{11} + B_{21}A_{12}$ can replace $A_{11}B_{11} + A_{12}B_{21}$, due to the lack of commutativity of matrix multiplication. This is unlike the usual multiplication of real numbers, where $a_{11}b_{11} + a_{12}b_{21} = b_{11}a_{11} + a_{12}b_{21} = a_{11}b_{11} + b_{21}a_{12}$.

For later use, let us consider powers of a block upper triangular matrix. When a block matrix consists of square diagonal blocks and all components below the diagonal blocks are zeros, we can express the k-th power of this block matrix in a simple form, as shown in Fact 2.2.

Fact 2.2 *Let a square matrix A be $\begin{bmatrix} B & C_1 \\ \mathbf{0} & D \end{bmatrix}$ where B and D are square matrices. Then, $A^k = \begin{bmatrix} B^k & C_k \\ \mathbf{0} & D^k \end{bmatrix}$ for some C_k's.*

Proof: $A^2 = \begin{bmatrix} B^2 & BC_1 + C_1D \\ \mathbf{0} & D^2 \end{bmatrix}$ with $C_2 = BC_1 + C_1D$. If we assume $A^{k-1} = \begin{bmatrix} B^{k-1} & C_{k-1} \\ \mathbf{0} & D^{k-1} \end{bmatrix}$, $A^k = \begin{bmatrix} B^k & BC_{k-1} + C_1D^{k-1} \\ \mathbf{0} & D^k \end{bmatrix}$ where $C_k = BC_{k-1} + C_1D^{k-1}$. ■

2.5 Inverse of Matrix

For matrix addition, an all-zero matrix is the identity of addition, and a matrix of which each entry's sign is flipped is the inverse of matrix addition. We have also learned of the identity matrix for matrix multiplication. In this section, we now study the inverse for matrix multiplication. We first define the inverse of a matrix as follows:

Definition 2.3 *Let A be an $m \times n$ matrix. A matrix B is a left-inverse of A if $BA = I_n$ and C is a right-inverse of A if $AC = I_m$. If a left-inverse of A is also a right-inverse of A, then we say that A is invertible and has an inverse.*

There are a few interesting observations derived from this definition:

- If B is a $k \times \ell$ matrix, then $n = k$ and $\ell = m$ for AB and BA to be well-defined. That is, B has to be of size $n \times m$. In fact, it must be $m = n$ if B is an inverse of A, as we will show later.
- Inverse is unique if it exists: if both B and C are inverses of A, then

$$B = BI_m = B(AC) = (BA)C = I_nC = C\,.$$

  We denote the inverse of A as A^{-1}.

- **A useful fact:** keep this in your mind as we will use it frequently throughout this book.

  If there exists $\mathbf{x} \neq \mathbf{0}$ satisfying $A\mathbf{x} = \mathbf{0}$, then A has no left-inverse and is not invertible.

- $B\mathbf{b}$ is a solution of $A\mathbf{x} = \mathbf{b}$ if B is a right-inverse of A.
- **A caution on using the left-inverse while solving $A\mathbf{x} = \mathbf{b}$**: Assume that a left-inverse B of A exists. Then, for $A\mathbf{x} = \mathbf{b}$, left multiplication of B to both sides results in $B(A\mathbf{x}) = B\mathbf{b}$ and consequently $\mathbf{x} = I_n\mathbf{x} = (BA)\mathbf{x} = B(A\mathbf{x}) = B\mathbf{b}$. However, for $\mathbf{x} = B\mathbf{b}$, $A\mathbf{x} = A(B\mathbf{b}) = (AB)\mathbf{b}$ may not reproduce $\mathbf{b}$ unless B is a right-inverse of A. This case frequently happens in regression analysis in statistics, and we are often satisfied with $B\mathbf{b}$ as an approximate solution. A typical example of a left-inverse which

may not be a right-inverse is the pseudoinverse in Fact 5.10, which can be used to derive an approximate solution to such a regression problem.

Example 2.2 Consider $A\mathbf{x} = \mathbf{b}$ when $A = \begin{bmatrix}1\\1\end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix}2\\1\end{bmatrix}$. 1×2 matrices $B = [\frac{1}{2}\ \frac{1}{2}]$, [1 0], and [0 1] are left-inverse matrices of A, since $BA = [1]$. However, $B\mathbf{b} = [\frac{3}{2}]$, [1], and [2] do not solve $A\mathbf{x} = \mathbf{b}$,[4] and A has no right-inverse. Among these multiple left-inverses, it is a standard practice to choose $[\frac{1}{2}\ \frac{1}{2}]$ for instance in regression analysis. Details are discussed in Fact 5.10 . ∎

- If $ad - bc \neq 0$, then $\begin{bmatrix}a & b\\c & d\end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}$. Check that this holds on your own.
- The inverse of a diagonal matrix is also diagonal: $A = \begin{bmatrix}d_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & d_n\end{bmatrix}$ $\Rightarrow\ A^{-1} = \begin{bmatrix}d_1^{-1} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & d_n^{-1}\end{bmatrix}$ if $d_i \neq 0$ for all i. If $d_i = 0$ for some i, $(AB)_{ij} = 0$ for $j = 1, \ldots, n$ and $AB \neq I$, for any B. That is, A is not invertible.
- If both A and B are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$. Check yourselves that $(AB)(B^{-1}A^{-1}) = I$ and $(B^{-1}A^{-1})(AB) = I$.
- $(A^{-1})^\top = (A^\top)^{-1}$, since $A^\top(A^{-1})^\top = (A^{-1}A)^\top = I^\top = I$.
- If A is symmetric and invertible, then A^{-1} is symmetric since $(A^{-1})^\top = (A^\top)^{-1} = A^{-1}$.

There is a simple yet useful observation on the inverse of a triangular matrix. It is particularly useful to familiarize yourself with proof techniques behind this result.

Theorem 2.1 *Assume A is an upper triangular matrix. Then, A is invertible if and only if every diagonal entry of A is non-zero. A^{-1} is also upper triangular*

[4] In fact, there is no solution that satisfies $A\mathbf{x} = \mathbf{b}$.

if it exists. If all diagonal entries of A are 1, then all the diagonal entries of A^{-1} are also 1.

Proof: We use mathematical induction on n, the size of a matrix A. As the induction hypothesis, we assume that the statement holds for matrices of size smaller than n. We note that this holds for all 1×1 matrices trivially, since any 1×1 matrix is invertible if and only if it is not zero. Let an $n \times n$ upper triangular matrix $A = \begin{bmatrix} A_{n-1} & \mathbf{u} \\ \mathbf{0}_{n-1}^\top & a \end{bmatrix}$, where A_{n-1} is an $(n-1) \times (n-1)$ upper triangular matrix, $\mathbf{u} \in \mathbb{R}^{n-1}$, and $a \in \mathbb{R}$. We may use $\mathbf{0}$ instead of $\mathbf{0}_{n-1}$ for brevity.

- "only if": Assume that the $n \times n$ upper triangular matrix A is invertible. If $a = 0$, then the last row of A vanishes, which makes the last row of AB vanish as well regardless of B. Because this contradicts to the invertibility of A, $a \neq 0$. Let $B = \begin{bmatrix} B_{n-1} & \mathbf{v} \\ \mathbf{w}^\top & b \end{bmatrix}$ be an inverse of A. From

$$AB = \begin{bmatrix} A_{n-1} & \mathbf{u} \\ \mathbf{0}^\top & a \end{bmatrix} \begin{bmatrix} B_{n-1} & \mathbf{v} \\ \mathbf{w}^\top & b \end{bmatrix} = \begin{bmatrix} A_{n-1}B_{n-1} + \mathbf{u}\mathbf{w}^\top & A_{n-1}\mathbf{v} + \mathbf{u}b \\ a\mathbf{w}^\top & ab \end{bmatrix} = I_n$$

  we need $a\mathbf{w}^\top = \mathbf{0}^\top$, which implies $\mathbf{w} = \mathbf{0}$ since $a \neq 0$. Then $\mathbf{u}\mathbf{w}^\top = \mathbf{0}_{n-1,n-1}$ and the first block of AB has to satisfy $A_{n-1}B_{n-1} = I_{n-1}$. On the other hand,

$$BA = \begin{bmatrix} B_{n-1} & \mathbf{v} \\ \mathbf{0}^\top & b \end{bmatrix} \begin{bmatrix} A_{n-1} & \mathbf{u} \\ \mathbf{0}^\top & a \end{bmatrix} = \begin{bmatrix} B_{n-1}A_{n-1} & B_{n-1}\mathbf{u} + \mathbf{v}a \\ \mathbf{0}^\top & ba \end{bmatrix} = I_n$$

  also implies $B_{n-1}A_{n-1} = I_{n-1}$. Therefore, A_{n-1} is an invertible upper triangular matrix of size $n-1$, and its diagonal should be non-zero by the induction hypothesis. Combining with $a \neq 0$, all diagonals of A are non-zero, and the "only if" statement holds for matrices of size n.

- "if": Assume that A has non-zero diagonals. Then, A_{n-1} is invertible by the induction hypothesis since A_{n-1} is an $(n-1) \times (n-1)$ upper triangular matrix with non-zero diagonals. If $B = \begin{bmatrix} A_{n-1}^{-1} & -a^{-1}A_{n-1}^{-1}\mathbf{u} \\ \mathbf{0}^\top & a^{-1} \end{bmatrix}$, then $BA = AB = I_n$ and $A^{-1} = B$. That is, A is invertible.

If A is invertible, $A^{-1} = \begin{bmatrix} A_{n-1}^{-1} & -a^{-1}A_{n-1}^{-1}\mathbf{u} \\ \mathbf{0}^\top & a^{-1} \end{bmatrix}$. Since A_{n-1}^{-1} is upper triangular by induction, A^{-1} is also upper triangular. Furthermore, if $a = 1$, the final

diagonal entry a^{-1} of A^{-1} is 1. Therefore, assuming the unit diagonal of A_{n-1}, all diagonal entries of A^{-1} are 1's by induction. ■

If you recall the relationship between the inverse and transpose of a matrix, you also see that Theorem 2.1 applies equally to lower triangular matrices. Therefore, all elementary matrices, resulting from Gaussian elimination, are invertible.

2.6 Triangular Factors and LU-Decomposition

It was not a coincidence that elementary matrices used in Gaussian elimination, which were multiplied to the coefficient matrix from left, were lower triangular. Let us consider the earlier example of Gaussian elimination.

$$\begin{bmatrix} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{bmatrix} \overset{①}{\rightsquigarrow} \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ -2 & 7 & 2 & 9 \end{bmatrix} \overset{②}{\rightsquigarrow} \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & 14 \end{bmatrix}$$
$$\overset{③}{\rightsquigarrow} \begin{bmatrix} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

In this process of elimination, we have multiplied the following three lower triangular elementary matrices:

$$①\text{: (eqn 2) - 2(eqn 1)}, \quad \tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad L_1 = \tilde{L}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$②\text{: (eqn 3) + (eqn 1)}, \quad \tilde{L}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad L_2 = \tilde{L}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

$$③\text{: (eqn 3) + (eqn 2)}, \quad \tilde{L}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad L_3 = \tilde{L}_3^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

We obtain a single lower triangular matrix that represents Gaussian elimination by multiplying these lower triangular matrices successively thanks to Fact 2.1. In Gaussian elimination, we alter the i-th row by adding a linear combination of the upper rows, i.e. the first to $(i-1)$-th rows, to the i-th row

itself. All the elementary matrices that correspond to these changes and their product then result in lower triangular matrices whose diagonal entries are all 1's. In this particular example, the resulting lower triangular matrix and its inverse are

$$\tilde{L} = \tilde{L}_3\tilde{L}_2\tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}$$

and

$$L = \tilde{L}^{-1} = \tilde{L}_1^{-1}\tilde{L}_2^{-1}\tilde{L}_3^{-1} = L_1L_2L_3 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}.$$

Gaussian elimination turns the coefficient matrix into an upper triangular matrix, and in this example, this matrix is

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix}.$$

After multiplying both sides by the inverse of the lower triangular matrix, we get

$$\begin{aligned} \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \\ &= LU, \end{aligned}$$

resulting in the product of the lower triangular matrix L and the upper triangular matrix U. We call this the LU-decomposition. The upper triangular matrix can be further decomposed as DU with an invertible diagonal matrix D and another upper triangular matrix U of which first non-zero entry is 1 for each row,[5] and we have $A = LDU$. It is called LDU-decomposition of the matrix A. The matrix in the above example has LDU-decomposition of

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & -8 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 1 & 1/4 \\ 0 & 0 & 1 \end{bmatrix}.$$

[5] The first non-zero entry may not be located at diagonal for a non-invertible U.

When A is not invertible, the upper triangular matrix U in LU- or LDU-decomposition of A is not invertible.

When A is invertible, with LU-decomposition we can solve the corresponding system of linear equations for any $\mathbf{b}$ by back-substitution as in

$$\begin{aligned} A\mathbf{x} = \mathbf{b} \quad &\Rightarrow \quad LU\mathbf{x} = \mathbf{b} \ \text{ or } \ L\mathbf{y} = \mathbf{b} \\ &\Rightarrow \quad \mathbf{y} = L^{-1}\mathbf{b} \\ &\Rightarrow \quad U\mathbf{x} = \mathbf{y} = L^{-1}\mathbf{b} \\ &\Rightarrow \quad \mathbf{x} = U^{-1}L^{-1}\mathbf{b}\,. \end{aligned}$$

As we mentioned before, linear systems with upper or lower triangular coefficient matrices can be efficiently solved by back-substitution.

We can also use LU-decomposition to compute the inverse of the coefficient matrix A. If U is invertible, $A^{-1} = U^{-1}L^{-1}$. We can thus perform Gaussian elimination on an $n \times 2n$ expanded coefficient matrix $[A|I]$, so that the first half results in an identity matrix, which transforms the latter half (the identity matrix) into the inverse of A:

$$U^{-1}L^{-1}\left[A\middle|I\right] = U^{-1}\left[U\middle|L^{-1}\right] = \left[I\middle|U^{-1}L^{-1}\right] = \left[I\middle|A^{-1}\right].$$

As an illustration, consider computing the following inverse:

$$A^{-1} = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}^{-1}.$$

First, we augment the original coefficient matrix A by attaching an identity matrix, resulting in

$$\left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 4 & -6 & 0 & 0 & 1 & 0 \\ -2 & 7 & 2 & 0 & 0 & 1 \end{array}\right].$$

We then multiply this augmented matrix with the lower triangular matrix $\tilde{L} = L^{-1}$ from Gaussian elimination, which transforms the original coefficient matrix A into the upper triangular matrix U. We continue by multiplying this matrix again from left with U^{-1}, making the coefficient matrix into a diagonal matrix. Finally, we turn it into an identity matrix by multiplying it with the inverse of this diagonal matrix.

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 4 & -6 & 0 & 0 & 1 & 0 \\ -2 & 7 & 2 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right]$$

$$\rightsquigarrow \left[\begin{array}{ccc} 1 & \frac{1}{8} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc|ccc} 2 & 1 & 1 & 1 & 0 & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 2 & 0 & \frac{3}{4} & \frac{3}{4} & \frac{1}{8} & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right]$$

$$\rightsquigarrow \left[\begin{array}{ccc} 1 & 0 & -\frac{3}{4} \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc|ccc} 2 & 0 & \frac{3}{4} & \frac{3}{4} & \frac{1}{8} & 0 \\ 0 & -8 & -2 & -2 & 1 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & \frac{3}{2} & -\frac{5}{8} & -\frac{3}{4} \\ 0 & -8 & 0 & -4 & 3 & 2 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right]$$

$$\rightsquigarrow \left[\begin{array}{ccc} \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{8} & 0 \\ 0 & 0 & 1 \end{array}\right] \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & \frac{3}{2} & -\frac{5}{8} & -\frac{3}{4} \\ 0 & -8 & 0 & -4 & 3 & 2 \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{3}{4} & -\frac{5}{16} & -\frac{3}{8} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{3}{8} & -\frac{1}{4} \\ 0 & 0 & 1 & -1 & 1 & 1 \end{array}\right]$$

This matrix, resulting from the concatenation of the original matrix and the identity matrix, is equivalent to the product of all the matrices that were multiplied from left. Because these multiplications transformed the coefficient matrix A into the identity matrix, this resulting augmented matrix represents the inverse of the coefficient matrix. That is,

$$A^{-1} = \begin{bmatrix} 3/4 & -5/16 & -3/8 \\ 1/2 & -3/8 & -1/4 \\ -1 & 1 & 1 \end{bmatrix}.$$

In short, this whole process can be expressed succinctly as

$$A^{-1} = (LDU)^{-1} = U^{-1}D^{-1}L^{-1},$$

although it requires Gaussian elimination for us to eventually obtain L, D and U.

Uniqueness of LU-Decomposition

If a square matrix A is invertible, then it can be decomposed as $A = LDU$ where triangular matrices have unit diagonal entries and the diagonal matrix has nonzero diagonal entries. This decomposition is unique. Assume two possible ways to factorize the matrix A; $A = L_1D_1U_1 = L_2D_2U_2$. By moving L_2, U_1, and D_1 to the other side, respectively, we get $L_2^{-1}L_1 = D_2U_2U_1^{-1}D_1^{-1}$. The left-hand and right-hand sides are lower and upper triangular, respectively, according to Theorem 2.1 and Fact 2.1. In other words, they are diagonal matrices. Because $L_2^{-1}L_1$ has unit diagonals, $L_2^{-1}L_1 = I$, and similarly, $U_2U_1^{-1} = I$. It follows that $D_1 = D_2$, and therefore $A = LDU$ is unique.

Let us further assume that A is a symmetric matrix and factorized into $A = LDU$ without any row swaps. Due to the symmetry of A, $LDU = U^\top DL^\top$

holds. According to the uniqueness of LU-decomposition above, $U = L^\top$, meaning that we can factorize A as $A = LDL^\top$.

LU-Decomposition with Row Exchanges

We can perform Gaussian elimination on a matrix, such as the one below, that would not admit Gaussian elimination in its original form:

$$A = \begin{bmatrix} 0 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

We do so by multiplying A from left with the following matrix, which results in swapping the first and second rows:[6]

$$Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

We get such a matrix by (repeatedly) swapping two rows of the identity matrix and call it a permutation matrix. Each row and column of a permutation matrix has exactly one 1 each, and all the other entries are 0's. (Refer to Appendix B for the details of permutation matrices.) This results in the following matrix QA:

$$QA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 6 \\ 0 & 2 & 3 \end{bmatrix},$$

which then can be factorized into $QA = LU$ using LU-decomposition.

2.7 LU-Decomposition by Partial Pivoting

As we have seen, any matrix A can be decomposed into a lower and an upper matrices by Gaussian elimination after some appropriate row swaps. In other words, once we know desired row swaps, represented as a permutation matrix Q, before starting Gaussian elimination, we can proceed with the decomposition of QA. It is however difficult to identify necessary row swaps in advance of Gaussian elimination, in practice. One naive but inefficient approach is to perform two rounds of Gaussian elimination. We identify the permutation matrix Q in the first round, and use it to perform the second round of Gaussian elimination on QA, resulting in L and U. This requires twice as much computation

[6] If we multiply A from right, it will swap the first and second columns instead.

and raises a question whether we can avoid this extra step of identifying the permutation matrix in advance.

Although we have seen that it is necessary to swap rows when the next pivot candidate was zero during Gaussian elimination, there are other reasons to swap rows in Gaussian elimination. For instance, it is widely known in numerical analysis that the numerical stability of Gaussian elimination can be improved with divisions by larger pivot entries. This suggests we choose as a pivot the largest absolute value among the entries below and on the diagonal in each column and swap rows to place this selected pivot on the diagonal.

Let Q_j be an identity matrix, or a permutation matrix that swaps the j-th row with another row below the j-th one, and E_j be an elementary matrix that eliminates all entries below the diagonal of the j-th column in

$$Q_j E_{j-1} Q_{j-1} \cdots E_1 Q_1 A.$$

E_j has the unit diagonal and non-zero entries only below the diagonal of the j-th column. We then write k steps of partial pivoting as

$$E_k Q_k E_{k-1} Q_{k-1} \cdots E_2 Q_2 E_1 Q_1 A = U, \tag{2.7}$$

where U is an upper triangular matrix.

Consider a matrix $\hat{E}_j$ with $j < k$ that satisfies

$$Q_k \cdots Q_{j+1} E_j = \hat{E}_j Q_k \cdots Q_{j+1}. \tag{2.8}$$

In both E_j and $\hat{E}_j$, the diagonal entries are all 1, and non-zero entries exist below the diagonal of the j-th column only. We see that this is the case by observing first that $Q_j = Q_j^\top = Q_j^{-1}$ and that Q_j swaps columns when multiplied to a matrix from the right. $E'_j = Q_{j+1} E_j Q_{j+1}$ shares the same diagonal entries as well as the same patterns of non-zero off-diagonal entries with E, because the left application of Q_{j+1} swaps the $(j+1)$-th row and another row below of E while the right application of Q_{j+1} the $(j+1)$-th column and another column after. By repeating this, we arrive at

$$\hat{E}_j = Q_k \cdots Q_{j+1} E_j (Q_k \cdots Q_{j+1})^\top = Q_k \cdots Q_{j+1} E_j Q_{j+1} \cdots Q_k.$$

We can therefore rewrite (2.7) as

$$E_k \hat{E}_{k-1} \cdots \hat{E}_2 \hat{E}_1 Q_k Q_{k-1} \cdots Q_2 Q_1 A = EQA = U, \quad \text{or} \quad QA = LU,$$

where Q is a permutation matrix, and E and L are lower triangular matrices with unit diagonals as we claimed, according to (2.8).

2.8 Inverse of Block Matrix

Let us consider the following 2×2 matrix A with non-zero a_{11}:

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

To eliminate a_{21}, we multiply an elementary matrix to the left of A as

$$\begin{bmatrix} 1 & 0 \\ -a_{21}a_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} - a_{21}a_{11}^{-1}a_{12} \end{bmatrix}.$$

Then, to convert a_{11} into 1, we scale the first row by multiplying a diagonal matrix to the left of A as

$$\begin{bmatrix} a_{11}^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} - a_{21}a_{11}^{-1}a_{12} \end{bmatrix} = \begin{bmatrix} 1 & a_{11}^{-1}a_{12} \\ 0 & a_{22} - a_{21}a_{11}^{-1}a_{12} \end{bmatrix}.$$

These two operations are achieved by multiplying once the following matrix

$$\begin{bmatrix} a_{11}^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -a_{21}a_{11}^{-1} & 1 \end{bmatrix} = \begin{bmatrix} a_{11}^{-1} & 0 \\ -a_{21}a_{11}^{-1} & 1 \end{bmatrix}.$$

This matrix representation of Gaussian elimination can be extended to block matrices.

Let A be a square matrix that can be represented as a block matrix as follows:

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

where A_{11} and A_{22} are also square matrices. If A_{11} is invertible, we can eliminate A_{21} using Gaussian elimination. We can illustrate this process by matrix multiplication:

$$\begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I_{11} & A_{11}^{-1}A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}. \tag{2.9}$$

Here I_{11} is an identity matrix of the same size as A_{11}. Similarly, we can eliminate A_{12} by Gaussian elimination if A_{22} is invertible, which can be expressed as

$$\begin{bmatrix} I_{11} & -A_{12}A_{22}^{-1} \\ \mathbf{0} & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & \mathbf{0} \\ A_{22}^{-1}A_{21} & I_{22} \end{bmatrix}. \tag{2.10}$$

Let $S_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$ to simplify the right-hand side of (2.9). We call S_{22} a Schur complement of A_{11} with respect to A.[7] The right-hand side of (2.9) simplifies to $\begin{bmatrix} I_{11} & A_{11}^{-1}A_{12} \\ \mathbf{0} & S_{22} \end{bmatrix}$, and with invertible S_{22} we can perform Gaussian elimination further as follows:[8]

$$\begin{bmatrix} I_{11} & -A_{11}^{-1}A_{12}S_{22}^{-1} \\ \mathbf{0} & S_{22}^{-1} \end{bmatrix} \begin{bmatrix} I_{11} & A_{11}^{-1}A_{12} \\ \mathbf{0} & S_{22} \end{bmatrix} = \begin{bmatrix} I_{11} & \mathbf{0} \\ \mathbf{0} & I_{22} \end{bmatrix}.$$

By plugging in (2.9), we get

$$\begin{bmatrix} I_{11} & -A_{11}^{-1}A_{12}S_{22}^{-1} \\ \mathbf{0} & S_{22}^{-1} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I_{11} & \mathbf{0} \\ \mathbf{0} & I_{22} \end{bmatrix},$$

from which we observe that the inverse of the original coefficient matrix A is the product of the two triangular matrices:

$$\begin{aligned} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} I_{11} & -A_{11}^{-1}A_{12}S_{22}^{-1} \\ \mathbf{0} & S_{22}^{-1} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S_{22}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S_{22}^{-1} \\ -S_{22}^{-1}A_{21}A_{11}^{-1} & S_{22}^{-1} \end{bmatrix} \end{aligned} \tag{2.11}$$

As we have demonstrated, two equations (2.9) and (2.10), arising from Gaussian elimination, are useful for performing thought experiments on various types of matrices in the form of block matrices.

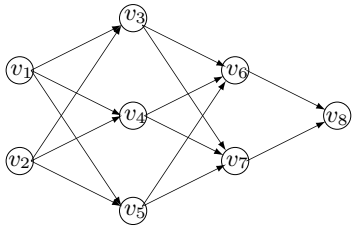
2.9 Application: Graphs and Matrices

We mathematically express as a graph or network the relationship between multiple objects in for instance a social network as well as engineering systems. We call v_i, in the figure below, a node, and you can imagine any object, that can have a relationship with other objects, as nodes, such as a person, organization, machine and computer. When two nodes are related to each other, we connect these two nodes with an edge. Such an edge could be either undirected or directed. In this book, we refer to an undirected edge as an edge, and to a directed

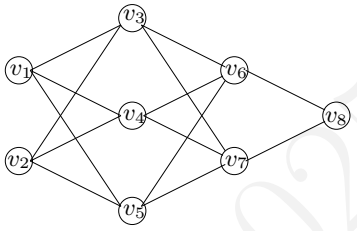
[7] Similarly, with an invertible A_{22}, a Schur complement of A_{22} with respect to A is $S_{11} = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

[8] In other words, perform the following replacements; $I_{11} \rightsquigarrow A_{11}$, $A_{11}^{-1}A_{12} \rightsquigarrow A_{12}$, $\mathbf{0} \rightsquigarrow A_{21}$, $S_{22} \rightsquigarrow A_{22}$, and apply (2.10).

edge as an arc, in order to distinguish them clearly.[9] It is intuitive to visualize such a graph but is challenging to manipulate it. We thus express a graph as a matrix to analyze the properties of and perform various manipulations of the graph.



(a) Directed Graph



(b) Undirected Graph

Figure 2.1: Examples of Graphs

With n nodes in a graph, we create an $n \times n$ matrix. Each row/column of this matrix corresponds to one node in the graph. If there is an arc between the i-th and j-th nodes, the (i, j)-th entry of the matrix takes the value 1. Otherwise, it is set to 0. For an edge, we regard it as two arcs with opposite origins and destinations. This matrix is called an adjacency matrix. The adjacency matrix of an undirected graph is thus symmetric by construction.

As an example, let us convert the graphs above into the adjacency matrices:

$$A_1 = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

By reordering the nodes, to be $v_3, v_4, v_5, v_8, v_1, v_2, v_6, v_7$, we get the following

[9] You can think of the road connecting two cities as an (undirected) edge, while a relationship between a cause and an effect as an arc.

block matrix as an adjacency matrix of the undirected graph:

$$\left[\begin{array}{ccccc|cccc} & & & & & 1 & 1 & 1 & 1 \\ & & & & & 1 & 1 & 1 & 1 \\ & & & & & 1 & 1 & 1 & 1 \\ & & & & & & & 1 & 1 \\ \hline 1 & 1 & 1 & & & & & & \\ 1 & 1 & 1 & & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ 1 & 1 & 1 & 1 & & & & & \end{array}\right] = \begin{bmatrix} \mathbf{0} & B \\ B^\top & \mathbf{0} \end{bmatrix}$$

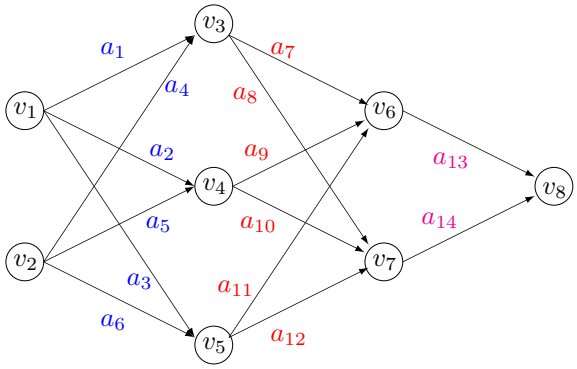
By examining this adjacency matrix, we observe that there are two groups of nodes, $\{v_3, v_4, v_5, v_8\}$ and $\{v_1, v_2, v_6, v_7\}$. There is no edge between nodes within each of these groups, but there are edges that connect nodes from these two groups. We call such a graph a bipartite graph. This is a simple demonstration of how we read out various properties of a graph by analyzing the graph's adjacency matrix.

Example 2.3 We can count the number of different ways to reach from one node to another in a graph by matrix multiplication. Suppose A is an adjacency matrix of a graph. The (i, j)-th entry of A^m corresponds to the number of ways to connect the i-th and j-th nodes of the graph. By definition, $A = A^1$ expresses whether there is an edge between two nodes, and equivalently, whether there is $(= 1)$ or is not $(= 0)$ a way to reach one node from another. In the case of A^2, $(A^2)_{ij}$ increments by one if there is a path connecting the i-th and j-th nodes via another node k, because $(A^2)_{ij} = \sum_{k=1}^{n} a_{ik}a_{kj}$. By mathematical induction, we can show that this interpretation extends to an arbitrary $m \geq 1$. See any textbook on graph theory for more details.

∎

Such analysis of an adjacency matrix is widely used in various disciplines, including applied mathematics, engineering and social sciences. In doing so, it is a usual and helpful practice to express and analyze an adjacency matrix as a block matrix.

Let us name the arcs of the directed graph in Figure 2.1a as follows.



A directed graph can also be expressed as an incidence matrix. Each row of the incident matrix corresponds to each node, and each column to each arc. We set a_{ij} to 1 if the j-th arc starts at the i-th node, and to -1 if the j-th arc terminates at the i-th node. Each column of any incidence matrix thereby has two non-zero entries; one 1 and one -1. An incidence matrix A corresponding to the directed graph above is then

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

Consider an example where this directed graph represents a network of airports, and each arc is associated with x_j that represents the number of flights from the i-th airport to the j-airport each day. $\mathbf{x} = (x_j)$ is then a vector representing all the flights in the sky each day. $A\mathbf{x}$ in turn represents the differences between the incoming and outgoing flights at all airports. For instance, let $\mathbf{b}$ satisfy $b_1 > 0, b_2 > 0, b_8 = -(b_1 + b_2) < 0, b_3 = \cdots = b_7 = 0$. When $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$, $\mathbf{x}$ corresponds to having all flights from the first two nodes, v_1 and v_2, eventually fly out to the final node v_8 without any loss of the flights in-between. Of course, we want to constraint x_j to be non-negative in practice. This is an example of using an incidence matrix to express the network flow and for representing the conservation of the flow, i.e. $A\mathbf{x} = \mathbf{b}$.

Chapter 3

Vector Spaces

Imagine a familiar physical quantity of force, which most of us have learned about earlier in our education. Given an object, we can apply force to manipulate this object. We can also apply two different forces to the object simultaneously, which would be equivalent to applying the sum of two forces, or the addition of these forces, to the object once. We can apply the same force twice to the object, which would be equivalent to applying double the force to the object. This thought experiment hints at a space of forces that can be added with each other and multiplied with a scalar. In fact, this is how we define a vector space in this chapter. A vector space is a set of things, called vectors, and these vectors can be added to each other and multiplied by a scalar. In this space, the distributive rule holds between addition and scalar multiplication.

Many things can be vectors, and naturally we can build many different vector spaces, such as a collection of points on a plane, a collection of real matrices of the same size, a collection of quadratic polynomials with real coefficients, a collection of random variables on a sample space and more. When we combine vector addition and scalar multiplication into one operation, we call it linear combination. With linear combination, we can ask mathematically interesting questions about a given vector space. Is a finite set of vectors minimal such that no vector in the set is a linear combination of other vectors? Are finite number of vectors many enough to represent every vector in the vector space as a linear combination? The answers to these two questions lead us to the concept of a basis, which is defined as a minimal set of vectors representing a whole vector space. From this definition of the basis, we can define the dimension of a vector space as the number of vectors in a basis. This allows us to compare two vector

spaces and conclude that two vector spaces of the same dimension are (roughly) equivalent.

By introducing a function between vector spaces called a linear transformation, we can tell more interesting stories. We soon see that a linear transformation corresponds to a matrix in a one-to-one fashion. Therefore, studying matrices equally extends our understanding of linear transformations. For a linear transformation described in terms of a matrix, the range of the transformation is defined as the column space of the matrix, and the kernel of the transformation is the null space of the matrix. Characterization of these spaces is a by-product of Gaussian elimination of the matrix. We refine the Gaussian elimination further to obtain the so-called row echelon form, whose pattern of non-zero entries is essential to finding the column and null spaces as well as the rank of the matrix. Many important observations on matrices and vector spaces are related to the rank. We also briefly look at special structures of matrices corresponding to geometric transformations like a rotation, a reflection, and a projection.

3.1 Vector Spaces and Subspaces

We define vector operations in a vector space by collecting all manipulations necessary for solving a linear system $A\mathbf{x} = \mathbf{b}$ as well as investigating the solutions to its associated linear system $A\mathbf{x} = \mathbf{0}$ called a homogeneous system. When you imagine a vector, you might think of a point in a familiar vector space of $\mathbb{R}^n$. A vector space is however a much more general concept, including for instance a set of all equal-sized matrices and a collection of all real-valued functions that share the same domain.

3.1.1 Operations in a Vector Space

There are two basic operations in a vector space; vector addition and scalar multiplication. All other operations are derived from these two basic operations.

1. Scalar multiplication: First, we must think of what a scalar is. In this book, we mostly consider a real-valued scalar in $\mathbb{R}$, although a scalar can be either real-valued or complex-valued ($\mathbb{C}$). For any scalar c and vectors $\mathbf{v}$, its scalar multiple $c\mathbf{v}$ is also a vector. The scalar multiplication is associative, i.e. $(c_1c_2)\mathbf{v} = c_1(c_2\mathbf{v})$. The multiplicative identity of scalars

denoted by 1 works as $1\mathbf{v} = \mathbf{v}$. In addition, for vectors $\mathbf{v}$ and $\mathbf{w}$, we denote $0\mathbf{v} = \mathbf{0}$ and $(-1)\mathbf{v} = -\mathbf{v}$ where 0 and -1 are the identity and its inverse of scalar addition.

2. Vector addition: The sum of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is also a vector, i.e., $\mathbf{v}_1 + \mathbf{v}_2$. Vector addition is both commutative and associative; $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$ and $\mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3) = (\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3$. The additive identity is $\mathbf{0}$ and is often self-evident given a vector space. For instance, some of these identities include an all-zero vector, an all-zero matrix, and a constant function that outputs only 0. We denote $\mathbf{w} + (-1)\mathbf{v} = \mathbf{w} - \mathbf{v}$ for simplicity.

3. Two distributive interactions between vector addition and scalar multiplication: $c(\mathbf{v}_1 + \mathbf{v}_2) = c\mathbf{v}_1 + c\mathbf{v}_2$ and $(c_1 + c_2)\mathbf{v} = c_1\mathbf{v} + c_2\mathbf{v}$. From these, we can derive the inverse of an arbitrary vector $\mathbf{v}$ for the vector addition. Since $\mathbf{v} - \mathbf{v} = 1\mathbf{v} + (-1)\mathbf{v} = (1-1)\mathbf{v} = 0\mathbf{v} = \mathbf{0}$, $-\mathbf{v} = (-1)\mathbf{v}$ is the additive inverse of the vector $\mathbf{v}$.

Definition 3.1 *A set $\mathbb{V}$ is a vector space if all vectors in $\mathbb{V}$ and scalars in $\mathbb{R}$ or $\mathbb{C}$ satisfy the operational rules above.*

Throughout the rest of this book, we consider a real-valued scalar ($\mathbb{R}$) unless specified otherwise. Some of the representative examples of vectors spaces include $\mathbb{R}^n$, $\mathbb{R}^\infty$, a space of a fixed-size matrices and a space of vector-valued functions. In particular, $\mathbb{R}^n$ is a standard finite-dimensional vector space. We will discuss more about the dimensionality later.

If a subset of vectors within a vector space satisfy the rules of a vector space themselves, we call this subset a subspace of the vector space. In this case, any linear combination of vectors in this subspace must be part of this subspace, where the linear combination of k vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and k scalars $c_1, \ldots, c_k$ is defined as

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \sum_{i=1}^{k} c_i\mathbf{v}_i.$$

We define a subspace of a given vector space in terms of the linear combination.

Definition 3.2 *A subspace of a vector space is a non-empty subset of the vector space such that all linear combinations stay in the subset.*

In order to show a non-empty subset $\mathbb{W}$ of a vector space $\mathbb{V}$ is a subspace, all we need to do is to check whether $\mathbb{W}$ is closed under vector addition and scalar multiplication. That is, we check whether

- $\mathbb{W} \subset \mathbb{V}$;
- $\mathbf{v}, \mathbf{w} \in \mathbb{W} \;\Rightarrow\; \mathbf{v} + \mathbf{w} \in \mathbb{W}$;
- $c \in \mathbb{R}, \mathbf{v} \in \mathbb{W} \;\Rightarrow\; c\mathbf{v} \in \mathbb{W}$.

A few examples of subspaces of $\mathbb{V}$ include $\{\mathbf{0}\}$ (potentially the smallest non-empty subspace), $\{c\mathbf{v} : c \in \mathbb{R}\}$ for a $\mathbf{v} \in \mathbb{V}$ (a 1-dimensional subspace) as well as $\{c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n : c_1, \ldots, c_n \in \mathbb{R}\}$ for fixed $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{V}$. A typical example of a non-subspace is $\{(x, y) : x \geq 0, y \geq 0\}$ in $\mathbb{R}^2$. In the case of a vector space consisting of matrices, some of the example subspaces include a set of all lower triangular matrices and a set of all symmetric matrices.

We can interestingly write scalar multiplication as either $c\mathbf{v}$ or $\mathbf{v}c$, when $\mathbf{v} \in \mathbb{R}^n$. The former, $c\mathbf{v}$, is a standard way to express scalar multiplication to a vector. On the other hand, we can think of $\mathbf{v}c$ as performing matrix multiplication $\mathbf{v}[c]$, where $\mathbf{v}$ is an $n \times 1$ matrix and c a 1×1 matrix. The latter with the associativity of matrix multiplication may help you identify useful expressions with more than two multiplicands, such as the one below:

$$(\mathbf{u}^\top \mathbf{v})\mathbf{w} = \mathbf{w}(\mathbf{u}^\top \mathbf{v}) = (\mathbf{w}\mathbf{u}^\top)\mathbf{v},$$

where $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Unlike the first two terms, which are both scalar multiplication, the right-most term is matrix-vector multiplication.

Let us introduce the following notation for summing sets of vectors:

Definition 3.3 *For any pair of subsets, A and B, of a vector space $\mathbb{V}$, we define the sum of A and B as*[a]

$$A + B = \{\mathbf{u} + \mathbf{v} : \mathbf{u} \in A, \mathbf{v} \in B\}.$$

If both $\mathbb{U}$ and $\mathbb{W}$ are subspaces of $\mathbb{V}$, and $\mathbb{U} \cap \mathbb{W} = \{\mathbf{0}\}$, we use $\mathbb{U} \oplus \mathbb{W}$ in place of $\mathbb{U} + \mathbb{W}$ and call it the direct sum.

[a] For brevity, we often shorten $\{\mathbf{v}\} + A = \mathbf{v} + A$ for $\mathbf{v} \in \mathbb{V}$ and $A \subset \mathbb{V}$.

From this definition, we can derive an important uniqueness property of the direct sum. If a vector in $\mathbb{U} \oplus \mathbb{W}$ can be expressed as $\mathbf{u}_1 + \mathbf{v}_1 = \mathbf{u}_2 + \mathbf{v}_2$ with

$\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{U}$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{W}$, then $\mathbf{u}_1 - \mathbf{u}_2 = \mathbf{v}_2 - \mathbf{v}_1 \in \mathbb{U} \cap \mathbb{W}$. Because $\mathbb{U} \cap \mathbb{W} = \{\mathbf{0}\}$ according to Definition 3.3, it holds that $\mathbf{u}_1 = \mathbf{u}_2$ and $\mathbf{v}_1 = \mathbf{v}_2$. In other words, there is a unique way to express each vector in $\mathbb{U} \oplus \mathbb{W}$ in terms of vectors from $\mathbb{U}$ and $\mathbb{W}$.

Fact 3.1 *A vector in a direct sum has a unique representation: For* $\mathbf{v} \in \mathbb{U} \oplus \mathbb{W}$*, there exists unique* $\mathbf{u} \in \mathbb{U}$ *and* $\mathbf{w} \in \mathbb{W}$ *such that* $\mathbf{v} = \mathbf{u} + \mathbf{w}$*.*

A two-dimensional Euclidean space with its two axes is a typical example used to demonstrate the relationship between the summand subspaces and their direct sum. A two dimensional Euclidean space can be expressed as a direct sum of two subspaces induced from the two axes, $\mathbb{R} \times \{0\}$ and $\{0\} \times \mathbb{R}$. That is, $\mathbb{R}^2 = (\mathbb{R} \times \{0\}) \oplus (\{0\} \times \mathbb{R})$, where the symbol $\times$ is called a Cartesian product and defined as, for any two sets A and B,

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

with a and b that could be real numbers, vectors, or even functions.[1] This notation becomes helpful later when we encounter a vector space expressed as a direct sum of subspaces.

3.1.2 Two Fundamental Subspaces induced by a Matrix

Let $A = [\,\mathbf{a}_1 \,|\, \cdots \,|\, \mathbf{a}_n]$, $\mathbf{a}_i \in \mathbb{R}^m$ be an $m \times n$ matrix. We can readily come up with two subspaces from this matrix; an m-dimensional column space and an n-dimensional null space.

- The column space of A, $\mathrm{Col}\,(A)$: the collection of linear combinations of columns of A.

  $$\mathrm{Col}\,(A) = \{v_1\mathbf{a}_1 + \cdots + v_n\mathbf{a}_n : v_1, \ldots, v_n \in \mathbb{R}\} = \{A\mathbf{v} : \mathbf{v} \in \mathbb{R}^n\} \subset \mathbb{R}^m\,.$$

  We enumerate a few (simple) properties of the column space.

  1. $A\mathbf{x} = \mathbf{b}$ is solvable if and only if $\mathbf{b} \in \mathrm{Col}\,(A)$;
  2. $\mathrm{Col}\,(I_n) = \mathbb{R}^n$;
  3. If A is an $n \times n$ invertible matrix, then $\mathrm{Col}\,(A) = \mathbb{R}^n$.
  4. $\mathrm{Col}\,(A)$ is a subspace of $\mathbb{R}^m$.

[1] Examples of the Cartesian product include $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ and $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$.

- The null space of A, $\text{Null}(A)$: the collection of vectors being mapped to $\mathbf{0}$ via the matrix A.

$$\text{Null}(A) = \{\mathbf{v} \in \mathbb{R}^n : A\mathbf{v} = \mathbf{0}\}.$$

  We often call $\text{Null}(A)$ a kernel of A. Here are a few (simple) properties of the null space.

  1. $\text{Null}(A) = \mathbb{R}^n$ if and only if $A = \mathbf{0}$;
  2. $\text{Null}(I_n) = \{\mathbf{0}\}$;
  3. If A is an $n \times n$ invertible matrix, then $\text{Null}(A) = \{\mathbf{0}\}$.
  4. $\text{Null}(A)$ is a subspace of $\mathbb{R}^n$.

When we multiply a matrix with another matrix, the former's column space often shrinks, unless the matrix being multiplied from the right is invertible. In that case, the column space does not change.

Lemma 3.1 *For any pair of matrices, A and B, where the number of columns of A and the number of rows of B coincide,* $\text{Col}(AB) \subset \text{Col}(A)$. *If B is invertible, then* $\text{Col}(AB) = \text{Col}(A)$.

Proof: For any $\mathbf{v}$, $(AB)\mathbf{v} = A(B\mathbf{v}) \in \text{Col}(A)$. Therefore, $\text{Col}(AB) \subset \text{Col}(A)$. Assume B is invertible. Set $C = AB$. Then, $A = CB^{-1}$ and $\text{Col}(A) = \text{Col}(CB^{-1}) \subset \text{Col}(C) = \text{Col}(AB)$ by the first part of the lemma. ■

The null space of a matrix determines the structure of the set of solutions to a linear system defined by the matrix. We obtain the solution set of any such linear system by shifting the null space.

Fact 3.2 *Assume $A\mathbf{x}^\star = \mathbf{b}$. Then,* $\{\mathbf{x} : A\mathbf{x} = \mathbf{b}\} = \{\mathbf{x}^\star + \mathbf{y} : \mathbf{y} \in \text{Null}(A)\} = \mathbf{x}^\star + \text{Null}(A)$.

Proof: Let $A\mathbf{x} = \mathbf{b}$. Then, $A(\mathbf{x} - \mathbf{x}^\star) = A\mathbf{x} - A\mathbf{x}^\star = \mathbf{b} - \mathbf{b} = \mathbf{0}$ and $\mathbf{x} - \mathbf{x}^\star = \mathbf{y} \in \text{Null}(A)$, which proves one direction of equality. For the other direction, $A(\mathbf{x}^\star + \mathbf{y}) = A\mathbf{x}^\star + A\mathbf{y} = \mathbf{b} + \mathbf{0} = \mathbf{b}$ for $\mathbf{y} \in \text{Null}(A)$. ■

3.2 Solving Linear Systems

Let us revisit the procedure to solve a linear system in Section 2.6. If we interpret $A\mathbf{x}$ as a linear combination of column vectors of A, we can regard solving $A\mathbf{x} = \mathbf{b}$

as finding a linear combination that matches $\mathbf{b}$. Therefore, more independent columns in the coefficient matrix A in $A\mathbf{x} = \mathbf{b}$ imply more $\mathbf{b}$'s for which there exist solutions.[2] An invertible matrix can be thought of as a matrix with the maximal number of independent columns. In this case, the linear system has a unique solution for any $\mathbf{b}$, i.e., $\mathbf{x} = A^{-1}\mathbf{b}$. On the other hand, if there exists $\mathbf{0} \neq \mathbf{y} \in \text{Null}\,(A)$,[3] there may be no solution or infinitely many solutions to the linear system depending on the choice of $\mathbf{b}$, according to Fact 3.2. In general, if the column space is a strict subspace of $\mathbb{R}^m$ (i.e., $\text{Col}\,(A) \subsetneq \mathbb{R}^m$), the linear system has a solution only when $\mathbf{b} \in \text{Col}\,(A)$.

In order to solve $A\mathbf{x} = \mathbf{b}$, we repeatedly eliminate coefficients of the linear system by adding/subtracting an equation to/from another equation, resulting in increasingly more zero entries in the coefficient matrix A. Gaussian elimination is where we do so by incrementally turning top-left coefficients of linear system to zeros. In the resulting coefficient matrix, all entries below a zeroed-out entry are all zeros, and we call a matrix in such a form a row echelon form.

3.2.1 Row Echelon Form

Let us perform the Gaussian elimination on

$$A = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 7 \\ -1 & -3 & 3 & 4 \end{bmatrix}.$$

- First pivoting: We use $\tilde{L}_1$ to eliminate all entries of the first column of A except for a_{11}.

$$\tilde{L}_1 A = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \textcircled{1} & 3 & 3 & 2 \\ 2 & 6 & 9 & 7 \\ -1 & -3 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 6 & 6 \end{bmatrix}$$

- Second pivoting: We use $\tilde{L}_2$ to eliminate all entries of the third column of $\tilde{L}_1 A$ below $(\tilde{L}_1 A)_{23}$.

$$\tilde{L}_2 \tilde{L}_1 A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & \textcircled{3} & 3 \\ 0 & 0 & 6 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} = U$$

[2] We will define and discuss the notion of independent vectors later in Definition 3.4.

[3] In other words, the columns of A are related to each other in a non-trivial way, satisfying $A\mathbf{y} = y_1\mathbf{a}_1 + \cdots + y_n\mathbf{a}_n = \mathbf{0}$.

The matrices multiplied from left are called elementary matrices, and they are lower triangular. We use pivots to refer to the entries that were used to eliminate the others in the corresponding columns. In the above example, $(A)_{11}$ and $(\tilde{L}_1 A)_{23}$ are pivot entries.

Some properties of the pivot entries are

1. Pivots are the first non-zero entries in their rows;
2. Below each pivot is a column of zeros after elimination;
3. Each pivot lies to the right of the pivot in the row above;
4. A pivot entry does not need to be 1.

After Gaussian elimination, we end up with an upper triangular matrix with all-zero rows at the bottom of the matrix. We call this form of a matrix a row echelon form.

We get the following matrix by multiplying all the elementary matrices above in order:

$$\tilde{L} = \tilde{L}_2 \tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 5 & -2 & 1 \end{bmatrix}.$$

According to Theorem 2.1, this matrix is invertible, and its inverse is

$$L = \tilde{L}^{-1} = \tilde{L}_1^{-1} \tilde{L}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix}.$$

Because $\tilde{L}A = U$ and thus $L^{-1}A = U$, we can write A as

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \text{lower triangular} \times \text{row echelon form.}$$

If it is necessary to swap rows during Gaussian elimination, we can apply permutation of rows and still end up with a row echolon form of a permuted matrix. This leads to the following result:

Fact 3.3 *For any $m \times n$ matrix A, there exist a permutation matrix Q, a lower triangular square matrix L with a unit diagonal, and an $m \times n$ echelon matrix U which is upper triangular, such that $QA = LU$.*

A Reduced Row Echelon Form

In a row echelon form, a pivot may not be 1, and entries above a pivot entry may be not 0. We can impose these conditions by first multiplying a matrix in row echelon form from left by an appropriate diagonal matrix to scale all pivots to be 1 and next multiplying the resulting matrix from left again with an appropriate upper triangular matrix to eliminate all non-zero entries above the pivot entries. We call the resulting matrix to be in a reduced row echelon form.[4]

Here, we try to obtain a reduced row echelon form of

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

by additional scaling and elimination.

- Scaling the second row: we scale the pivot 3 to 1 by multiplying it with a diagonal matrix from left.

$$\tilde{D}\,L^{-1}QA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- Eliminating non-zero entries above the pivot elements:

$$\tilde{U}\,\tilde{D}L^{-1}QA = \begin{bmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & \textcircled{1} & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = R$$

We can rearrange all the steps taken so far, i.e. $\tilde{U}\tilde{D}L^{-1}QA = R$, as

$$QA = L\tilde{D}^{-1}\tilde{U}^{-1}R = LDUR, \tag{3.1}$$

where every matrix multiplied to A (or modified A), including Q, L, D and U, is invertible. Recall that the product of two upper triangular matrix, UR, is also an upper triangular matrix due to Fact 2.1.

We can illustrate the matrix entries of a row echelon form and a reduced row echelon form as:

[4] A reduced row echelon form is unique up to permutation after Gaussian elimination.

$$U = \begin{bmatrix} \bullet & * & * & * & * & * & * \\ 0 & \bullet & * & * & * & * & * \\ 0 & 0 & 0 & 0 & \bullet & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & \bullet \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[\Longrightarrow]{\text{scaling and reduction}} R = \begin{bmatrix} 1 & 0 & *' & *' & 0 & *' & 0 \\ 0 & 1 & *' & *' & 0 & *' & 0 \\ 0 & 0 & 0 & 0 & 1 & *' & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3.2.2 Pivot Variables and Free Variables

We call variables (elements) in $\mathbf{x}$ that correspond to the rows in R (in a reduced row echelon form) that contain pivot entries pivot variables, and the rest of the variables in $\mathbf{x}$ free variables. Among many different ways to find a solution to $R\mathbf{x} = 0$, one systematic way is to assign (literally) arbitrary values to free variables and determine the values of pivot variables.

Let us continue from the previous example

$$\begin{bmatrix} \textcircled{1} & 3 & 0 & -1 \\ 0 & 0 & \textcircled{1} & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

where u and w are pivot variables, and v and y are free variables. We can express the pivot variables as functions of the free variables, as follows

$$\begin{cases} u &= -3v &+y \\ w &= &-y \end{cases}$$

We can therefore readily determine the values of the pivot variables once we assign arbitrary values to the free variables. Although we will only define the notion of dimension rigorously in Definition 3.7, we can imagine that the number of free variables is the dimension of the solution space of $R\mathbf{x} = \mathbf{0}$.

We can derive a few interesting properties.

- $A\mathbf{x} = \mathbf{0} \iff R\mathbf{x} = \mathbf{0}$: Because Q, L, D and U are all invertible in (3.1), these two homogeneous systems are equivalent.

- We can express a vector in Null (A) by replacing each pivot variable with

its equivalent expression in terms of free variables, as follows

$$\begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \begin{bmatrix} -3v + y \\ v \\ -y \\ y \end{bmatrix} = v \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + y \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

This can be thought of as a 2-dimensional plane in the 4-dimensional Euclidean space, $\mathbb{R}^4$, geometrically. We can also express it as the column space; Col $\begin{bmatrix} -3 & 1 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}$, where the first column is a solution to $A\mathbf{x} = \mathbf{0}$ given $v = 1$ and $y = 0$, and the second column given $v = 0$ and $y = 1$.

By inspecting these representations based on pivot and free variables, we make the following observations. Let A be an $m \times n$ matrix. Then,

- the number of pivots $\leq \min(m, n)$;
- if $n > m$, there exists at least one free variable, and $A\mathbf{x} = \mathbf{0}$ has at least one non-zero solution.

Because it will be useful in later sections, we summarize the last property above into the following lemma.

Lemma 3.2 *If a matrix A has more columns than rows, $A\mathbf{x} = \mathbf{0}$ has a non-zero solution. Equivalently, if $A\mathbf{x} = \mathbf{0}$ does not have any non-zero solution, then A has at least as many rows as there are columns.*

Solving $A\mathbf{x} = \mathbf{b}$, $U\mathbf{x} = \mathbf{c}$ and $R\mathbf{x} = \mathbf{d}$

We are now ready to take the final step of solving a linear system. We assume that we can obtain a row echelon form without swapping rows, that is, there is no need to multiply a permutation matrix to proceed with Gaussian elimination.

First, we multiply both sides of $A\mathbf{x} = \mathbf{b}$ with a lower triangular matrix L^{-1} to obtain a linear system expressed in terms of a row-echelon-form coefficient matrix, i.e.

$$L^{-1}(A\mathbf{x}) = L^{-1}(\mathbf{b}) \quad \iff \quad U\mathbf{x} = \mathbf{c}.$$

Consider the example above

$$A\mathbf{x} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 7 \\ -1 & -3 & 3 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \mathbf{b}\,.$$

This corresponds to

$$L^{-1}\,A\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 5 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 7 \\ -1 & -3 & 3 & 4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 5 & -2 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = L^{-1}\mathbf{b},$$

resulting in

$$U\mathbf{x} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} b_1 \\ b_2 - 2b_1 \\ b_3 - 2b_2 + 5b_1 \end{bmatrix} = \mathbf{c}.$$

We therefore see that there exists a solution to $U\mathbf{x} = \mathbf{c}$ (equivalently, $A\mathbf{x} = \mathbf{b}$) if and only if $b_3 - 2b_2 + 5b_1 = 0$.

With $b_3 - 2b_2 + 5b_1 = 0$, we can rewrite the linear system as

$$\begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} b_1 \\ b_2 - 2b_1 \\ 0 \end{bmatrix}.$$

By setting all free variables to 0 ($v = 0$ and $y = 0$), we get a particular solution $\mathbf{x}_p = (3b_1 - b_2, 0, \frac{1}{3}b_2 - \frac{2}{3}b_1, 0)^\top$, because $3w = b_2 - 2b_1$, $w = \frac{1}{3}b_2 - \frac{2}{3}b_1$ and $u + 3w = b_1$, $u = 3b_1 - b_2$.

Now we solve the following system, which is equivalent to the homogeneous system, $A\mathbf{x} = \mathbf{0}$, of the original linear system:

$$\begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ y \end{bmatrix} = \mathbf{0}.$$

As have seen earlier, we get the following two independent solutions:

- $v = 1, y = 0$: $\mathbf{x}_1 = (-3, 1, 0, 0)^\top$
- $v = 0, y = 1$: $\mathbf{x}_2 = (1, 0, -1, 1)^\top$

We can then express any solution with the corresponding α and β as

$$\mathbf{x} = \mathbf{x}_p + \alpha\mathbf{x}_1 + \beta\mathbf{x}_2 = \begin{bmatrix} 3b_1 - b_2 \\ 0 \\ \frac{1}{3}b_2 - \frac{2}{3}b_1 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

We encourage you to show that $A\mathbf{x} = \mathbf{b}$ for this $\mathbf{x}$ yourself.

To summarize, assume U has r pivots when U is a row echelon form of A. That is, it satisfies $U\mathbf{x} = \mathbf{c}$ which is equivalent to the original system $A\mathbf{x} = \mathbf{b}$. Because the last $(m - r)$ rows of U are all zeros, the last $(m - r)$ elements of $\mathbf{c}$ must be all zeros as well, for the linear system to have a solution. If so, $(n - r)$ elements in $\mathbf{x}$ are free variables.

U may look different as we swap rows during Gaussian elimination. The number of pivots in a matrix is however maintained and is called the matrix's rank. See Section 3.4 for more details.

3.3 Linear Independence, Basis, and Dimension

We introduce concepts of linear independence, spanning a subspace, a basis for a subspace, and the dimension of a subspace, which are fundamental to linear algebra.

Linear Independence

When we refer to a set of vectors as linearly independent vectors, we are saying that it is a minimal set of non-redundant vectors. More formally,

Definition 3.4 *For vectors $\mathbf{v}_i \in \mathbb{V}$ and scalars c_i, suppose $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$ holds only when $c_1 = \cdots = c_n = 0$. Then, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent. If a linear combination vanishes for some non-zero c_i's, then $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly dependent, and some $\mathbf{v}_i$ can be represented as a linear combination of the others.*

Based on this definition, it is important for you to understand the following properties and examples.

- $\{\mathbf{0}\}$ is linear dependent.

- $A = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} = [\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \mathbf{a}_3 \,|\, \mathbf{a}_4]$: Columns of A are linearly dependent,
  because $(-3)\mathbf{a}_1 + 1\mathbf{a}_2 + 0\mathbf{a}_3 + 0\mathbf{a}_4 = \mathbf{0}$. Rows of A are also linearly dependent, because if we denote $B = [\mathbf{b}_1 \,|\, \mathbf{b}_2 \,|\, \mathbf{b}_3] = A^\top$, $5\mathbf{b}_1 + (-2)\mathbf{b}_2 + 1\mathbf{b}_3 = \mathbf{0}$.
- $A = \begin{bmatrix} 3 & 4 & 2 \\ 0 & 1 & 5 \\ 0 & 0 & 2 \end{bmatrix}$: Columns of A are linearly independent.
- $A = [\,\mathbf{a}_1 | \mathbf{a}_2 | \cdots | \,\mathbf{a}_n]$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$:

$$\begin{aligned} &\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\} \text{ is linearly independent} \\ &\quad \Leftrightarrow \quad x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{0} \text{ implies } \mathbf{x} = \mathbf{0} \\ &\quad \Leftrightarrow \quad A\mathbf{x} = \mathbf{0} \text{ implies } \mathbf{x} = \mathbf{0} \\ &\quad \Leftrightarrow \quad \mathrm{Null}\,(A) = \{\mathbf{0}\} \end{aligned}$$

- Non-zero rows of a row-echelon-form U are linearly independent. Similarly, columns containing pivots are linearly independent.
- A set of n vectors in $\mathbb{R}^m$ is always linearly dependent if $n > m$ because of Lemma 3.2.

Spanning a Subspace

When a number of vectors can express each and every vector in a vector space as their linear combination, we say these vectors span the vector space.

Definition 3.5 *For vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, their span is a minimal subspace containing $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and is described formally as*

$$\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \left\{c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n : c_i \in \mathbb{R}\right\}.$$

If $\mathbb{V} = \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then we say that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ spans $\mathbb{V}$.

Here are two oft-used vector spaces spanned by a finite set of vectors:

- For $A = [\,\mathbf{a}_1 | \mathbf{a}_2 | \cdots | \,\mathbf{a}_n]$, $\mathrm{Col}\,(A) = \mathrm{span}\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$.
- For $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ where the i-th entry is 1, $\mathbb{R}^n = \mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

Spanning vectors need not be linearly independent. When they are not linearly independent, there are many different ways to linearly combine spanning vectors to represent each vector in the spanned space. Moreover, linearly independent vectors span a vector space with a unique linear combination for each vector within. Fact 3.4 below clarifies this point.

Fact 3.4 *Assume that* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *is linearly independent. If a vector* $\mathbf{v}$ *can be represented as a linear combination of these vectors, i.e.,* $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$*, the coefficients* x_i*'s are unique.*

Proof: Suppose that $\mathbf{v} = y_1\mathbf{v}_1 + \cdots + y_n\mathbf{v}_n$ holds for some scalars y_i's. If we subtract the latter from the former, we get

$$(x_1 - y_1)\mathbf{v}_1 + \cdots + (x_n - y_n)\mathbf{v}_n = \mathbf{0}\,.$$

Because of the linear independence, $x_i - y_i = 0$, or equivalently $x_i = y_i$, for all $i = 1, \ldots, n$. ■

We can further observe that any vector outside a subspace spanned by linearly independent vectors is linearly independent of the spanning vectors.

Fact 3.5 *If* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *is linearly independent and* $\mathbf{v} \notin \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$*,* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{v}\}$ *is also linearly independent.*

Proof: Notice that $\mathbf{v} \neq \mathbf{0}$. Suppose that $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n + c_{n+1}\mathbf{v} = \mathbf{0}$ holds for some scalars c_i's. If $c_{n+1} \neq 0$, then we have

$$\mathbf{v} = -\frac{c_1}{c_{n+1}}\mathbf{v}_1 - \cdots - \frac{c_n}{c_{n+1}}\mathbf{v}_n \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}\,,$$

which contradicts the assumption. Hence $c_{n+1} = 0$, and the linear independence of spanning vectors implies $c_1 = \cdots = c_n = 0$. ■

Basis for a Vector Space

The independence property is about the minimality, and the spanning property is about the sufficiency. Then a natural question is on minimally sufficient vectors to span a space.

Definition 3.6 *A basis for a vector space $\mathbb{V}$ is a set of vectors satisfying both of the following properties:*

1. **(independence)** *The vectors in the set are linearly independent;*
2. **(spanning)** *The vectors in the set span the space $\mathbb{V}$.*

A vector in a basis is called a basic vector.

If linearly dependent vectors span a vector space, we can always find to span the same vector space. On the other hand, if some linearly independent vectors do not span a target vector space, we can incrementally add linearly independent vectors one at a time until they are sufficient to span the target space, due to Fact 3.5.

- Because there is a unique linear combination to represent an arbitrary vector in a vector space using basic vectors, we can treat the coefficient of such linear combination as a coordinate. That is, given a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, we can represent an arbitrary vector $\mathbf{v} \in \mathbb{V}$ uniquely as

$$\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n \tag{3.2}$$

  where x_i's are coefficients of $\mathbf{v}$ with respect to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. So, we may conveniently regard $(x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ as $\mathbf{v}$.

  For this correspondence to hold, it should be one-to-one. We see that this is true from the spanning property, which states that any arbitrary vector can be represented as a linear combination of basic vectors, and the independence property, which states that such representation is unique. In summary, when $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent,

$$\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \overset{1\text{-to-}1}{\Longleftrightarrow} (x_1, \ldots, x_n) \in \mathbb{R}^n\,. \tag{3.3}$$

- There can be many bases for a vector space. When $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^n$ is a basis of the vector space $\mathbb{R}^n$, let $B = [\mathbf{v}_1 | \cdots | \mathbf{v}_n]$. For any arbitrary invertible matrix P, we obtain a new basis for the vector space by taking the column vectors of BP. For instance, if we multiply $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, whose column vectors span $\mathbb{R}^2$, i.e., $\mathbb{R}^2 = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$, with an

invertible matrix $P = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$, we get $BP = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}$, implying $\mathbb{R}^2 = \text{span}\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\}$.

Example 3.1 When $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis of a vector space $\mathbb{V}$, consider

$$\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3, \ldots, \mathbf{v}_1 + \cdots + \mathbf{v}_n\}.$$

Because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis, $c_n = c_{n-1} = \cdots = c_1 = 0$ when

$$(c_1 + \cdots + c_n)\mathbf{v}_1 + (c_2 + \cdots + c_n)\mathbf{v}_2 + \cdots + (c_{n-1} + c_n)\mathbf{v}_{n-1} + c_n\mathbf{v}_n = \mathbf{0}.$$

By rearranging the terms, we get

$$c_1\mathbf{v}_1 + c_2(\mathbf{v}_1 + \mathbf{v}_2) + \cdots + c_n(\mathbf{v}_1 + \cdots + \mathbf{v}_n)$$

and thus, $\mathcal{B}$ is linearly independent. Furthermore, for any i,

$$\mathbf{v}_i = (\mathbf{v}_1 + \cdots + \mathbf{v}_i) - (\mathbf{v}_1 + \cdots + \mathbf{v}_{i-1}) \in \text{span}\,\mathcal{B}$$

and, hence $\mathbb{V} \subset \text{span}\,\mathcal{B}$ holds. Therefore, $\mathcal{B}$ is another basis for $\mathbf{V}$. ∎

Dimension of a Vector Space

Although there are many bases for a vector space, the number of basic vectors within each basis remains identical, according to the following theorem.

Theorem 3.1 *If both* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *and* $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ *are bases for a vector space* $\mathbb{V}$*, then* $n = m$*.*

Proof: Since both sets of vectors span $\mathbb{V}$, there exist a_{ij}'s such that

$$\mathbf{v}_j = a_{1j}\mathbf{w}_1 + \cdots + a_{mj}\mathbf{w}_m = \sum_{i=1}^{m} a_{ij}\mathbf{w}_i \quad \text{for all } j = 1, \ldots, n\,.$$

Let us set an $m \times n$ matrix $A = (a_{ij})$. Assume $A\mathbf{x}^* = \mathbf{0}$ for some vector $\mathbf{x}^* \in \mathbb{R}^n$. Then, $\sum_{j=1}^{n} x_j^* a_{ij} = 0$ for all $i = 1, \ldots, m$. For this $\mathbf{x}^*$,

$$\sum_{j=1}^{n} x_j^* \mathbf{v}_j \quad = \quad \sum_{j=1}^{n} x_j^* \Big(\sum_{i=1}^{m} a_{ij}\mathbf{w}_i \Big) = \sum_{j=1}^{n}\sum_{i=1}^{m} x_j^* a_{ij}\mathbf{w}_i = \sum_{i=1}^{m}\sum_{j=1}^{n} x_j^* a_{ij}\mathbf{w}_i$$

$$= \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} x_j^* a_{ij} \Big) \mathbf{w}_i = \mathbf{0},$$

which implies $x_j^* = 0$ for all j since $\mathbf{v}_i$'s are linearly independent, that is, $\mathbf{x}^* = \mathbf{0}$. Therefore, $m \geq n$ by Lemma 3.2. If we change the roles of $\mathbf{v}_i$ and $\mathbf{w}_i$, then we have $n \geq m$. ■

This allows us to use the number of basic vectors in a basis to quantify the size of a vector space.

Definition 3.7 *The dimension of a vector space $\mathbb{V}$ is the number of basic vectors in its basis.*

We encourage you to think of the following properties and why they hold.

- $\dim(\mathbb{R}^n) = n$.
- $(k+1)$ vectors in a k-dimensional vector space are linearly dependent.
- Any spanning set of vectors can be reduced to a basis, i.e., a minimal spanning set.

A few observations follow.

Lemma 3.3 *In a finite-dimensional vector space, any linearly independent set of vectors can be extended to a basis.*

Proof: Consider k linearly independent vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbb{V}$, where $\mathbb{V}$ is finite-dimensional. Let $k < \dim \mathbb{V} < \infty$. If $\mathbb{V} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, this is contradictory, as $\dim \mathbb{V} = k$. Thus, there exists $\mathbf{v} \in \mathbb{V}$ that satisfies $\mathbf{v} \notin \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. According to Fact 3.5, $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}\}$ is a set of $(k+1)$ linearly independent vectors, that include $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. We repeat this process of adding one vector at a time, until we obtain the basis of $\mathbb{V}$ that contains all the initial linearly independent vectors. ■

Fact 3.6 *Let $\mathbb{V}$ be a finite-dimensional vector space. $\mathbb{W}_1$ and $\mathbb{W}_2$ are two subspaces of $\mathbb{V}$. Then,*

$$\dim(\mathbb{W}_1 \cap \mathbb{W}_2) \geq \dim \mathbb{W}_1 + \dim \mathbb{W}_2 - \dim \mathbb{V}$$

holds.

Proof: Denote $\dim \mathbb{V} = n$, $\dim \mathbb{W}_1 = n_1$, and $\dim \mathbb{W}_2 = n_2$. Let $\dim(\mathbb{W}_1 \cap \mathbb{W}_2) = k$ and $\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis of $\mathbb{W}_1 \cap \mathbb{W}_2$. Because $\mathcal{B}$ is linearly independent, we can find the bases of $\mathbb{W}_1$ and $\mathbb{W}_2$, respectively, according to Lemma 3.3. We use $\mathcal{B} \cup \{\mathbf{w}_{k+1}, \ldots, \mathbf{w}_{n_1}\}$ and $\mathcal{B} \cup \{\mathbf{u}_{k+1}, \ldots, \mathbf{u}_{n_2}\}$ to denote their bases.

- Let $\mathbf{u} = z_{k+1}\mathbf{u}_{k+1} + \cdots + z_{n_2}\mathbf{u}_{n_2} \in \mathbb{W}_1 \cap \mathbb{W}_2$. Since $\mathcal{B}$ is a basis of $\mathbb{W}_1 \cap \mathbb{W}_2$,

  $$\mathbf{u} = x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k \quad \text{or} \quad z_{k+1}\mathbf{u}_{k+1} + \cdots + z_{n_2}\mathbf{u}_{n_2} - x_1\mathbf{v}_1 - \cdots - x_k\mathbf{v}_k = \mathbf{0}.$$

  Because $\mathcal{B} \cup \{\mathbf{u}_{k+1}, \ldots, \mathbf{u}_{n_2}\}$ is a basis, $x_1 = \cdots = x_k = z_{k+1} = \cdots z_{n_2} = 0$, and $\mathbf{u} = \mathbf{0}$.

- Consider the following zero-vector

  $$\underbrace{x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k}_{=\,\mathbf{v}} + \underbrace{y_{k+1}\mathbf{w}_{k+1} + \cdots + y_{n_1}\mathbf{w}_{n_1}}_{=\,\mathbf{w}} + \underbrace{z_{k+1}\mathbf{u}_{k+1} + \cdots + z_{n_2}\mathbf{u}_{n_2}}_{=\,\mathbf{u}}$$

  By re-arranging the terms, we get $\mathbf{u} = -(\mathbf{v} + \mathbf{w}) \in \mathbb{W}_1$, which implies $\mathbf{u} \in \mathbb{W}_1 \cap \mathbb{W}_2$. As we have shown above, $\mathbf{u} = \mathbf{0}$, which is equivalent to $z_{k+1} = \cdots = z_{n_2} = 0$. Together with the fact that $\mathcal{B} \cup \{\mathbf{w}_{k+1}, \ldots, \mathbf{w}_{n_1}\}$ is a basis, $x_1 = \cdots = x_k = y_{k+1} = \cdots = y_{n_1} = 0$, since $\mathbf{v} + \mathbf{w} = \mathbf{0}$. In other words, $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_{n_1}, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_{n_2}\}$ consists of $(n_1 + n_2 - k)$ linearly independent vectors, which implies $n_1 + n_2 - k \leq n$.

Therefore, $\dim(\mathbb{W}_1 \cap \mathbb{W}_2) \geq n_1 + n_2 - n$. ■

3.4 Rank of a Matrix

How many linearly independent columns do we get in a matrix A with r pivot elements? To answer this question, let us start with a row echelon form U of the matrix A. U has r pivots, and assume that there are at most M linearly independent columns in U.

- Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ contain r columns of U, including its pivots. For the convenience, we let the pivot of $\mathbf{u}_i$ be the i-th entry, say p_i. The r-th equation in the linear system, $x_1\mathbf{u}_1 + \cdots + x_r\mathbf{u}_r = \mathbf{0}$, is $0x_1 + \cdots + 0x_{r-1} + p_r x_r = 0$. Because the pivot p_r of $\mathbf{u}_r$ is not 0, x_r vanishes, which allows us to shorten the equation into $x_1\mathbf{u}_1 + \cdots + x_{r-1}\mathbf{u}_{r-1} = \mathbf{0}$ in $r-1$ unknowns. By repeating this process, we end up with $x_1 = \cdots = x_r = 0$, and therefore $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is linearly independent. That is, $M \geq r$.

- Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ be a set of k arbitrarily selected columns of U, with $k > r$. Since the r pivots lie in the first r rows of U, all $m-r$ entries below in each $\mathbf{u}_i$ are zeros. With this, we see that $x_1\mathbf{u}_1 + \cdots + x_k\mathbf{u}_k = \mathbf{0}$, a linear system of m equations, is in fact a system of r linear equations in k variables. There are thus solutions that are not $\mathbf{0}$, according to Lemma 3.2, which implies that more than r columns of U can not be linearly independent. Thus, $M \leq r$ holds.

Therefore, $M = r$, and the maximum number of linearly independent columns in U, a row echelon form of A, is r.

Let us continue with the original matrix A. Because Q and $\tilde{L}$, in $\tilde{L}QA = U$, are both invertible, the following equivalence holds

$$A\mathbf{x} = \mathbf{0} \iff U\mathbf{x} = \mathbf{0}.$$

According to this equivalence, the relationship among the column vectors of A, that is whether they are linearly independent, holds the same among the column vectors of U as well. From this, we now know that the maximum number of linearly independent columns of a matrix coincides with the number of pivots of the same matrix. We call this number the rank of a matrix.

Definition 3.8 (Rank of a Matrix) *Let U be the row echelon form of a matrix A. If U has r pivots, then the rank of A is r. We denote it as* $\operatorname{rank} A$.

Because both U, the row echelon form of A, and R, the reduced row echelon form of A, are row echelon forms of themselves, respectively, the ranks of A, U and R are same.

We turn our attention to the maximum number of linearly independent rows of A and its relation to the rank. Because A and QA are equivalent up to the ordering of the rows, without loss of generality, it is enough to consider the case of $Q = I$. That is, we consider the case of $A = LU$. Let $\operatorname{rank} A = r$, i.e., there are r pivots in A. It is clear that the r rows on the top of the row echelon form U are linearly independent, and the rest of the rows are all zeros. Therefore, U has exactly r linearly independent rows.

Let us create $\hat{A}$ and $\hat{U}$ from A and U, respectively, by collecting only the first r rows of these matrices. We also create an invertible lower triangular matrix L_r from L by collecting the first r rows and r columns. These matrices are related

to each other by $\hat{A} = L_r\hat{U}$.[5] Since the rows of $\hat{U}$ are linearly independent, $\mathbf{y}^\top\hat{A} = \mathbf{y}^\top L_r\hat{U} = \mathbf{0}$ implies $\mathbf{y}^\top L_r = \mathbf{0}$. Because L_r is invertible, $\mathbf{y} = \mathbf{0}$. In short, there are at least r linearly independent rows in A, because the first r rows of A are linearly independent.

Let us consider the other direction. We can see that $A = LU = \tilde{L}\hat{U}$ where $\tilde{L}$ is created by collecting the first r columns of L, because the last $(m-r)$ rows of U are all zeros. With $B = \tilde{L}L_r^{-1}$, we get

$$A = \tilde{L}\hat{U} = \tilde{L}(L_r^{-1}L_r)\hat{U} = \tilde{L}L_r^{-1}\hat{A} = B\hat{A}.$$

Let us show that k rows of A are linearly dependent for $k > r$. Select k arbitrary rows of A and call this $k \times n$ matrix A'. B' is constructed by selecting k corresponding rows from B. Then, $A' = B'\hat{A}$ holds. Because B' is a $k \times r$ matrix, there exists $\mathbf{y} \neq \mathbf{0}$ that satisfies $\mathbf{y}^\top B' = \mathbf{0}^\top$ according to Lemma 3.2, and thereby $\mathbf{y}^\top A' = \mathbf{y}^\top B'\hat{A} = \mathbf{0}^\top\hat{A} = \mathbf{0}^\top$, implying that the rows in A' are not linearly independent. There are therefore at most $r = \operatorname{rank} A$ linearly independent rows in A.

Putting these two parts together, we arrive at the conclusion that $\operatorname{rank} A$ is the maximum number of linearly independent rows or columns of A. We can further conclude that

$$\operatorname{rank} A = \operatorname{rank} A^\top. \tag{3.4}$$

Think of a subspace $\operatorname{Col}(A)$ spanned by the columns of a rank-r $m \times n$ matrix $A = [\mathbf{a}_1 | \cdots | \mathbf{a}_n]$, with the first r column vectors being linearly independent. With $k > r$, $\{\mathbf{a}_1, \ldots, \mathbf{a}_r, \mathbf{a}_k\}$ is not linearly independent, which means there exist x_i's that satisfy $x_1\mathbf{a}_1 + \cdots + x_r\mathbf{a}_r + x_k\mathbf{a}_k = \mathbf{0}$ with $x_k \neq 0$. Then, $\operatorname{Col}(A) \subset \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$, because

$$\mathbf{a}_k = -\frac{x_1}{x_k}\mathbf{a}_1 - \cdots - \frac{x_r}{x_k}\mathbf{a}_r \in \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}.$$

That is, $\operatorname{Col}(A) = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$, and $\dim \operatorname{Col}(A) = \operatorname{rank} A$. In summary,

Lemma 3.4 *For* $\mathbf{a}_i \in \mathbb{R}^m$,

$$\dim \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \operatorname{rank}[\mathbf{a}_1 | \cdots | \mathbf{a}_n]$$

or, in a matrix form,

$$\dim \operatorname{Col}(A) = \operatorname{rank} A$$

for any matrix A.

[5] This holds because the i-th row of LU is the linear combination of the first $(i-1)$ rows of U when L is a lower triangular matrix with a unit diagonal.

This applies equally to the vector space spanned by the rows of A, as $\dim \mathrm{Col}\,(A^\top) = \mathrm{rank}\, A^\top = \mathrm{rank}\, A$. Also, the multiplication of matrices does not increase the rank of a product.

Fact 3.7 *Suppose A and B are $n \times m$ and $m \times n$ matrices, respectively. Then,* $\mathrm{rank}(AB) \leq \mathrm{rank}(A)$ *and* $\mathrm{rank}(AB) \leq \mathrm{rank}(B)$.

Proof: According to Lemma 3.1, $\mathrm{Col}\,(AB) \subset \mathrm{Col}\,(A)$. When two vector spaces, $\mathbb{W}_1$ and $\mathbb{W}_2$, satisfy $\mathbb{W}_1 \subset \mathbb{W}_2$, we can establish the relationship between their dimensions as $\dim \mathbb{W}_1 \leq \dim \mathbb{W}_2$. It then follows that $\mathrm{rank}(AB) \leq \mathrm{rank}\, A$ due to Lemma 3.4. The second inequality follows from this, because $\mathrm{rank}(AB) = \mathrm{rank}((AB)^\top) = \mathrm{rank}(B^\top A^\top) \leq \mathrm{rank}(B^\top) = \mathrm{rank}(B)$. ■

3.5 Four Fundamental Subspaces

We introduced two subspaces related to a matrix, the column space and the null space in Section 3.1.2. Consider these subspaces of the transpose of a given matrix, and we have the following four subspaces related to a rank-r $m \times n$ matrix A:

1. Column space $\mathrm{Col}\,(A) \subset \mathbb{R}^m$. $\dim \mathrm{Col}\,(A) = r$;
2. Null space $\mathrm{Null}\,(A) \subset \mathbb{R}^n$. $\dim \mathrm{Null}\,(A) = n - r$;
3. Row space $\mathrm{Row}\,(A) = \mathrm{Col}\,(A^\top) \subset \mathbb{R}^n$. $\dim \mathrm{Row}\,(A) = r$;
4. Left null space $\mathrm{LeftNull}\,(A) = \mathrm{Null}\,(A^\top) \subset \mathbb{R}^m$. $\dim \mathrm{LeftNull}\,(A) = m - r$.

We already characterized the dimensions of the first and third subspaces in Lemma 3.4. Let us take a look at the dimension of null spaces. From Gaussian elimination, we showed $A\mathbf{x} = \mathbf{0} \;\Leftrightarrow\; U\mathbf{x} = \mathbf{0}$, which implies that $\mathrm{Null}\,(A) = \mathrm{Null}\,(U)$. By assigning 1 to one free variable in $U\mathbf{x} = \mathbf{0}$ and 0 to all the other free variables, we get as many linearly independent vectors in the null space as there are free variables, and they form a basis of the null space. Therefore, $\dim \mathrm{Null}\,(A) = n - r$, because $\dim \mathrm{Null}\,(U)$ coincides with the number of free variables of U which is $n - r$. From these, we arrive at the rank-nullity theorem:

$$\mathrm{rank}(A) + \dim \mathrm{Null}\,(A) \quad = \quad \dim \mathrm{Col}\,(A) + \dim \mathrm{Null}\,(A)$$

$$= \quad \text{the number of columns of } A. \tag{3.5}$$

Example 3.2 Let us find the four fundamental subspaces of $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = U = R$. The four subspaces can be written down for this simple matrix, as follows.

1. Column space: $\text{Col}\,(A) = \text{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$.
2. Null space: $\text{Null}\,(A) = \text{span}\left\{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right\}$.
3. Row space: $\text{Row}\,(A) = \text{Col}\,(A^\top) = \text{span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}$.
4. Left null space: $\text{LeftNull}\,(A) = \text{Null}\,(A^\top) = \text{span}\left\{\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$.

∎

Illustrating the Construction of Fundamental Spaces from a Row Echelon Form

It is often more straightforward to identify the four fundamental subspaces of A from its row echelon form U. We study how we can do so in the following example. Let $A = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 7 \\ -1 & -3 & 3 & 4 \end{bmatrix}$. Then, its row echelon form is $U = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

1. Finding the Row space $\text{Col}\,(A^\top)$: The basis of $\text{Col}\,(A^\top)$ consists of nonzero rows of U, that is, $\text{Col}\,(A^\top) = \text{Col}\,(U^\top)$. This follows from Lemma 3.1 and the fact that L is invertible and $A^\top = U^\top L^\top$.

2. Finding the Column space $\mathrm{Col}\,(A)$: A basis of $\mathrm{Col}\,(U)$ consists of columns that contain pivot entries in U. Because $A\mathbf{x} = \mathbf{0} \Leftrightarrow U\mathbf{x} = \mathbf{0}$, linear independence of some columns of A is equivalent to linear independence of the corresponding columns of U. From this, we find that $\dim \mathrm{Col}\,(A) = \dim \mathrm{Col}\,(U) = \mathrm{rank}\, A = r$. In other words, we can form a basis of $\mathrm{Col}\,(A)$ by collecting as many linearly independent columns of A as there are pivots in U. As the columns that contain pivots in U are linearly independent, the corresponding columns in A are also linearly independent. That is, these columns form a basis of $\mathrm{Col}\,(A)$. In the example above, they are the first and third columns of A.

3. Finding the null space $\mathrm{Null}\,(A)$: Because $A\mathbf{x} = \mathbf{0} \Leftrightarrow U\mathbf{x} = \mathbf{0}$ from Gaussian elimination, $\mathrm{Null}\,(A) = \mathrm{Null}\,(U)$. The number of free variables in U, which is in turn $\dim \mathrm{Null}\,(U)$, is $n - r$. By assigning 1 to one free variable in $U\mathbf{x} = \mathbf{0}$ and 0 to all the other free variables, we obtain as many linearly independent vectors as free variables, that span the null space, and they form its basis.

How to Find a Basis of a Spanned Subspace in Euclidean Space

One way to find a basis of a vector space spanned by a set of vectors is to stack those vectors row-wise to construct a matrix, perform Gaussian elimination on this matrix and collect all non-zero rows in the row echelon form as a basis. We can also use all pivot columns of a row echelon form of a matrix constructed by horizontally stacking given vectors. Unlike the first method, it is important to notice that we collect the columns of A corresponding to the pivots of U to form a basis.

3.6 Existence of Inverse Matrix

Consider a rank-m, $m \times n$ matrix A which can be written down as $A = LU$ with $m < n$. In this case, there are m pivot columns. We choose an $n \times n$ permutation matrix Q such that we can use the first m columns of UQ to construct an invertible submatrix $\hat{U}$. If it is not necessary to swap columns,

take $Q = I_n$. For instance, we start with

$$L^{-1}A = U = \begin{bmatrix} *' & * & * & * & * & * & \cdots \\ 0 & 0 & *' & * & * & * & \cdots \\ 0 & 0 & 0 & *' & * & * & \cdots \\ 0 & 0 & 0 & 0 & \ddots & * & \vdots \\ 0 & 0 & 0 & 0 & 0 & *' & \cdots \end{bmatrix}$$

where $*'$ stands for a non-zero entry, and multiply it with a permutation matrix to move the second column to the end, as in

$$L^{-1}AQ = UQ = \left[\begin{array}{cccccc|cc} *' & * & * & \dots & * & * & * & \cdots \\ 0 & *' & * & \dots & * & * & * & \cdots \\ 0 & 0 & *' & \dots & * & * & * & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & *' & * & * & \cdots \\ 0 & 0 & \dots & 0 & 0 & *' & * & \cdots \end{array}\right] = \left[\hat{U} \,\middle|\, G\right].$$

Consider any $(n-m) \times m$ matrix H that satisfies $GH = \mathbf{0}$. Then, with

$$C = Q \begin{bmatrix} \hat{U}^{-1} \\ H \end{bmatrix} L^{-1},$$

it holds that

$$AC = (AQ)(Q^{-1}C) = (LUQ)(Q^{-1}C) = L \left[\hat{U} \,\middle|\, G\right] \begin{bmatrix} \hat{U}^{-1} \\ H \end{bmatrix} L^{-1} = LI_mL^{-1} = I_m,$$

because

$$\left[\hat{U} \,\middle|\, G\right] \begin{bmatrix} \hat{U}^{-1} \\ H \end{bmatrix} = \hat{U}\hat{U}^{-1} + G\,H = I_m.$$

Therefore, C is a right-inverse of A. Note that C depends on H and there may be many H's, including $H = \mathbf{0}$, that satisfy $GH = \mathbf{0}$. In other words, there can be many right inverses of A.

If we had to swap rows in the process of Gaussian elimination and can use a permutation matrix Q' to represent all row swaps, we use $Q'A$ instead of A in the derivation above. The right inverse C above then satisfies $Q'AC = I_m$ and thereby $AC = {Q'}^{-1}$, resulting in CQ' being a right inverse of A.

On the other hand, there is no left inverse of A. If there were, there exists B such that $BA = I_n$. Because there are m columns in B, meaning that $\text{rank}(B) \leq$

m and subsequently that $\text{rank}(BA) \leq \text{rank}(B) \leq m$. This is contradictory, since $\text{rank}(I_n) = n > m$.

Here are some additional, useful properties:

- Similarly to the case above, when $\text{rank}(A) = n < m$, left-inverses exist but there is no right inverse. We can confirm it by working with the transpose of the matrix.
- Consider a case where $\text{rank}(A) = m = n$. This is equivalent to the case of $\text{rank}\, A = m < n$ above however without G and H. In that case, the right inverse of A is $C = Q\hat{U}^{-1}L^{-1}$. Because $L^{-1}AQ = \hat{U}$,

$$CA = Q\hat{U}^{-1}L^{-1}A = Q\hat{U}^{-1}(L^{-1}AQ)Q^{-1} = Q\hat{U}^{-1}\hat{U}Q^{-1} = I_m\,,$$

  meaning that C is also the left inverse. That is, C is the inverse of A.

These cases can be summarized into the following theorem.

Theorem 3.2 *For an $m \times n$ matrix A, the inverse of A exists only when $\text{rank}(A) = m = n$. If $\text{rank}(A) = m \leq n$, the right-inverse of A exists. If $\text{rank}(A) = n \leq m$, the left-inverse of A exists.*

The following relationships hold between the existence of right- and left-inverses of A and the characteristics of the solutions for $A\mathbf{x} = \mathbf{b}$. For any $m \times n$ matrix A,

- $\text{rank}(A) = m$ implies the existence of a solution for $A\mathbf{x} = \mathbf{b}$, since $C\mathbf{b}$ is a solution if C is a right-inverse of A. Alternatively, we can also derive the existence from $\text{Col}\,(A) = \mathbb{R}^m$ because the row and column ranks coincide, and there exists at least one solution for $A\mathbf{x} = \mathbf{b}$ since $\mathbf{b} \in \text{Col}\,(A)$.
- $\text{rank}(A) = n$ implies the uniqueness of the solution for $A\mathbf{x} = \mathbf{b}$ if it exists. By multiplying a left-inverse B to both sides from the left, we get $(BA)\mathbf{x} = \mathbf{x} = B\mathbf{b}$ if a solution $\mathbf{x}$ exists. This implies the uniqueness. Alternatively, because $\dim \text{Col}\,(A) = n$, the columns of A are linearly independent, and therefore, the solution to $A\mathbf{x} = \mathbf{b}$ is unique, if it exists.

The rank of a matrix is upper-bounded by the smaller of the numbers of rows and columns. When the rank of a matrix is maximal, we obtain the following additional properties.

Fact 3.8 *Let A be an $m \times n$ matrix.*

1. $\operatorname{rank} A = n$ *case: Then the rank of* $A^\top A$ *is also* n*, and* $A^\top A$ *is invertible.* $(A^\top A)^{-1}A^\top$ *is a left-inverse of* A *and* $A(A^\top A)^{-1}$ *is a right-inverse of* $A^\top$*.*

2. $\operatorname{rank} A = m$ *case: Then the rank of* $AA^\top$ *is also* m*, and* $AA^\top$ *is invertible.* $A^\top(AA^\top)^{-1}$ *is a right-inverse of* A *and* $(AA^\top)^{-1}A$ *is a left-inverse of* $A^\top$*.*

Proof: For $\operatorname{rank} A = n$ case, it is enough to show that the $n \times n$ matrix $A^\top A$ has a trivial nullspace $\{\mathbf{0}\}$, according to the rank-nullity theorem ((3.5)). Assume $A^\top A\mathbf{x} = \mathbf{0}$. By multiplying $\mathbf{x}$ on the both sides of the equation, we get $\mathbf{x}^\top A^\top A\mathbf{x} = 0$. If we denote $\mathbf{y} = A\mathbf{x}$, then $\mathbf{x}^\top A^\top A\mathbf{x} = \mathbf{y}^\top\mathbf{y} = 0$. Since $\mathbf{y}^\top\mathbf{y} = \sum_{i=1}^m y_i^2 = 0$, each $y_i = 0$ and $\mathbf{y} = \mathbf{0}$, that is, $A\mathbf{x} = \mathbf{0}$. Hence $\mathbf{x} \in \operatorname{Null}(A)$. However the dimension balance (3.5) implies $\dim \operatorname{Null}(A) = 0$ and $\operatorname{Null}(A) = \{\mathbf{0}\}$. Therefore $\operatorname{rank} A^\top A = n$ and $A^\top A$ is invertible. It is clear that $(A^\top A)^{-1}A^\top$ is a left-inverse of A and $A(A^\top A)^{-1}$ is a right-inverse of $A^\top$.

For the other case, take $B = A^\top$ and apply the first result to B. ■

Consider the following simple example for finding right inverses and checking their relationships to a left inverse.

Example 3.3 Let $A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}$. Given arbitrary c_{31} and c_{32}, $AC = I_2$, where $C = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \\ c_{31} & c_{32} \end{bmatrix}$. That is, C is a right inverse of A. We already showed earlier that such a matrix does not admit a left inverse. If we multiply A with C from left, we get

$$CA = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \\ c_{31} & c_{32} \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4c_{31} & 5c_{32} & 0 \end{bmatrix}.$$

Regardless of c_{31} and c_{32}, C is not a left inverse.

Because $\operatorname{rank} A = 2$, we can derive a right inverse of A according to Fact 3.8, as follows:

$$C^* = A^\top(AA^\top)^{-1} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}^\top \begin{bmatrix} 1/16 & 0 \\ 0 & 1/25 \end{bmatrix} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \\ 0 & 0 \end{bmatrix}.$$

In this particular right inverse C^*, $c_{31} = c_{32} = 0$, and we refer to this right inverse as the pseudo-inverse of A, as we will learn later in Section 5.8.

3.7 Rank-one Matrices

All columns of a rank-one matrix are scalar multiplications of each other. This fact allows us to represent a rank-one matrix as $\mathbf{u}\mathbf{v}^\top$ with an appropriate choice of $\mathbf{u}$ and $\mathbf{v}$. It is not trivial to check whether the rank of a matrix is one, but it is relatively straightforward to find these two vectors, $\mathbf{u}$ and $\mathbf{v}$, once we know that its rank is one. For instance, it takes effort to tell the rank of the following matrix:

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & 2 & 2 \\ 8 & 4 & 4 \\ -2 & -1 & -1 \end{bmatrix},$$

but once we know its rank is 1, it is not too challenging to find out that we can express this matrix as

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

$\mathbf{u}$ and $\mathbf{v}$ are not uniquely determined, since we can multiply one with a scalar and the other with its reciprocal without changing the resulting matrix. It is sometimes useful to represent an arbitrary matrix or matrix-matrix product in terms of rank-one matrices. Consider $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^n$, which is a vector whose elements are all zeroes except for the i-th element. Consider a matrix A whose i-th column is $\mathbf{a}_i$, such that $A = [\,\mathbf{a}_1 | \mathbf{a}_2 | \cdots | \mathbf{a}_n\,]$. We can express this matrix as a sum of rank-one matrices, where each summand rank-one matrix is $\mathbf{a}_i\mathbf{e}_i^\top$. That is,

$$A = \mathbf{a}_1\mathbf{e}_1^\top + \cdots + \mathbf{a}_n\mathbf{e}_n^\top = \sum_{i=1}^{n} \mathbf{a}_i\mathbf{e}_i^\top. \tag{3.6}$$

We can generalize this further to show that a product of two matrices can also be expressed as the sum of rank-one matrices.

Lemma 3.5 *Assume an $m \times n$ matrix $A = [\,\mathbf{a}_1|\mathbf{a}_2|\cdots|\,\mathbf{a}_n\,]$ for $\mathbf{a}_i \in \mathbb{R}^m$ and an $\ell \times n$ matrix $B = [\,\mathbf{b}_1|\mathbf{b}_2|\cdots|\,\mathbf{b}_n\,]$ for $\mathbf{b}_i \in \mathbb{R}^\ell$. Then,*

$$AB^\top = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i^\top . \tag{3.7}$$

Proof: We first rewrite A and B using (3.6) as $A = \sum_{i=1}^n \mathbf{a}_i \mathbf{e}_i^\top$ and $B = \sum_{i=1}^n \mathbf{b}_i \mathbf{e}_i^\top$, respectively. We then notice that $\mathbf{e}_i^\top \mathbf{e}_j = 1$ when $i = j$ and otherwise 0. Then,

$$\begin{aligned} AB^\top &= \Big(\sum_{i=1}^{n} \mathbf{a}_i \mathbf{e}_i^\top\Big)\Big(\sum_{j=1}^{n} \mathbf{b}_j \mathbf{e}_j^\top\Big)^\top \\ &= \Big(\sum_{i=1}^{n} \mathbf{a}_i \mathbf{e}_i^\top\Big)\Big(\sum_{j=1}^{n} \mathbf{e}_j \mathbf{b}_j^\top\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{a}_i \mathbf{e}_i^\top \mathbf{e}_j \mathbf{b}_j^\top \\ &= \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i^\top . \end{aligned}$$

■

We generalize it even further by considering a case where a diagonal matrix is inserted between A and $B^\top$:

Corollary 3.1 *Assume an $m \times n$ matrix $A = [\,\mathbf{a}_1|\mathbf{a}_2|\cdots|\,\mathbf{a}_n\,]$ for $\mathbf{a}_i \in \mathbb{R}^m$, an $\ell \times n$ matrix $B = [\,\mathbf{b}_1|\mathbf{b}_2|\cdots|\,\mathbf{b}_n\,]$ for $\mathbf{b}_i \in \mathbb{R}^\ell$, and an $n \times n$ diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then,*

$$A\Lambda B^\top = \sum_{i=1}^{n} \lambda_i \mathbf{a}_i \mathbf{b}_i^\top . \tag{3.8}$$

Proof: Because $A\Lambda = [\,\lambda_1\mathbf{a}_1|\lambda_2\mathbf{a}_2|\cdots|\,\lambda_n\mathbf{a}_n\,]$, we can apply Lemma 3.5 to $(A\Lambda)B^\top$ to prove this statement. ■

We will see later in this book how it is useful to represent a matrix, or a matrix resulting from matrix multiplication, as a sum of rank-one matrices.

3.8 Linear Transformation

In this section, we consider transformations of vectors between vector spaces that seamlessly work with vector addition and scalar multiplication. That is,

we consider transformations that are compatible with vector addition and scalar multiplication, under which adding two transformed vectors is equivalent to transforming the sum of these two vectors as well as multiplying a transformed vector with a scalar is also equivalent to transforming a vector scaled by the same scalar. We call such a transformation a linear transformation, and this can be visualized as a line that passes through the origin in the 2-dimensional space and a plane that passes through the original in the 3-dimensional space. We can more precisely define linear transformation as:

Definition 3.9 *We call a function $T : \mathbb{V} \to \mathbb{W}$ defined between two vector spaces a linear transformation if it satisfies the following properties given two arbitrary vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$:*[a]

- $T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2)$;
- $T(\alpha\mathbf{v}_1) = \alpha T(\mathbf{v}_1)$.

We call such a linear transformation a linear map as well.

[a] These two properties can be combined into one condition; "

$$T(\alpha\mathbf{v}_1 + \mathbf{v}_2) = \alpha T(\mathbf{v}_1) + T(\mathbf{v}_2)$$

for any scalar $\alpha \in \mathbb{R}$ and a pair of arbitrary vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$."

This definition of linear transformation does not involve a basis of either $\mathbb{V}$ nor $\mathbb{W}$, which brings up a question of whether there exists an efficient way to describe any linear transformation beyond specifying how each and every vector from $\mathbb{V}$ maps to a vector in $\mathbb{W}$. In order to answer this question and find such an efficient way, we restrict our attention to finite-dimensional vector spaces. That is, we assume that $\dim(\mathbb{V}) = n$ and $\dim(\mathbb{W}) = m$.

3.8.1 Matrix Representation of Linear Transformation

An arbitrary vector $\mathbf{v}$ in $\mathbb{V}$ can be expressed as a linear combination of the basic vectors with a unique coefficient set, $x_1, \ldots, x_n$, given a basis $\mathcal{B}_\mathbb{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.[6] That is, $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$. If T were a linear transformation,

$$T(\mathbf{v}) = T(x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n) = x_1T(\mathbf{v}_1) + \cdots + x_nT(\mathbf{v}_n),$$

[6] This is because basic vectors are linearly independent and they span the vector space.

which implies that we can describe T given the basis $\mathcal{B}_{\mathbb{V}}$ by describing $T(\mathbf{v}_j)$ of $\mathbb{W}$ that corresponds to each basic vector $\mathbf{v}_j$ of $\mathbb{V}$. In other words, we just need to determine $\{T(\mathbf{v}_j) : j = 1, \ldots, n\}$ and identify $\{x_j : j = 1, \ldots, n\}$, that satisfies $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$, in order to map an arbitrary $\mathbf{v} \in \mathbb{V}$ to $\mathbb{W}$ via T. With those information, we can evaluate $T(\mathbf{v})$ by $\sum_{j=1}^{n} x_j T(\mathbf{v}_j)$.

Let $\mathcal{B}_{\mathbb{W}} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis of $\mathbb{W}$. Because $T(\mathbf{v}_j) \in \mathbb{W}$, we can represent each $T(\mathbf{v}_j)$ as a linear combination of the basic vectors $\mathbf{w}_i$'s and their coefficients $\{a_{1j}, \ldots, a_{mj}\} \subset \mathbb{R}$:

$$T(\mathbf{v}_j) = a_{1j}\mathbf{w}_1 + \cdots + a_{mj}\mathbf{w}_m = \sum_{i=1}^{m} a_{ij}\mathbf{w}_i.$$

Combining two representations, we can then express $T(\mathbf{v})$, for an arbitrary $\mathbf{v} \in \mathbb{V}$, as

$$\begin{aligned} T(\mathbf{v}) &= x_1 T(\mathbf{v}_1) + \cdots + x_n T(\mathbf{v}_n) = \sum_{j=1}^{n} x_j T(\mathbf{v}_j) \\ &= \sum_{j=1}^{n} x_j \Big(\sum_{i=1}^{m} a_{ij}\mathbf{w}_i \Big) \\ &= \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} x_j \Big) \mathbf{w}_i. \end{aligned}$$

By denoting the coordinate of $T(\mathbf{v})$ under $\mathcal{B}_{\mathbb{W}}$ as $(y_1, \ldots, y_m)$, we have a purely algebraic equations in real numbers:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j \,, \;\; i = 1, \ldots, m,$$

because $T(\mathbf{v}) = \sum_{i=1}^{m} y_i \mathbf{w}_i$. Putting these all together, we observe that

$$\mathbf{y} = A\mathbf{x},$$

for an n-dimensional real vector $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ and an m-dimensional real vector $\mathbf{y} = (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$, given an $m \times n$ matrix $A = (a_{ij})$. These correspondences are illustrated in the following schematic diagram:

$$\begin{array}{lcl} \mathbb{V} \ni \mathbf{v} & \xrightarrow{\text{linear transform } T} & \mathbf{w} = T(\mathbf{v}) \in \mathbb{W} \\ \mathcal{B}_{\mathbb{V}}: \mathbf{v} = \sum_{j=1}^{n} x_j \mathbf{v}_j & & \mathcal{B}_{\mathbb{W}}: \mathbf{w} = \sum_{i=1}^{m} y_i \mathbf{w}_i \\ \mathbb{R}^n \ni \mathbf{x} & \xrightarrow{\text{multiplication by matrix } A} & \mathbf{y} = A\mathbf{x} \in \mathbb{R}^m \end{array}$$

There are a few interesting aspects we should keep in our mind about this relationship:

- $\mathbb{V}$ and $\mathbb{W}$ are not defined in the context of $\mathbb{R}^n$ and $\mathbb{R}^m$. They are general vector spaces that may consist of any vectors, such as n-th order polynomials.
- Other than that they are linearly independent, there is no more qualification on vectors forming a basis.
- The coordinates $x_1, \ldots, x_n$ tells us about the relationship between the individual vectors and a basis and does not say anything about T. Even with the same vector spaces $\mathbb{V}$, $\mathbb{W}$ and a fixed linear transformation T, the transformation matrix A changes if we choose to use a different basis for $\mathbb{V}$. Even when we fix the basis of $\mathbb{V}$, the matrix A also changes if we choose another basis for $\mathbb{W}$. In short, the matrix that represents T depends on the choice of the bases for the domain and range.
- We are often familiar with standard bases, $\mathcal{B}_{\mathbb{V}} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^n$ and $\mathcal{B}_{\mathbb{W}} = \{\mathbf{e}'_1, \ldots, \mathbf{e}'_m\} \subset \mathbb{R}^m$, for $\mathbb{V} = \mathbb{R}^n$ and $\mathbb{W} = \mathbb{R}^m$, respectively. In this case, T is represented by a transformation matrix $A = (a_{ij})$, where $a_{ij} = \big(T(\mathbf{e}_j)\big)_i = T(\mathbf{e}_j)^\top \mathbf{e}'_i$. This is only a special case and should not be considered as a general case nor a representative case.

Example 3.4 Let A be an arbitrary $m \times n$ matrix, and also $\mathbb{V} = \mathbb{R}^n$ and $\mathbb{W} = \mathbb{R}^m$. Define a transformation T from $\mathbb{V}$ to $\mathbb{W}$ to be $T(\mathbf{x}) = A\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n = \mathbb{V}$. Then, T is a linear transformation.

1. Let us introduce the standard bases $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^n$ and $\{\mathbf{e}'_1, \ldots, \mathbf{e}'_m\} \subset \mathbb{R}^m$ for $\mathbb{V}$ and $\mathbb{W}$, respectively. Under these standard bases, the matrix corresponding to T is A.
2. As we saw in Example 3.1, another basis for $\mathbb{R}^n$ is $\{\mathbf{e}_1, \mathbf{e}_1+\mathbf{e}_2, \ldots, \mathbf{e}_1+\cdots+\mathbf{e}_n\}$. When $\mathbf{z} = (z_1, \ldots, z_n)^\top$ is a coefficient vector representing $\mathbf{x} \in \mathbb{R}^n$ in the basis $\mathcal{B}$,

$$\begin{aligned} \mathbf{x} &= z_1\mathbf{e}_1 + z_2(\mathbf{e}_1 + \mathbf{e}_2) + \cdots + z_n(\mathbf{e}_1 + \cdots + \mathbf{e}_n) \\ &= (z_1 + \cdots + z_n)\mathbf{e}_1 + \cdots + (z_{n-1} + z_n)\mathbf{e}_{n-1} + z_n\mathbf{e}_n \\ &= U\mathbf{z}, \end{aligned}$$

where

$$U = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & \cdots & 1 & 1 \\ 0 & 0 & \ddots & \ddots & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Then, the transformation matrix for T is given as AU, with a basis $\{\mathbf{e}'_1, \ldots, \mathbf{e}'_m\}$ for $\mathbb{W}$.

∎

Example 3.5 A set $\mathbb{V}$ includes all polynomials of degree 2 or less, that is, $\mathbb{V} = \{a_0 + a_1 t + a_2 t^2 : a_0, a_1, a_2 \in \mathbb{R}\}$. Let T map a polynomial $f(t)$ to its derivative $f'(t)$.

1. It is easy to check that $\mathbb{V}$ is a vector space over the scalar set $\mathbb{R}$. Check it yourself.

2. Set $\mathcal{B}_{\mathbb{V}} = \{1, t, t^2\}$. It is clear that $\mathbb{V} = \operatorname{span} \mathcal{B}_{\mathbb{V}}$. To see the linear independence, assume $c_0 + c_1 t + c_2 t^2 = 0$ for all t. Because we get $c_0 = c_1 = c_2 = 0$ by sequentially trying 0, 1 and 2 for t, $\mathcal{B}_{\mathbb{V}}$ is linearly independent. Therefore, $\mathcal{B}_{\mathbb{V}}$ is a basis for $\mathbb{V}$ and $\dim \mathbb{V} = |\mathcal{B}_{\mathbb{V}}| = 3$.

3. T is a linear map from $\mathbb{V}$ into $\mathbb{V}$. That is, $T(a_0 + a_1 t + a_2 t^2) = a_1 + 2a_2 t \in \mathbb{V}$. T is linear since the differentiation is linear, that is, $T\big(af(t) + bg(t)\big) = \big(af(t) + bg(t)\big)' = af'(t) + bg'(t) = aT(f(t)) + bT(g(t))$.

4. Under the basis $\mathcal{B}_{\mathbb{V}}$,

$$a_0 + a_1 t + a_2 t^2 \in \mathbb{V} \iff (a_1, a_1, a_2) \in \mathbb{R}^3 .$$

Through T, the entries of the matrix representing T are decided as follows:

$$\begin{aligned} T(1) &= 0 = 0 \cdot 1 + 0 \cdot t + 0 \cdot t^2 \iff a_{11} = 0, a_{21} = 0, a_{31} = 0 \\ T(t) &= 1 = 1 \cdot 1 + 0 \cdot t + 0 \cdot t^2 \iff a_{12} = 1, a_{22} = 0, a_{32} = 0 \\ T(t^2) &= 2t = 0 \cdot 1 + 2 \cdot t + 0 \cdot t^2 \iff a_{13} = 0, a_{23} = 2, a_{33} = 0. \end{aligned}$$

3×3 matrix A representing T from $\mathcal{B}_{\mathbb{V}}$ into $\mathcal{B}_{\mathbb{V}}$ is given by $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$.

Example 3.6 A set $\mathbb{V}$ includes all polynomials of degree n or less, that is, $\mathbb{V} = \{a_0 + a_1 t + \ldots + a_n t^n : a_0, a_1, \ldots, a_n \in \mathbb{R}\}$. Let T map a polynomial $f(t)$ to its derivative $f'(t)$. n is a fixed integer.

1. It is easy to check that $\mathbb{V}$ is a vector space over the scalar set $\mathbb{R}$. Try it yourself.

2. Let $\mathcal{B}_\mathbb{V} = \{1, t, t^2, \ldots, t^n\}$. It is clear that $\mathbb{V} = \operatorname{span} \mathcal{B}_\mathbb{V}$. To see the linear independence, assume $a_0 + a_1 t + \cdots + a_n t^n = 0$ for all t. By plugging in various values into t, we get $a_0 = a_1 = \cdots = a_n = 0$, and thus $\mathcal{B}_\mathbb{V}$ is linearly independent. Therefore, $\mathcal{B}_\mathbb{V}$ is a basis for $\mathbb{V}$ and $\dim \mathbb{V} = |\mathcal{B}_\mathbb{V}| = n + 1$.

3. T can be thought of as a map from $\mathbb{V}$ into $\mathbb{V}$. If we set $\mathbb{W} = \{a_0 + a_1 t + \ldots + a_{n-1} t^{n-1} : a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}\}$, T is also a map from $\mathbb{V}$ onto $\mathbb{W}$. It can be also shown that T is a linear map in both cases. $\mathcal{B}_\mathbb{W} = \{1, t, t^2, \ldots, t^{n-1}\}$ is a basis for $\mathbb{W}$.

4. For each $k = 0, 1, \ldots, n$,

$$\begin{aligned} T(t^k) \quad &= \quad k t^{k-1} = 0 \cdot 1 + 0 \cdot t + \cdots + 0 \cdot t^{k-2} + k \cdot t^{k-1} \\ &\qquad + 0 \cdot t^k + \cdots + 0 \cdot t^n \\ &\iff a_{1(k+1)} = 0, \ldots, a_{(k-1)(k+1)} = 0, a_{k(k+1)} = k, \\ &\qquad a_{(k+1)(k+1)} = 0, \ldots, a_{(n+1)(k+1)} = 0 \,. \end{aligned}$$

   The $(n+1) \times (n+1)$ matrix A representing T from $\mathcal{B}_\mathbb{V}$ into $\mathcal{B}_\mathbb{V}$ is then $A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & n-1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$. The $n \times (n+1)$ matrix A' representing T from $\mathcal{B}_\mathbb{V}$ onto $\mathcal{B}_\mathbb{W}$ is $A' = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & n-1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & n \end{bmatrix}$. Observe that the matrix representations under different bases may be different.

Composition of Linear Transformations

Let $\mathbb{U}$, $\mathbb{V}$ and $\mathbb{W}$ be vector spaces with their dimensions, n, m and ℓ, respectively. We consider two linear transformations/maps, $S : \mathbb{U} \to \mathbb{V}$ and $T : \mathbb{V} \to \mathbb{W}$. Furthermore, let $A = (a_{kj})$ and $B = (b_{ik})$ be the transformation matrices representing S and T, respectively, with respect to three bases $\mathcal{B}_{\mathbb{U}} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, $\mathcal{B}_{\mathbb{V}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, and $\mathcal{B}_{\mathbb{W}} = \{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$. We now derive an $\ell \times n$ matrix C that represents the composition of S and T, $T \circ S : \mathbb{U} \to \mathbb{W}$, just like what we did before.

We can express $S(\mathbf{u}_j)$ for a basic vector $\mathbf{u}_j$ of $\mathbb{U}$ as a linear combination of basic vectors of $\mathbb{V}$ with appropriate a_{kj}'s,

$$S(\mathbf{u}_j) = \sum_{k=1}^{m} a_{kj}\mathbf{v}_k.$$

We can similarly write $T(\mathbf{v}_k)$ for a basic vector $\mathbf{v}_k$ of $\mathbb{V}$ as a linear combination of basic vectors of $\mathbb{W}$ as

$$T(\mathbf{v}_k) = \sum_{i=1}^{\ell} b_{ik}\mathbf{w}_i.$$

We then compose these two, as follows:

$$\begin{aligned}
(T \circ S)(\mathbf{u}_j) &= T\Big(\sum_{k=1}^{m} a_{kj}\mathbf{v}_k\Big) \\
&= \sum_{k=1}^{m} a_{kj} T(\mathbf{v}_k) \\
&= \sum_{k=1}^{m} a_{kj} \sum_{i=1}^{\ell} b_{ik}\mathbf{w}_i \\
&= \sum_{i=1}^{\ell} \Big(\sum_{k=1}^{m} b_{ik}a_{kj}\Big)\mathbf{w}_i \,.
\end{aligned}$$

This can be rewritten as

$$(T \circ S)(\mathbf{u}_j) = \sum_{i=1}^{\ell} c_{ij}\mathbf{w}_i$$

where

$$c_{ij} = \sum_{k=1}^{m} b_{ik}a_{kj} = (BA)_{ij} \,.$$

That is, $C = BA$.

The following theorem summarizes this result.

Theorem 3.3 *Let A and B be the matrix representations of linear transformations S and T, respectively, with respect to some bases. For the same bases, the matrix representation of $T \circ S$ is BA, and $T \circ S$ is a linear transformation corresponds to a matrix representation BA.*

3.8.2 Interpretable Linear Transformations

There are some cases where the transformation matrix A, in $T(\mathbf{x}) = A\mathbf{x}$, is intuitively interpretable.

- Scaling: A matrix in the form of $\alpha I = \begin{bmatrix} \alpha & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \alpha \end{bmatrix}$, multiplies each element of $\mathbf{x}$ by a scalar α, as $T(\mathbf{x}) = \alpha I\mathbf{x} = \alpha\mathbf{x}$.
- Rotation: Take as an example a 2×2 matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}$ which rotates $\mathbf{x}$ counter-clock-wise 90°. We can generalize such a matrix as
$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$
such that $R_\theta\mathbf{x}$ rotates $\mathbf{x}$ by θ. Using Theorem 3.3, we see that $R_\theta R_\phi$ corresponds to rotating a vector by ϕ and then by θ, which is equivalent to rotating the same vector by $\theta + \phi$. From this, we see that
$$R_\phi R_\theta = R_{\theta+\phi}.$$
From this relationship, we can further derive that $R_\theta^{-1} = R_{-\theta}$, because $R_\theta R_{-\theta} = I$. We can also easily check that $R_{-\theta} = R_\theta^\top$.
- Projection: Consider $P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ which projects a vector $\mathbf{x}$ on the x axis, as $T(\mathbf{x}) = P\mathbf{x} = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$. We can generalize such a matrix to perform projection of a vector on a line that passes the origin at the angle θ. We use P_θ to

refer to the transformation matrix corresponding to this particular linear map T. Two basic vectors in $\mathbb{R}^2$, $\mathbf{e}_1 = (1,0)^\top$ and $\mathbf{e}_2 = (0,1)^\top$, are mapped to $T(\mathbf{e}_1) = (\cos\theta\cos\theta, \sin\theta\cos\theta)^\top$ and $T(\mathbf{e}_2) = (\cos\theta\sin\theta, \sin\theta\sin\theta)^\top$, respectively.

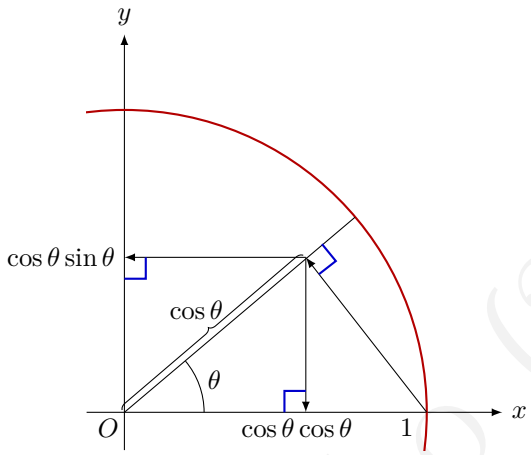


Figure 3.1: Projection of $\mathbf{e}_1$ onto the direction of angle θ

We can then get

$$P_\theta = \begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix},$$

because

$$\begin{aligned} T(\mathbf{x}) &= x_1 T(\mathbf{e}_1) + x_2 T(\mathbf{e}_2) \\ &= x_1 \begin{bmatrix} \cos\theta\cos\theta \\ \sin\theta\cos\theta \end{bmatrix} + x_2 \begin{bmatrix} \sin\theta\cos\theta \\ \sin\theta\sin\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

Since a projected vector should remain as is even if it is projected once more, it must hold that $P^2 = P$ for any projection matrix P. Indeed, P_θ above satisfies this condition.[7] Another interesting property of projection is that $I - P$ is also a projection matrix if P were.

[7] It is interesting to note that P_θ is a rank-one matrix and can be expressed as

$$P_\theta = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}^\top.$$

From this, we can also derive $P_\theta^2 = P_\theta$.

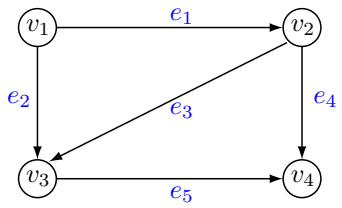
- Reflection: We can reflect a vector $\mathbf{x}$ on the other side of a line that passes the origin at the angle of 45° by multiplying it with $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, as $T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$.

  Let us generalize this to work with any given angle θ by defining H_θ. We first realize that the mid-point vector between $\mathbf{x}$ and $T(\mathbf{x})$ must be identical to $\mathbf{x}$ projected to the reflecting line. That is, $\frac{1}{2}(\mathbf{x}+T(\mathbf{x})) = P_\theta\mathbf{x}$. From this, we can derive H_θ by $H_\theta = 2P_\theta - I$, since $T(\mathbf{x}) = 2P_\theta\mathbf{x} - \mathbf{x} = (2P_\theta - I)\mathbf{x}$.

  Even more generally, consider a reflection matrix for any arbitrary subspace of $\mathbb{R}^n$. Given an $n \times n$ projection matrix P that projects a vector onto this subspace, (i.e. it is symmetric and satisfies $P^2 = P$) we can construct a reflection matrix H by $H = 2P - I$. This matrix satisfies $H^\top = H$ and $H^2 = (2P - I)^2 = 4P^2 - 4P + I = I$. Conversely, given an $n \times n$ matrix H that satisfies $H^\top = H$ and $H^2 = I$, we can obtain the projection matrix P by $P = \frac{1}{2}(H + I)$. This matrix P then satisfies $P^\top = P$ and $P^2 = \frac{1}{4}(H^2+2H+I) = \frac{1}{4}(2H+2I) = P$. Because $I-P$ is a projection matrix when P is also a projection matrix, $2(I-P)-I = I-2P$ is also a reflection matrix.

3.9 Application: Analysis of Graphs

We discussed the incidence matrix associated to a directed graph in Section 2.9. We investigate various aspects of this incidence matrix. Let us consider the following simple directed graph.



The incidence matrix of this graph is given by

$$A = \begin{array}{c|ccccc|} & e_1 & e_2 & e_3 & e_4 & e_5 \\ v_1 & -1 & -1 & & & \\ v_2 & 1 & & -1 & -1 & \\ v_3 & & 1 & 1 & & -1 \\ v_4 & & & & 1 & 1 \end{array}.$$

Denote $\mathbf{1} = (1, 1, 1, 1)^\top$ and the i-th column of A by $\mathbf{c}_i$. Then, we observe the followings.

1. $A^\top \mathbf{1} = \mathbf{0}$, that is, $\mathbf{1} \in \mathrm{Null}\,(A^\top)$. This property holds for all incidence matrices of directed graphs since their each column contains exactly one 1 and one -1 representing a directed arc.

2. A dependence relation $\mathbf{c}_1 - \mathbf{c}_2 + \mathbf{c}_3 = \mathbf{0}$ implies that $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ is linearly dependent and $(1, -1, 1, 0, 0)^\top \in \mathrm{Null}\,(A)$. If we discard the direction of arcs in the graph, arcs e_1, e_2, e_3 corresponding to columns $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$, respectively, constitute a cycle or loop. The correspondence between a cycle in a graph and a dependence relation among column vectors of its incidence matrix holds for all graphs.

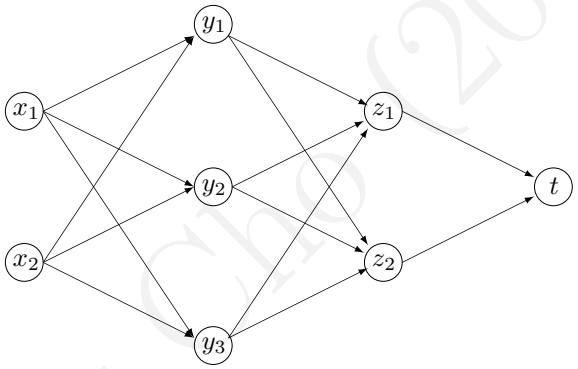
3. If any two nodes are connected by arcs discarding their directions, the graph is called a connected graph. If the graph has no cycle, the graph is called acyclic. If a graph has both properties above, that is, a graph is acyclic and connected, we call the graph a tree. It can be shown that the columns of the incidence matrix corresponding to arcs in a tree are linearly independent. Furthermore, it can be also shown that a tree with n nodes has $n - 1$ arcs by mathematical induction.

4. $\mathrm{rank}\, A = n - 1$ for an incidence matrix of a connected directed graph with n nodes. First, we get $\mathrm{rank}\, A \leq n - 1$ from $A^\top \mathbf{1} = \mathbf{0}$. The column vectors corresponding to $n - 1$ arcs constituting a tree in the graph are linearly independent, and $\mathrm{rank}\, A \geq n - 1$. A connected graph always has at least one tree.

5. $\mathrm{Null}\,(A^\top) = \mathrm{span}\{\mathbf{1}\}$ since $A^\top \mathbf{1} = \mathbf{0}$ and $\dim \mathrm{Null}\,(A^\top) = 1$ from $\mathrm{rank}\, A = n - 1$.

3.10 Application: Neural Networks

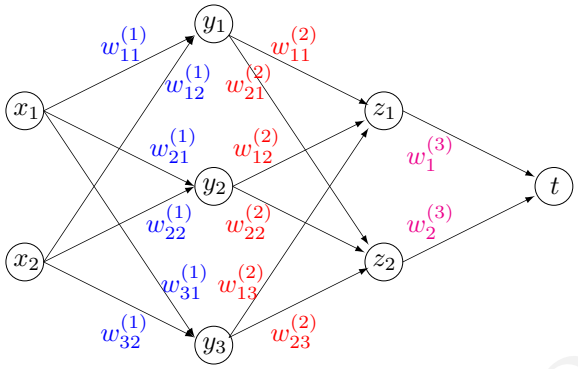
In machine learning, a neural network is often implemented as alternation between linear and nonlinear transformations, and linear transformations are often expressed as weight matrices. Training such a neural network corresponds to adjusting these linear transformations so as to minimize the difference between observations and neural network's predictions. When visualizing a layered neural network as a directed graph, or a network, we use a node to represent a nonlinear transformation and arcs correspond to linear transformations. For instance, we can visualize a neural network that takes as input a 2-dimensional input $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$, as below.



Computation happens left-to-right in this neural network. The input (x_1, x_2) is linearly and nonlinearly transformed into the first intermediate quantities (y_1, y_2, y_3). These are linearly and nonlinearly transformed into the next intermediate quantities (z_1, z_2). The final two are linearly transformed once more to form the final output t. It is visibly apparent from the figure above how this neural network is layered; (x_1, x_2) is the input layer, (y_1, y_2, y_3) and (z_1, z_2) are two hidden layers, and (t) is the output layer. Such a layered structure enables efficient computation even with a neural network with many nodes, such as by using a general-purpose graphics processing unit (GPU).

Consider linear transformations within this neural network. The first linear transformation from (x_1, x_2) to (y_1, y_2, y_3) can be expressed as a 3×2 matrix, according to Section 3.8.1, since it is linear transformation from a 2-dimensional space to a 3-dimensional space. Let this matrix be $W^{(1)} = (w_{ij}^{(1)})$. The next linear transformation can be expressed as a 2×2 matrix $W^{(2)} = (w_{ij}^{(2)})$, and the final one as a 1×2 matrix $W^{(3)} = [w_1^{(3)} w_2^{(3)}]$. Each arc in the directed graph is associated with one of the entries of these transformation matrices. We can

visualize this by putting the associated matrix element, to which we often refer as a parameter, on top of the corresponding arcs, as below.



We use $\sigma^{(k)}$ to denote the k-th nonlinear transformation. Such nonlinear transformation is often called an activation function, and it is often applied point-wise, that is, it is applied to each node in the layer independently. Let us use $\hat{}$ to denote the value of each node prior to nonlinear transformation, such as $\hat{y}_i, \hat{z}_i$ and $\hat{t}$. In this particular example, these pre-activation values are computed by

$$\hat{y}_1 = w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2, \;\; \hat{y}_2 = w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2, \;\; \hat{y}_3 = w_{31}^{(1)} x_1 + w_{32}^{(1)} x_2.$$

We apply the activation function to these pre-activation values to get the final values of the first layer: $y_1 = \sigma^{(1)}(\hat{y}_1)$, $y_2 = \sigma^{(1)}(\hat{y}_2)$, $y_3 = \sigma^{(1)}(\hat{y}_3)$. It is a common practice to apply the activation function to a vector, which is equivalent to applying the activation function to each element of the vector. This simplifies the equation above into

$$\hat{\mathbf{y}} = W^{(1)}\mathbf{x} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \\ w_{31}^{(1)} & w_{32}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \sigma^{(1)}(\hat{\mathbf{y}}).$$

The same procedures applies equally to the next layer: $\hat{z}_1 = w_{11}^{(2)} y_1 + w_{12}^{(2)} y_2 + w_{13}^{(2)} y_3$, $\hat{z}_2 = w_{21}^{(2)} y_1 + w_{22}^{(2)} y_2 + w_{23}^{(2)} y_3$ and $z_1 = \sigma^{(2)}(\hat{z}_1)$, $z_2 = \sigma^{(2)}(\hat{z}_2)$. This is simplified into

$$\hat{\mathbf{z}} = W^{(2)}\mathbf{y} = \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \mathbf{z} = \sigma^{(2)}(\hat{\mathbf{z}}).$$

In the final layer, we first compute $\hat{t} = w_1^{(3)} z_1 + w_2^{(3)} z_2$ and apply the activation function to get $t = \sigma^{(3)}(\hat{t})$. This is equivalent to

$$\hat{t} = W^{(3)}\mathbf{z} = \begin{bmatrix} w_{11}^{(3)} & w_{12}^{(3)} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad t = \sigma^{(3)}(\hat{t}).$$

In summary, this neural network performs the following computation:

$$t = \sigma^{(3)}\big(W^{(3)}\,\sigma^{(2)}\big(W^{(2)}\,\sigma^{(1)}\big(W^{(1)}\mathbf{x}\big)\big)\big).$$

Learning corresponds to adjusting $w_{ij}^{(k)}$ to minimize the difference between the output from the neural network and a desired (target) output. Modern neural networks sometimes have billions or even hundreds of billions of parameters.

Earlier, it was usual to use a so-called sigmoid function $(1 + e^{-x})^{-1}$, which is a bounded S-shaped curve, as an activation function. This choice however made learning greatly challenging. It has become more common in recent years to use a rectified linear function $\max\{x, 0\}$ instead, which is considered one of the major reasons behind the explosive growth of deep learning since 2012.

3.10.1 Flexibility of Neural Network Representations

An interesting and consequential question we can ask is how nonlinear a neural network that implement linear combination followed by such a simple activation function. Consider a neural network $f(x, y; \theta)$ in Figure 3.2.

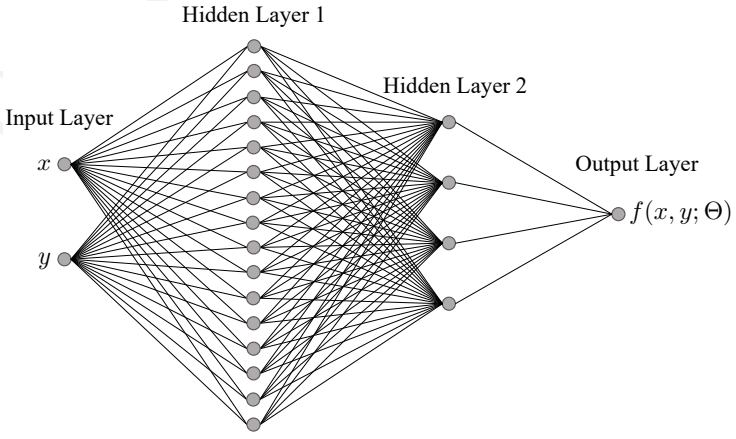


Figure 3.2: A simple neural network with 2 hidden layers and 121 learning parameters

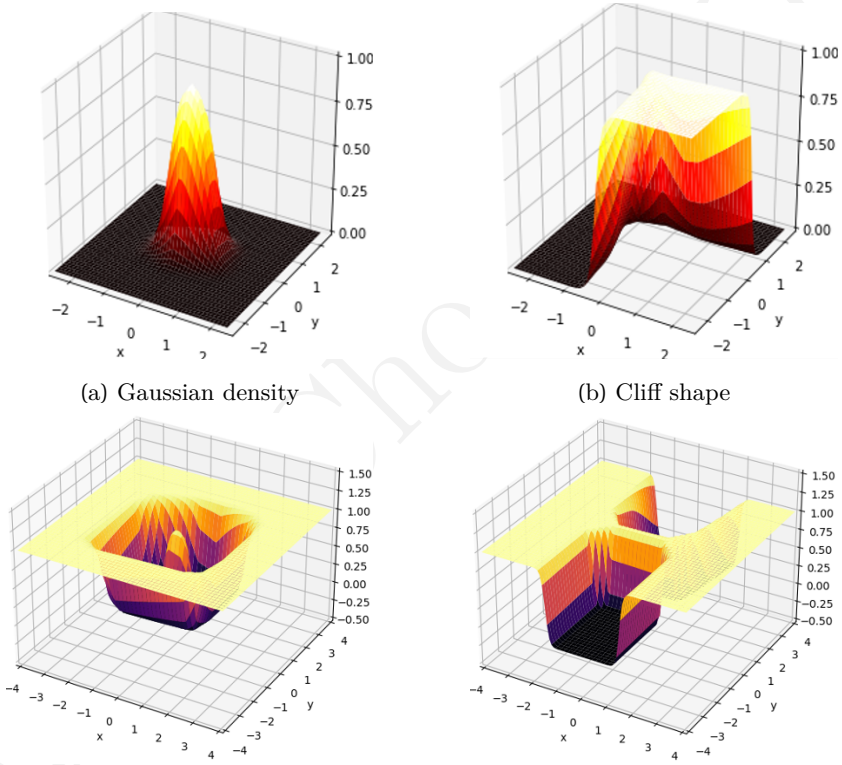
The output $f(x, y; \Theta)$ is determined by two hidden layers of sizes 16 and 4, respectively, given an input (x, y). These 20 nodes in the hidden layers use rectified linear functions as their activation functions, and we use a sigmoid function at the output layer. There are 100 arcs and 21 nodes, excluding the input nodes, resulting in 100 weights and 21 biases.[8] Let $\Theta \in \mathbb{R}^{121}$ a collection of these parameters. As we alter Θ, the output of the neural network given the same input (x, y) changes. In other words, Θ determines the function expressed by the neural network.

We obtained four functions, respectively corresponding to $\Theta_1, \ldots, \Theta_4$, by solving the following minimization problem using four real data:

$$\min_{\Theta \in \mathbb{R}^{121}} \sum_{i=1}^{n} \left\| f(x_i, y_i; \Theta) - f_i \right\|.$$

In Figure 3.3, we plot these four functions represented by four parameter sets. The diversity and complexity of these functions demonstrate that we can get vastly different, highly nonlinear functions by simply varying the parameters of the same neural network.

[8] A bias refers to an extra scalar added to the pre-activation value of each node. We omitted it earlier for brevity.



(a) Gaussian density

(b) Cliff shape

(c) Splash shape

(d) Bridge shape

Figure 3.3: Various output representations by a simple neural network

Chapter 4

Orthogonality and Projections

Based on our definitions of vectors and vector spaces, we are now ready to start asking questions about their relationships. For instance, we may ask whether two vectors are perpendicular to each other, and in answering so, we use a notion of an inner product between two vectors. When this inner product is zero, these two vectors are perpendicular to each other. There is a popular question in physics, that is related to the perpendicularity between vectors: What is effective force on an object when we apply force to an object moving in a direction not parallel to the applied force? We decompose force as a sum of two vectors; one vector along the moving direction and the other along its perpendicular direction, using the Pythagorean rule. These two perpendicular directions span the space containing the original force. We develop the same idea through so-called orthogonal projection in a more general vector space and derive many useful results about inner products and orthogonal projections in this chapter.

4.1 Inner Products

We start by introducing the inner product between two vectors and the norm of a vector. The inner product is defined to be linear with respect to each of these vectors.

Definition 4.1 *An inner product $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ between two vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, in a vector space $\mathbb{V}$ is a real-valued function that satisfies the following properties:*

1. $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle$;
2. $\langle c\mathbf{v}_1, \mathbf{v}_2 \rangle = c\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ *for any real number* c;
3. $\langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_3 \rangle = \langle \mathbf{v}_1, \mathbf{v}_3 \rangle + \langle \mathbf{v}_2, \mathbf{v}_3 \rangle$ *for any* $\mathbf{v}_3 \in \mathbb{V}$;
4. $\langle \mathbf{v}_1, \mathbf{v}_1 \rangle > 0$ *if and only if* $\mathbf{v}_1 \neq \mathbf{0}$.

When an inner product is defined, we use it to define a norm induced from the inner product as

$$|\mathbf{v}| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \quad \text{for each } \mathbf{v} \in \mathbb{V},$$

just like the absolute value of a real number. If the norm of a vector $\mathbf{v}$ is 1, that is, $\langle \mathbf{v}, \mathbf{v} \rangle = 1$, we call $\mathbf{v}$ a unit vector.

From the second property of an inner product in Definition 4.1, we see $\langle \mathbf{0}, \mathbf{v}_1 \rangle = 0$. We can show that $\langle \mathbf{v}_1, \mathbf{v}_2 + \mathbf{v}_3 \rangle = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_1, \mathbf{v}_3 \rangle$ from the first and third properties. An inner product is therefore linear in either of two vector arguments. In other words, an inner product is bilinear. Because $\alpha\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ is an inner product for any positive scalar α, we can easily see that we can derive infinitely many different inner products from just one. Similarly, the sum of two inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle_1 + \langle \mathbf{v}_1, \mathbf{v}_2 \rangle_2$, is also an inner product. We study how to characterize all possible inner products in Theorem 4.1.

From the definition, we see that there might be many norms on a single vector space since many inner products exist. Generally, a norm of a vector in a vector space $\mathbb{V}$ is a real-valued function f that satisfies the following three properties:

- $f(\mathbf{v}) \geq 0$ for all $\mathbf{v} \in \mathbb{V}$, and $f(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$;
- Scalar multiplication: $f(c\mathbf{v}) = |c| f(\mathbf{v})$ for all $c \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{V}$;
- Triangle inequality: $f(\mathbf{v} + \mathbf{w}) \leq f(\mathbf{v}) + f(\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{V}$.

Together with Fact 4.2 below, we can show that the norm defined using an inner product satisfies the above three properties, and hence it is truly a norm.

Let us study an inner product further first by showing that the Cauchy-Schwarz inequality holds in a vector space as well.

Fact 4.1 *The Cauchy-Schwarz inequality holds between an inner product and the norm induced by the inner product:*

$$\left|\langle \mathbf{v}_1, \mathbf{v}_2 \rangle\right| \leq |\mathbf{v}_1| \cdot |\mathbf{v}_2| \,.$$

Proof: For any real number $t \in \mathbb{R}$,

$$\begin{aligned} |t\mathbf{v}_1 + \mathbf{v}_2|^2 &= \langle t\mathbf{v}_1 + \mathbf{v}_2, t\mathbf{v}_1 + \mathbf{v}_2 \rangle \\ &= \langle t\mathbf{v}_1 + \mathbf{v}_2, t\mathbf{v}_1 \rangle + \langle t\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= \langle t\mathbf{v}_1, t\mathbf{v}_1 \rangle + \langle \mathbf{v}_2, t\mathbf{v}_1 \rangle + \langle t\mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= t^2 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + t\langle \mathbf{v}_2, \mathbf{v}_1 \rangle + t\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= t^2 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + 2t\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \,. \end{aligned}$$

Since $|t\mathbf{v}_1 + \mathbf{v}_2| \geq 0$ for any t, the quadratic equation on the right-hand side cannot have two different solutions. That is,

$$0 \geq \langle \mathbf{v}_1, \mathbf{v}_2 \rangle^2 - \langle \mathbf{v}_1, \mathbf{v}_1 \rangle \langle \mathbf{v}_2, \mathbf{v}_2 \rangle = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle^2 - |\mathbf{v}_1|^2 \cdot |\mathbf{v}_2|^2 \,.$$

■

From the Cauchy-Schwarz inequality, we can derive the triangle inequality with which we find the definition of the norm from an inner product much more natural.

Fact 4.2 *The triangle inequality holds for the norm induced from an inner product $\langle \cdot, \cdot \rangle$:*

$$|\mathbf{v}_1 + \mathbf{v}_2| \leq |\mathbf{v}_1| + |\mathbf{v}_2|$$

and furthermore the positive homogeneity holds:

$$|c\mathbf{v}| = c|\mathbf{v}| \quad \textit{for any positive real number } c \,.$$

Proof:

$$\begin{aligned} |\mathbf{v}_1 + \mathbf{v}_2|^2 &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 \rangle \\ &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 \rangle + \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + \langle \mathbf{v}_2, \mathbf{v}_1 \rangle + \langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + 2\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \end{aligned}$$

$$\begin{aligned} &\leq |\mathbf{v}_1|^2 + 2|\mathbf{v}_1| \cdot |\mathbf{v}_2| + |\mathbf{v}_2|^2 \\ &= \big(|\mathbf{v}_1| + |\mathbf{v}_2|\big)^2 . \end{aligned}$$

$$|c\mathbf{v}_1| = \sqrt{\langle c\mathbf{v}, c\mathbf{v}\rangle} = \sqrt{c^2\langle \mathbf{v}, \mathbf{v}\rangle} = c\sqrt{\langle \mathbf{v}, \mathbf{v}\rangle} .$$

■

As the first example of an inner product, consider the following vector space consisting of polynomial functions.

Example 4.1 Let $\mathbb{V}$ be a vector space of polynomials of degrees less than or equal to n. For two polynomials $f(t)$, $g(t) \in \mathbb{V}$, we define an inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(t)g(t)dt .$$

For $f(t) = t$ and $g(t) = t^2$,

- $\int_{-1}^{1} f(t)g(t)dt = \int_{-1}^{1} g(t)f(t)dt$ is linear in f and g. $\int_{-1}^{1} f(t)^2 dt = 0$ implies $f \equiv 0$ since f is a polynomial which is continuous. Hence, $\langle \cdot, \cdot \rangle$ is an inner product.
- $\langle f, g\rangle = \int_{-1}^{1} t \cdot t^2 dt = \int_{-1}^{1} t^3 dt = 0$.
- $|f| = \sqrt{\langle f, f\rangle} = \sqrt{\int_{-1}^{1} t^2 dt} = \sqrt{\frac{1}{3}t^3\Big|_{-1}^{1}} = \sqrt{\frac{2}{3}}$.
- $|g| = \sqrt{\langle g, g\rangle} = \sqrt{\int_{-1}^{1} t^4 dt} = \sqrt{\frac{1}{5}t^5\Big|_{-1}^{1}} = \sqrt{\frac{2}{5}}$.
- $|f - g| = \sqrt{|f|^2 - 2\langle f, g\rangle + |g|^2} = \sqrt{\frac{2}{3} - 2 \cdot 0 + \frac{2}{5}} = \sqrt{\frac{16}{15}}$.

▮

Let us analyze an inner product defined over a vector space. Rather than considering all possible vector pairs within the vector space, we focus on pairs of linearly independent vectors.

Lemma 4.1 *Assume non-zero, linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in a vector space $\mathbb{V}$. Define a $k \times k$ matrix $A = (a_{ij})$, where $a_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j\rangle$. Then, A is symmetric and invertible.*

Proof: Symmetry is naturally derived from the definition of the inner product. To check the invertibility of A, we need to show that $A\mathbf{x} = \mathbf{0}$ has no non-trivial solution. Let $\mathbf{x} \in \mathbb{R}^k$ satisfy $A\mathbf{x} = \mathbf{0}$. The i-th equation of $A\mathbf{x} = \mathbf{0}$ is

$$a_{i1}x_1 + \cdots + a_{ik}x_k = x_1\langle \mathbf{v}_i, \mathbf{v}_1\rangle + \cdots + x_k\langle \mathbf{v}_i, \mathbf{v}_k\rangle = 0.$$

Rearranging the equation using the linearity of inner product gives

$$\langle \mathbf{v}_i, x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k\rangle = 0 \quad \text{for } i = 1, \ldots, k\,.$$

If we add the equations after multiplying x_i to i-th equation, the linearity again allows

$$0 = \sum_{i=1}^{k} x_i\langle \mathbf{v}_i, x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k\rangle = \langle x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k,\, x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k\rangle\,.$$

Then the definition of inner product imposes $x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k = \mathbf{0}$, which implies $\mathbf{x} = \mathbf{0}$ by the linear independence of $\mathbf{v}_i$'s. Hence, the null space of A is $\{\mathbf{0}\}$, and A is invertible. ■

Characterization of an Inner Product

Using Lemma 4.1, let us see how we can characterize an inner product in terms of inner products of basic vector pairs.

Consider a basis $\mathcal{B}_{\mathbb{V}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{V}$. For any two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{V}$, there exist two unique vectors $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ in $\mathbb{R}^n$, respectively, such that

$$\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n \quad \text{and} \quad \mathbf{w} = y_1\mathbf{v}_1 + \cdots + y_n\mathbf{v}_n\,.$$

Then, the bilinearity of inner product implies

$$\langle \mathbf{v}, \mathbf{w}\rangle = \Big\langle \sum_{i=1}^{n} x_i\mathbf{v}_i, \sum_{j=1}^{n} y_j\mathbf{v}_j \Big\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_iy_j\langle \mathbf{v}_i, \mathbf{v}_j\rangle\,.$$

If we set $a_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j\rangle$ and construct an $n \times n$ symmetric matrix $A = (a_{ij})$,

$$\langle \mathbf{v}, \mathbf{w}\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_iy_ja_{ij} = \mathbf{x}^\top A\mathbf{y}\,.$$

Since $\mathbf{v} = \mathbf{0} \in \mathbb{V}$ if and only if $\mathbf{x} = \mathbf{0} \in \mathbb{R}^n$, $\langle \mathbf{v}, \mathbf{v}\rangle = \mathbf{x}^\top A\mathbf{x} > 0$ if and only if $\mathbf{x} \neq \mathbf{0}$. Hence, the A can be characterized as

an $n \times n$ real symmetric matrix such that $\mathbf{x}^\top A\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$.

Therefore, a quadratic form taking positive values for any non-zero vectors characterizes an inner product. The matrix appearing in the quadratic form also characterizes the inner product. This property is called positive definiteness and formalized by the following Definition 4.2.[1]

Definition 4.2 *A square matrix A is positive definite if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.*

Conversely, with a positive definite matrix A, a bilinear function, defined as

$$\Big\langle \sum_{i=1}^{n} x_i \mathbf{v}_i, \sum_{j=1}^{n} y_j \mathbf{v}_j \Big\rangle = \mathbf{x}^\top A \mathbf{y}\,, \tag{4.1}$$

induces an inner product by Lemma 4.2.

We can summarize this observation in the following theorem:

Theorem 4.1 *An inner product in an n-dimensional vector space is characterized by an $n \times n$ symmetric positive definite matrix as in (4.1).*

To complete the proof of Theorem 4.1, we must introduce the following lemma.

Lemma 4.2 *Let $\mathbb{V}$ be a vector space and A be an $n \times n$ symmetric positive definite matrix. Fix an arbitrary basis of $\mathbb{V}$, such as $\mathcal{B}_\mathbb{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. For any two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{V}$, let $\langle \mathbf{v}, \mathbf{w}\rangle = \mathbf{x}^\top A\mathbf{y}$, where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ satisfy $\mathbf{v} = \sum_{i=1}^n x_i\mathbf{v}_i$ and $\mathbf{w} = \sum_{i=1}^n y_i\mathbf{v}_i$. Then, $\langle \mathbf{v}, \mathbf{w}\rangle$ is an inner product of $\mathbb{V}$.*

Proof: Let $\mathbf{v} = \sum_{i=1}^n x_i\mathbf{v}_i$, $\mathbf{w} = \sum_{i=1}^n y_i\mathbf{v}_i$, and $\mathbf{u} = \sum_{i=1}^n z_i\mathbf{v}_i$.

- $\langle \mathbf{v}, \mathbf{w}\rangle = \mathbf{x}^\top A\mathbf{y} = (\mathbf{x}^\top A\mathbf{y})^\top = \mathbf{y}^\top A^\top \mathbf{x} = \mathbf{y}^\top A\mathbf{x} = \langle \mathbf{w}, \mathbf{v}\rangle$;
- $\langle c\mathbf{v}, \mathbf{w}\rangle = (c\mathbf{x})^\top A\mathbf{y} = c(\mathbf{x}^\top A\mathbf{y}) = c\langle \mathbf{w}, \mathbf{v}\rangle$;
- $\langle \mathbf{v} + \mathbf{u}, \mathbf{w}\rangle = (\mathbf{x} + \mathbf{z})^\top A\mathbf{y} = \mathbf{x}^\top A\mathbf{y} + \mathbf{z}^\top A\mathbf{y} = \langle \mathbf{v}, \mathbf{w}\rangle + \langle \mathbf{u}, \mathbf{w}\rangle$
- $\langle \mathbf{v}, \mathbf{v}\rangle = \mathbf{x}^\top A\mathbf{x} > 0$ if $\mathbf{x} \neq \mathbf{0}$ and equivalently $\mathbf{v} \neq \mathbf{0}$.

■

[1] We learn positive definite matrices in more detail in Chapter 7.

Example 4.2 Let $\mathbb{V}$ be a collection of polynomials of degree less than 3. That is, $\mathbb{V} = \{a_0 + a_1 x + a_2 x^2 : a_i \in \mathbb{R}\}$. We discussed that $\mathbb{V}$ is a vector space. Define $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)dx$ for $f, g \in \mathbb{V}$. $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{V}$ since integration is linear in integrand. Let us characterize this inner product using a positive definite matrix with respect to a basis $\{1, x, x^2\}$.

$$\begin{aligned}
\langle 1, 1\rangle &= \int_{-1}^{1} 1 \cdot 1 dx = 2\,, \quad \langle 1, x\rangle = \int_{-1}^{1} 1 \cdot x dx = 0 \\
\langle 1, x^2\rangle &= \int_{-1}^{1} 1 \cdot x^2 dx = \frac{2}{3}\,, \quad \langle x, x\rangle = \int_{-1}^{1} x \cdot x dx = \frac{2}{3} \\
\langle x, x^2\rangle &= \int_{-1}^{1} x \cdot x^2 dx = 0\,, \quad \langle x^2, x^2\rangle = \int_{-1}^{1} x^2 \cdot x^2 dx = \frac{2}{5}
\end{aligned}$$

The matrix corresponding to this inner product is given by $\begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix}$. ∎

Example 4.3 Let $\mathbb{V} = \mathbb{R}^3$ with a bilinear function $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^\top A \mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ where $A = \begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix}$. From

$$\begin{aligned}
\mathbf{x}^\top A \mathbf{x} &= [x_1\ x_2\ x_3] \begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
&= 2x_1^2 + \frac{4}{3}x_1 x_3 + \frac{2}{3}x_2^2 + \frac{2}{5}x_3^2 \\
&= 2\big(x_1 + \frac{1}{3}x_3\big)^2 + \frac{2}{3}x_2^2 + \frac{8}{45}x_3^2\,,
\end{aligned}$$

we see that $\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x}\rangle = 0$ implies $\mathbf{x} = \mathbf{0}$. $\langle \mathbf{x}, \mathbf{y}\rangle$ is also bilinear. Therefore, it is an inner product. This inner product is equivalent to the one from Example 4.2 above, as

$$\begin{aligned}
\langle [1\ 0\ 0], [1\ 0\ 0]\rangle &= 2\,, \quad \langle [1\ 0\ 0], [0\ 1\ 0]\rangle = 0 \\
\langle [1\ 0\ 0], [0\ 0\ 1]\rangle &= \frac{2}{3}\,, \quad \langle [0\ 1\ 0], [0\ 1\ 0]\rangle = \frac{2}{3} \\
\langle [0\ 1\ 0], [0\ 0\ 1]\rangle &= 0\,, \quad \langle [0\ 0\ 1], [0\ 0\ 1]\rangle = \frac{2}{5}\,.
\end{aligned}$$

∎

Of course, in a Euclidean space which is frequently used by and familiar to us, we can define a variety of inner products. Among these inner products, we call the one defined below the standard inner product.

Definition 4.3 *We define the standard inner product, or sometimes dot product, between two vectors,* $\mathbf{x} = (x_1, \ldots, x_n)^\top$ *and* $\mathbf{y} = (y_1, \ldots, y_n)^\top$*, in the n-dimensional Euclidean space,* $\mathbb{R}^n$*, as*

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \cdots + x_n y_n = \mathbf{x}^\top \mathbf{y}. \tag{4.2}$$

The norm is thus defined as $|\mathbf{x}| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{x_1^2 + \cdots + x_n^2}$*, and we refer to it as the Euclidean norm.*

The standard inner product in $\mathbb{R}^n$ corresponds to (4.1) with $A = I_n$. $\langle \mathbf{x}, \mathbf{y} \rangle = 2\sum_{k=1}^n x_k y_k$, $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n k x_k y_k$, and $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^n \frac{1}{k} x_k y_k$ are examples of inner products in $\mathbb{R}^n$.

Example 4.4 Let us see that the standard inner product is indeed an inner product in $\mathbb{R}^n$ by showing that the dot product satisfies the properties of an inner product, following elementary arithmetic steps. Given $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $c \in \mathbb{R}$,

- $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = x_1 y_1 + \cdots + x_n y_n = y_1 x_1 + \cdots + y_n x_n = \mathbf{y}^\top \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle$;
- $\langle c\mathbf{x}, \mathbf{y} \rangle = (c\,\mathbf{x})^\top \mathbf{y} = (cx_1) y_1 + \cdots + (cx_n) y_n = c(x_1 y_1 + \cdots + x_n y_n) = c\,(\mathbf{x}^\top \mathbf{y}) = c\langle \mathbf{x}, \mathbf{y} \rangle$;
- $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = (\mathbf{x} + \mathbf{y})^\top \mathbf{z} = (x_1 + y_1) z_1 + \cdots + (x_n + y_n) z_n = (x_1 z_1 + \cdots + x_n z_n) + (y_1 z_1 + \cdots + y_n z_n) = \mathbf{x}^\top \mathbf{z} + \mathbf{y}^\top \mathbf{z} = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$;
- $|\mathbf{x}|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\top \mathbf{x} = x_1^2 + \cdots + x_n^2 \geq 0$. If $|\mathbf{x}| = 0$, all $x_i = 0$ and $\mathbf{x} = \mathbf{0}$. It is straightforward to show the converse.

∎

4.2 Orthogonal Vectors and Subspaces

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ be two non-zero vectors in $\mathbb{R}^n$. We ask whether two directions defined respectively by $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to

each other. By assuming $\mathbf{x}$ and $\mathbf{y}$ are not parallel to each other,[2] there exists a unique plane that contains both the origin $\mathbf{0}$, $\mathbf{x}$ and $\mathbf{y}$. This plane, which is a 2-dimensional subspace, is spanned by $\mathbf{x}$ and $\mathbf{y}$, i.e., $P = \{a\mathbf{x} + b\mathbf{y} : a, b \in \mathbb{R}\}$. Within this subspace, $\mathbf{0}$, $\mathbf{x}$ and *by* form a triangle, and the lengths of three edges are defined as the Euclidean norms, $|\mathbf{x}|$, $|\mathbf{y}|$ and $|\mathbf{x}-\mathbf{y}|$. When $|\mathbf{x}|^2+|\mathbf{y}|^2 = |\mathbf{x}-\mathbf{y}|^2$, we say the Pythagorean relationship holds and that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other. Because the Euclidean norm is defined with the standard inner product,

$$\begin{aligned} 0 &= |\mathbf{x}|^2 + |\mathbf{y}|^2 - |\mathbf{x}-\mathbf{y}|^2 \\ &= x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 - (x_1-y_1)^2 - \cdots - (x_n-y_n)^2 \\ &= 2(x_1y_1 + \cdots + x_ny_n) \\ &= 2\,\mathbf{x}^\top\mathbf{y}, \end{aligned}$$

which implies that $\mathbf{x}^\top\mathbf{y} = 0$ is a necessary and sufficient condition for $\mathbf{x}$ and $\mathbf{y}$ being orthogonal to each other.

Instead of orthogonality, we can think of the angle between a pair of vectors, $\mathbf{x}$ and $\mathbf{y}$. From a well-known result from 2-dimensional geometry,

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

when the angle between $\mathbf{x}$ and $\mathbf{y}$ is θ. From this, we can derive the angle between any arbitrary pair of vectors as

$$\cos\theta = \frac{\mathbf{x}^\top\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}. \tag{4.3}$$

We often refer to it as the cosine similarity between two vectors and use it to measure the similarity between two vectors in terms of their angles while ignoring their norms.

In this section, we generalize the orthogonality of two vectors to work with a general inner product, beyond the standard inner product. This allows us to derive a variety of interesting results later.

[2] $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$ if they are parallel.

Definition 4.4 *In a finite-dimensional vector space $\mathbb{V}$ with an inner product $\langle \cdot, \cdot \rangle$,*

1. *We say $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal and use $\mathbf{v}_1 \perp \mathbf{v}_2$ to express the orthogonality, if*
$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0.$$
2. *If every pair of vectors among $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is orthogonal, i.e., $\mathbf{v}_i \perp \mathbf{v}_j$ for $i \neq j$, we say that they are (mutually) orthogonal.*
3. *When unit vectors are (mutually) orthogonal, we say they are orthonormal.*
4. *If a basis consists of orthonormal basic vectors, we call it an orthonormal basis.*

Let us now connect this generalized notion of orthogonality with the traditional notion of orthogonality we just described earlier. We start by showing that this new definition of orthogonality extends the Pythagorean relationship among the edges of an right triangle to an arbitrary vector space.

Fact 4.3 *Show that two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal if and only if the Pythagorean relation holds:*
$$|\mathbf{v}_1|^2 + |\mathbf{v}_2|^2 = |\mathbf{v}_1 + \mathbf{v}_2|^2 .$$

Proof: Since $|\mathbf{v}_1 + \mathbf{v}_2|^2 = \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 \rangle = \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + 2\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle$ for any two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, we have
$$|\mathbf{v}_1 + \mathbf{v}_2|^2 = |\mathbf{v}_1|^2 + 2\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + |\mathbf{v}_2|^2 ,$$
which proves the statement. ■

We can also show that orthogonal vectors are linearly independent as well.

Fact 4.4 *If non-zero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are mutually orthogonal, then they are linearly independent.*

Proof: Consider $c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$ for some real c_i's. Since $\mathbf{v}_i$'s are mutually orthogonal,
$$\langle \mathbf{v}_i, \mathbf{0} \rangle \quad = \quad \langle \mathbf{v}_i, c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n \rangle$$

$$= c_1\langle \mathbf{v}_i, \mathbf{v}_1\rangle + \cdots + c_i\langle \mathbf{v}_i, \mathbf{v}_i\rangle + \cdots + c_n\langle \mathbf{v}_i, \mathbf{v}_n\rangle$$
$$= c_i\langle \mathbf{v}_i, \mathbf{v}_i\rangle .$$

Since $\langle \mathbf{v}_i, \mathbf{v}_i\rangle > 0$ and $\langle \mathbf{v}_i, \mathbf{0}\rangle = 0$, $c_i = 0$. ■

We can draw the following observation about orthonormal vectors.

Fact 4.5 *Let* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *be an orthonormal basis for a vector space* $\mathbb{V}$. *For any vector* $\mathbf{v} \in \mathbb{V}$,

$$\mathbf{v} = \langle \mathbf{v}, \mathbf{v}_1\rangle \mathbf{v}_1 + \cdots + \langle \mathbf{v}, \mathbf{v}_n\rangle \mathbf{v}_n = \sum_{i=1}^{n} \langle \mathbf{v}, \mathbf{v}_i\rangle \mathbf{v}_i$$

holds and the representation is unique.

Proof: Since the basis span the space, we have $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n = \sum_{i=1} x_i\mathbf{v}_i$ for some real x_i's. From the orthogonality and normality,

$$\langle \mathbf{v}, \mathbf{v}_j\rangle = \langle \sum_{i=1}^{n} x_i\mathbf{v}_i, \mathbf{v}_j\rangle = \sum_{i=1}^{n} x_i\langle \mathbf{v}_i, \mathbf{v}_j\rangle = x_j\langle \mathbf{v}_j, \mathbf{v}_j\rangle = x_j , \quad \text{for each } j = 1, \ldots, n$$

and we obtain the desired representation. ■

Fact 4.5 implies that we can perfectly represent an arbitrary vector as inner-products against basic vectors in an orthonormal basis. That is, each inner product $\langle \mathbf{v}, \mathbf{v}_i\rangle$ serves as a coordinate along the corresponding orthonormal basic vector.

Example 4.5 For the Euclidean vector space $\mathbb{R}^n$, define the standard basic vectors

$$\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^n , \quad i = 1, \ldots, n$$

whose i-th element is 1, and all others are 0. The standard basis in $\mathbb{R}^n$ is $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

1. The standard basis in $\mathbb{R}^n$ is orthonormal since $|\mathbf{e}_i|^2 = \mathbf{e}_i^\top \mathbf{e}_i = 1$, $\mathbf{e}_i^\top \mathbf{e}_j = 0$ for $i \neq j$;

2. For $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, $x_i = \langle \mathbf{x}, \mathbf{e}_i\rangle = \mathbf{x}^\top \mathbf{e}_i$ and $\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n$, which confirms Fact 4.5 for the standard basis and the standard inner product in $\mathbb{R}^n$ since $\mathbf{x}^\top \mathbf{e}_i = x_i$ and $\mathbf{x} = (x_1, 0, \ldots, 0)^\top + \cdots + (0, \ldots, 0, x_n)^\top = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$.

In mathematics, it is usual to say that a property is satisfied by two sets, if it holds between every pair of elements from two sets. Along this line, we can generalize the notion of orthogonality to subspaces.

Definition 4.5 *We say two subspaces, $\mathbb{U}$ and $\mathbb{W}$, of a vector space $\mathbb{V}$ are orthogonal and use $\mathbb{U} \perp \mathbb{W}$ to denote the orthogonality, if*

$$\mathbf{u} \perp \mathbf{w} \quad \textit{for all} \quad \mathbf{u} \in \mathbb{U}, \mathbf{w} \in \mathbb{W}.$$

From this definition of orthogonality of subspaces, we can define the orthogonal complement.

Definition 4.6 *An orthogonal complement of a subspace $\mathbb{W}$ of a vector space $\mathbb{V}$ is defined as*

$$\mathbb{W}^{\perp} = \{\mathbf{v} \in \mathbb{V} : \mathbf{v} \perp \mathbf{w} \quad \textit{for all } \mathbf{w} \in \mathbb{W}\}.$$

Due to the linearity of an inner product, $\mathbb{W}^{\perp}$ is also a subspace of $\mathbb{V}$.

Fact 4.6 *$\mathbb{W}^{\perp}$ is also a subspace of a vector space $\mathbb{V}$ if $\mathbb{W}$ is a subspace of $\mathbb{V}$.*

Proof: By definition, $\mathbb{W}^{\perp} \subset \mathbb{V}$ and $\mathbf{0} \in \mathbb{W}^{\perp}$. With $\alpha, \beta \in \mathbb{R}$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{W}^{\perp}$, $\langle \alpha\mathbf{v}_1 + \beta\mathbf{v}_2, \mathbf{w} \rangle = \alpha\langle \mathbf{v}_1, \mathbf{w} \rangle + \beta\langle \mathbf{v}_2, \mathbf{w} \rangle = 0$ for any $\mathbf{w} \in \mathbb{W}$. That is, $\alpha\mathbf{v}_1 + \beta\mathbf{v}_2 \in \mathbb{W}^{\perp}$. Therefore, $\mathbb{W}^{\perp}$ is a subspace of $\mathbb{V}$. ■

We should warn you that two orthogonal subspaces may not be the orthogonal complement of each other. For instance, two subspaces of $\mathbb{R}^3$ under the standard inner product, $\mathbb{U} = \text{span}\{(1, 0, 0)\}$ and $\mathbb{W} = \text{span}\{(0, 1, 0)\}$, are orthogonal, but the orthogonal complement of the former, $\mathbb{U}^{\perp} = \text{span}\{(0, 1, 0), (0, 0, 1)\}$, is a proper superset of $\mathbb{W}$. In other words, $\mathbb{U} \perp \mathbb{W}$ but $\mathbb{U}^{\perp} \supsetneq \mathbb{W}$. We present and study specific examples of orthogonal complements in Euclidean spaces, such as $\text{Col}\,(A^{\top})^{\perp} = \text{Null}\,(A)$ and $\text{Null}\,(A)^{\perp} = \text{Col}\,(A^{\top})$, later in Section 4.6.

From the example above, where two 1-dimensional subspaces were produced from two orthogonal vectors, respectively, we can see that two orthogonal subspaces do not overlap with each other except for the origin in general.

Fact 4.7 *Let $\mathbb{W}$ be a subspace of a vector space $\mathbb{V}$, and $\mathbb{W}^\perp$ be the orthogonal complement of $\mathbb{W}$ in $\mathbb{V}$. Then, $\mathbb{W} \cap \mathbb{W}^\perp = \{\mathbf{0}\}$.*

Proof: Let $\mathbf{w} \in \mathbb{W} \cap \mathbb{W}^\perp$. Since $\mathbf{w} \in \mathbb{W}$ and $\mathbf{w} \in \mathbb{W}^\perp$, we conclude $\langle \mathbf{w}, \mathbf{w} \rangle = 0$ by the definition of the orthogonal complement. The definition of the inner product implies that $\mathbf{w} = \mathbf{0}$. ■

According to Definition 3.3, Fact 4.7 guarantees $\mathbb{W} + \mathbb{W}^\perp = \mathbb{W} \oplus \mathbb{W}^\perp$. Therefore, the summands are unique once a vector is represented as a sum of two vectors from a subspace and its orthogonal complements by Fact 3.1. We summarize this observation as the following theorem.

Theorem 4.2 *Let $\mathbb{W}$ be a subspace of a vector space $\mathbb{V}$, and $\mathbb{W}^\perp$ be the orthogonal complement of $\mathbb{W}$ in $\mathbb{V}$. If every $\mathbf{v} \in \mathbb{V}$ has a decomposition of $\mathbf{v} = \mathbf{w} + \mathbf{z}$ where $\mathbf{w} \in \mathbb{W}$ and $\mathbf{z} \in \mathbb{W}^\perp$, then this decomposition is unique and $\mathbb{V} = \mathbb{W} \oplus \mathbb{W}^\perp$.*

As an illustrating example of Theorem 4.2, let us consider $\mathbb{R}^3$.

Example 4.6 Let $\mathbb{V} = \mathbb{R}^3$ and $\mathbb{W} = \mathbb{R} \times \{0\} \times \{0\}$. Then, $\mathbb{W}^\perp = \{0\} \times \mathbb{R} \times \mathbb{R}$. Check yourself that $\mathbb{W} \cap \mathbb{W}^\perp = \{\mathbf{0}\}$. For $(x, y, z)^\top \in \mathbb{V}$, $(x, y, z)^\top = (x, 0, 0)^\top + (0, y, z)^\top \in \mathbb{W} + \mathbb{W}^\perp$. Since $\mathbb{W} \cap \mathbb{W}^\perp = \{\mathbf{0}\}$, $\mathbb{W} + \mathbb{W}^\perp = \mathbb{W} \oplus \mathbb{W}^\perp$, and this representation is unique according to Fact 3.1. ▮

In Section 4.5, we will further show that any arbitrary vector in a finite-dimensional vector space $\mathbb{V}$ can be expressed as a direct sum of two subspaces, any subspace and its orthogonal complement. That is, any vector can be uniquely represented as the sum of two vectors from a subspace and its orthogonal complement.

Another problem we frequently run into is to determine whether a vector $\mathbf{v}$ is orthogonal to $\mathbb{W} = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$. Instead of checking whether $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for every $\mathbf{w} \in \mathbb{W}$, it is enough to check whether $\langle \mathbf{v}, \mathbf{w}_j \rangle = 0$ for each $\mathbf{w}_j$.

Lemma 4.3 *Let $\mathbb{V}$ be a vector space and $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ be a set of some vectors in $\mathbb{V}$. For any $\mathbf{v} \in \mathbb{V}$,*

$$\mathbf{v} \perp \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \quad \textit{if and only if} \quad \mathbf{v} \perp \mathbf{w}_j \ \textit{ for } j = 1, \ldots, k.$$

Proof: "only if" part is clear, since all $\mathbf{w}_j \in \mathbb{W} = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$. For "if" direction, assume $\mathbf{w} \in \mathbb{W}$. $\mathbf{w}$ should then be written as $\mathbf{w} = x_1\mathbf{w}_1 + \cdots + x_k\mathbf{w}_k$ for some $x_1, \ldots, x_k \in \mathbb{R}$. So,

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, x_1\mathbf{w}_1 + \cdots + x_k\mathbf{w}_k \rangle = x_1\langle \mathbf{v}, \mathbf{w}_1 \rangle + \cdots + x_k\langle \mathbf{v}, \mathbf{w}_k \rangle = 0,$$

and hence $\mathbf{v} \perp \mathbb{W}$. ■

Keep in your mind that we did not put any other condition, such as linear independence, on $\mathbf{w}_1, \ldots, \mathbf{w}_k$ to get this result.

4.3 Orthogonal Projection

4.3.1 Projection to the Direction of a Vector

Consider a vector $\mathbf{w} \neq \mathbf{0}$ in a vector space $\mathbb{V}$. Among all the points on a line passing through the origin along the direction $\mathbf{w}$, let us pick one that is closest to another vector $\mathbf{v}$. We use the norm of a vector, induced by an inner product, to measure the distance. First, we parametrize all the points (vectors) on the line along the direction of $\mathbf{w}$ using λ:

$$\{\lambda \mathbf{w} : \lambda \in \mathbb{R}\}.$$

The distance from $\mathbf{v}$ to an arbitrary point on this line is then $|\lambda \mathbf{w} - \mathbf{v}|$. We can rewrite this as the following:

$$\begin{aligned} |\lambda \mathbf{w} - \mathbf{v}|^2 &= \langle \lambda \mathbf{w} - \mathbf{v}, \lambda \mathbf{w} - \mathbf{v} \rangle \\ &= \lambda^2 \langle \mathbf{w}, \mathbf{w} \rangle - 2\lambda \langle \mathbf{v}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle. \end{aligned}$$

This quadratic form is minimized with

$$\lambda^* = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle}.$$

The vector on a line along $\mathbf{w}$, that is nearest to $\mathbf{v}$ is then[3]

$$\lambda^* \mathbf{w} = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} = \left\langle \mathbf{v}, \frac{1}{|\mathbf{w}|} \mathbf{w} \right\rangle \frac{1}{|\mathbf{w}|} \mathbf{w}.$$

According to our intuition from Euclidean spaces, the vector that represents the shortest distance, $\mathbf{v} - \lambda^* \mathbf{w}$, and $\mathbf{w}$ should be orthogonal. Indeed so, this holds in an arbitrary vector space equipped with an inner product as well, since

$$\langle \mathbf{v} - \lambda^* \mathbf{w}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle - \lambda^* \langle \mathbf{w}, \mathbf{w} \rangle = 0.$$

We thus call $\lambda^* \mathbf{w}$ the orthogonal projection of $\mathbf{v}$ onto $\mathbf{w}$.[4] In a vector space with an inner product, we project a vector onto a line and obtain the nearest vector

[3] Because $\mathbf{w}$'s role here is to identify the direction, its norm/magnitude is not essential. In order to make it more concise, we can start from a unit vector $\mathbf{w}$, i.e., $|\mathbf{w}| = \langle \mathbf{w}, \mathbf{w} \rangle = 1$.

[4] Since we only deal with orthogonal projection in this book, we will sometimes omit orthogonal and simply say projection, for the brevity.

on the line to the original vector. We represent such orthogonal projection of $\mathbf{v}$ onto $\mathbf{w}$ as

$$\mathbf{p}_{\mathbf{w}}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} \,. \tag{4.4}$$

If $\mathbf{w}$ was a unit vector, i.e. $\langle \mathbf{w}, \mathbf{w} \rangle = 1$, the representation can be simplified as

$$\mathbf{p}_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{w} \rangle \mathbf{w} \,. \tag{4.5}$$

Because $\langle \mathbf{v}, \mathbf{w} \rangle$ is linear in $\mathbf{v}$, the orthogonal projection $\mathbf{p}_{\mathbf{w}}(\cdot)$ is also linear. In addition, the resulting vector from orthogonal projection remains the same even after further orthogonal projection onto the same direction, since

$$\mathbf{p}_{\mathbf{w}}(\mathbf{p}_{\mathbf{w}}(\mathbf{v})) = \frac{\langle \mathbf{p}_{\mathbf{w}}(\mathbf{v}), \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} = \frac{\langle \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} = \mathbf{p}_{\mathbf{w}}(\mathbf{v}) \,.$$

Because of this property, orthogonal projection is idempotent.[5] Orthogonal projection is the only idempotent one among linear transformations. We show later that a symmetric, idempotent matrix identifies an orthogonal projection.

We have introduced an orthogonal projection onto a one-dimensional subspace spanned by a single vector. We now generalize it to the orthogonal projection onto a subspace as follows.

Definition 4.7 *An orthogonal projection onto a subspace $\mathbb{W}$ is a linear transform that maps any vector to another vector in a subspace $\mathbb{W}$ such that the direction connecting the two vectors is orthogonal to the subspace $\mathbb{W}$. We denote the projection by $\mathbf{P}_{\mathbb{W}}$. In the Euclidean vector space, a matrix is called an orthogonal projection matrix if the matrix represents a projection.*

Orthogonal projection of any vector as defined in in Definition 4.7 is unique by Theorem 4.2. Orthogonal projection is also linear. It is easy to show that scalar multiplication is preserved by $\mathbf{P}_{\mathbb{W}}$, and if $\mathbf{P}_{\mathbb{W}}(\mathbf{v}_1) = \mathbf{w}_1$ and $\mathbf{P}_{\mathbb{W}}(\mathbf{v}_2) = \mathbf{w}_2$, then $\mathbf{P}_{\mathbb{W}}(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{w}_1 + \mathbf{w}_2$ since $\mathbf{v}_1 + \mathbf{v}_2 - (\mathbf{w}_1 + \mathbf{w}_2) = (\mathbf{v}_1 - \mathbf{w}_1) + (\mathbf{v}_2 - \mathbf{w}_2) \in \mathbb{W}^{\perp}$. This linearity allows us to represent orthogonal projection in the Euclidean space as matrix-vector multiplication.

Similarly to the earlier case with orthogonal projection onto a one-dimensional space, we can ask what $\mathbf{P}_{\mathbb{W}}(\mathbf{w})$ is for a vector $\mathbf{w} \in \mathbb{W}$. If we choose $\mathbf{w}$ as a candidate of $\mathbf{P}_{\mathbb{W}}(\mathbf{w})$, the direction $\mathbf{w} - \mathbf{P}_{\mathbb{W}}(\mathbf{w}) = \mathbf{0}$ is trivially orthogonal to $\mathbb{W}$.

[5] We call a matrix or a function idempotent when the square of the matrix or the composition of the function with itself are the same as the original one.

Thus, a projection of a vector in $\mathbb{W}$ onto $\mathbb{W}$ is just the vector itself, and because of the uniqueness of orthogonal projection, $\mathbf{P}_{\mathbb{W}}(\mathbf{w}) = \mathbf{w}$ holds for $\mathbf{w} \in \mathbb{W}$. In other words, $\mathbf{P}_{\mathbb{W}}(\mathbf{P}_{\mathbb{W}}(\mathbf{v})) = \mathbf{P}_{\mathbb{W}}(\mathbf{v})$ for any vector $\mathbf{v}$, since $\mathbf{P}_{\mathbb{W}}(\mathbf{v}) \in \mathbb{W}$, that is, $\mathbf{P}_{\mathbb{W}} \circ \mathbf{P}_{\mathbb{W}} = \mathbf{P}_{\mathbb{W}}$. Orthogonal projection onto a subspace spanned by more than one vectors is thus idempotent, and we can state it as $P^2 = P$ for a corresponding projection matrix P.

Fact 4.8 *In the n-dimensional Euclidean space $\mathbb{R}^n$, the orthogonal projection matrix onto a vector $\mathbf{w}$ is*

$$P = \frac{1}{\mathbf{w}^\top \mathbf{w}} \mathbf{w}\mathbf{w}^\top. \tag{4.6}$$

This projection matrix is symmetric and idempotent, as shown earlier, that is, $P^\top = P$ and $P^2 = P$.

Proof: Consider a $\mathbf{v} \in \mathbb{R}^n$. By rearranging (4.4) in $\mathbb{R}^n$, we observe

$$\mathbf{p}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle} \mathbf{w} = \frac{1}{\mathbf{w}^\top \mathbf{w}} (\mathbf{v}^\top \mathbf{w}) \mathbf{w} = \frac{1}{\mathbf{w}^\top \mathbf{w}} \mathbf{w} (\mathbf{w}^\top \mathbf{v}) = \frac{1}{\mathbf{w}^\top \mathbf{w}} (\mathbf{w}\mathbf{w}^\top) \mathbf{v}$$

where the last term is a multiplication of a matrix $\frac{1}{\mathbf{w}^\top \mathbf{w}} \mathbf{w}\mathbf{w}^\top$ and a vector $\mathbf{v}$. Therefore, the projection matrix is a rank-one matrix, as in

$$P = \frac{1}{\mathbf{w}^\top \mathbf{w}} \mathbf{w}\mathbf{w}^\top .$$

It is thus trivial that P is symmetric, and $P = P^2$, because

$$P^2 = \frac{1}{(\mathbf{w}^\top \mathbf{w})^2} \big(\mathbf{w}\mathbf{w}^\top\big)\big(\mathbf{w}\mathbf{w}^\top\big) = \frac{1}{(\mathbf{w}^\top \mathbf{w})^2} \mathbf{w}\big(\mathbf{w}^\top \mathbf{w}\big)\mathbf{w}^\top = \frac{\mathbf{w}^\top \mathbf{w}}{(\mathbf{w}^\top \mathbf{w})^2} \mathbf{w}\mathbf{w}^\top = P\,.$$

■

4.3.2 Projection onto Orthonormal Vectors

Let $\mathbb{W}$ be a subspace of a finite-dimensional vector space $\mathbb{V}$, and $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ be an orthonormal basis of $\mathbb{W}$. That is, $\langle \mathbf{q}_j, \mathbf{q}_j \rangle = 1$ and $\langle \mathbf{q}_i, \mathbf{q}_j \rangle = 0$ for $i \neq j$. We can write arbitrary $\mathbf{w} \in \mathbb{W}$ as a linear combination of these basic vectors:

$$\mathbf{w} = x_1 \mathbf{q}_1 + \cdots + x_k \mathbf{q}_k = \sum_{i=1}^{k} x_i \mathbf{q}_i \quad \in \mathbb{W}.$$

For a given $\mathbf{v} \in \mathbb{V}$, we consider a problem of finding $\mathbf{w} \in \mathbb{W}$ that is orthogonal to $\mathbf{v} - \mathbf{w}$. This is equivalent to finding its coordinate x_i, and the inner product in the vector space plays an important role in determining these coordinate values.

According to Lemma 4.3, saying that the direction of $\mathbf{v}-\mathbf{w}$ is orthogonal to the subspace $\mathbb{W}$ is equivalent to saying that $\mathbf{v}-\mathbf{w}$ is orthogonal to every $\mathbf{q}_j$. We can thus get $x_j = \langle \mathbf{q}_j, \mathbf{v}\rangle$ from $\langle \mathbf{q}_j, \mathbf{v}-\mathbf{w}\rangle = 0$, because

$$\langle \mathbf{q}_j, \mathbf{v}-\mathbf{w}\rangle = \Big\langle \mathbf{q}_j, \mathbf{v}-\sum_{i=1}^{k} x_i\mathbf{q}_i\Big\rangle = \langle \mathbf{q}_j, \mathbf{v}\rangle - \sum_{i=1}^{k} x_i\langle \mathbf{q}_j, \mathbf{q}_i\rangle = \langle \mathbf{q}_j, \mathbf{v}\rangle - x_j = 0.$$

That is, if we set

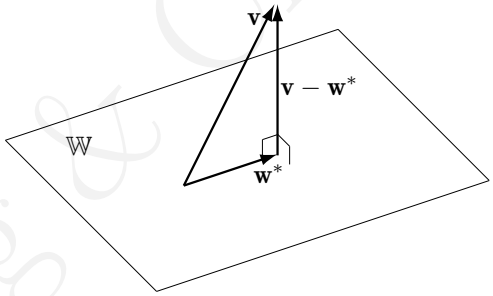
$$\mathbf{w}^* = \langle \mathbf{v}, \mathbf{q}_1\rangle\mathbf{q}_1 + \cdots + \langle \mathbf{v}, \mathbf{q}_k\rangle\mathbf{q}_k = \sum_{i=1}^{k}\langle \mathbf{v}, \mathbf{q}_i\rangle\mathbf{q}_i\,,$$

$\mathbf{v}-\mathbf{w}^*$ is orthogonal to all the basic vectors of $\mathbb{W}$ and therefore it is orthogonal to $\mathbb{W}$.

We can decompose $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{w}^* + (\mathbf{v}-\mathbf{w}^*).$$

Because $\mathbf{w}^* \in \mathbb{W}$ and $\mathbf{v}-\mathbf{w}^* \in \mathbb{W}^\perp$, we see that this decomposition is unique, according to Theorem 4.2.



In other words, there exists a unique line orthogonal to $\mathbb{W}$ that meets $\mathbf{v}$. According to Definition 4.7, this lines passes the vector in $\mathbb{W}$, and this vector is the projection of $\mathbf{v}$ onto $\mathbb{W}$. We use $\mathbf{P}_{\mathbb{W}}(\mathbf{v})$ to refer to this vector and can express it as

$$\mathbf{P}_{\mathbb{W}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{q}_1\rangle\mathbf{q}_1 + \cdots + \langle \mathbf{v}, \mathbf{q}_k\rangle\mathbf{q}_k = \sum_{i=1}^{k}\langle \mathbf{v}, \mathbf{q}_i\rangle\mathbf{q}_i, \tag{4.7}$$

where $\mathbb{W} = \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$.

It is crucial that these spanning vectors are orthonormal, for us to obtain this succient representation of $\mathbf{P}_{\mathbb{W}}$. When there is only one basic vector, i.e., $k = 1$,

(4.7) and (4.5) are equivalent. Because $\langle \mathbf{v}, \mathbf{q}_i \rangle \mathbf{q}_i$ is linear in $\mathbf{v}$, the orthogonal projection $\mathbf{P}_{\mathbb{W}}(\mathbf{v})$ is also linear in $\mathbf{v}$, and we can derive the projection matrix from (4.7).

Fact 4.9 *Let $\mathbf{P}_{\mathbb{W}}$ be an orthogonal projection onto a subspace $\mathbb{W}$ spanned by orthonormal basic vectors, $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} \subset \mathbb{R}^n$. Define Q as the $n \times k$ matrix whose i-th column is $\mathbf{q}_i$. Then $Q^\top Q = I_k$, and the projection matrix of $\mathbf{P}_{\mathbb{W}}$ can be expressed as*

$$P = QQ^\top. \tag{4.8}$$

This matrix is symmetric and satisfies $P^2 = P$.

Proof: Let $\mathbf{v} \in \mathbb{R}^n$ be a vector we want to project. Using the standard inner-product, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$, we can express the projection in the form of matrix-vector multiplication, as follows:

$$\begin{aligned} \mathbf{P}_{\mathbb{W}}(\mathbf{v}) &= \mathbf{q}_1 \langle \mathbf{q}_1, \mathbf{v} \rangle + \cdots + \mathbf{q}_k \langle \mathbf{q}_k, \mathbf{v} \rangle \\ &= \mathbf{q}_1 \mathbf{q}_1^\top \mathbf{v} + \cdots + \mathbf{q}_k \mathbf{q}_k^\top \mathbf{v} \\ &= \left(\mathbf{q}_1 \mathbf{q}_1^\top + \cdots + \mathbf{q}_k \mathbf{q}_k^\top \right) \mathbf{v} \\ &= QQ^\top \mathbf{v} \qquad \text{by Lemma 3.5.} \end{aligned}$$

The symmetry and $P^2 = P$ can be easily checked. ■

Fact 4.9 shows that a projection matrix P is symmetric and satisfies $P^2 = P$. We will see in Fact 4.16 that any matrix with these two properties is a projection matrix.

4.3.3 Projection onto Independent Vectors

Let a subspace $\mathbb{W}$ of a finite-dimensional vector space $\mathbb{V}$ be spanned by linearly independent vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$. They may not be orthonormal. We desire to get $\mathbf{w} \in \mathbb{W}$, that satisfies $(\mathbf{v} - \mathbf{w}) \perp \mathbb{W}$ to project the vector $\mathbf{v} \in \mathbb{V}$ onto the subspace $\mathbb{W}$. Because $\mathbb{W} = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$,

$$(\mathbf{v} - \mathbf{w}) \perp \mathbb{W} \quad \Leftrightarrow \quad (\mathbf{v} - \mathbf{w}) \perp \mathbf{w}_i \ \text{ for all } i = 1, \ldots, k$$

according to Lemma 4.3. We can represent $\mathbf{w}$ as $\mathbf{w} = x_1 \mathbf{w}_1 + \cdots + x_k \mathbf{w}_k$ with an appropriate choice of coefficients, $\mathbf{x} = (x_1, \ldots, x_k)^\top \in \mathbb{R}^k$. We combine these two to get

$$0 \quad = \quad \langle \mathbf{w}_i, \mathbf{w} - \mathbf{v} \rangle$$

$$\begin{aligned} &= \langle \mathbf{w}_i,\ x_1\mathbf{w}_1 + \cdots + x_k\mathbf{w}_k - \mathbf{v}\rangle \\ &= \langle \mathbf{w}_i,\ x_1\mathbf{w}_1 + \cdots + x_k\mathbf{w}_k\rangle - \langle \mathbf{w}_i, \mathbf{v}\rangle \\ &= x_1\langle \mathbf{w}_i, \mathbf{w}_1\rangle + \cdots + x_k\langle \mathbf{w}_i, \mathbf{w}_k\rangle - \langle \mathbf{w}_i, \mathbf{v}\rangle. \end{aligned}$$

That is, we have to find $(x_1, x_2, \ldots, x_k)$ satisfying

$$x_1\langle \mathbf{w}_i, \mathbf{w}_1\rangle + \cdots + x_k\langle \mathbf{w}_i, \mathbf{w}_k\rangle = \langle \mathbf{w}_i, \mathbf{v}\rangle \quad i = 1, \ldots, k. \tag{4.9}$$

Unlike the case of orthonormal spanning vectors earlier, we cannot determine the coordinates one at a time. Instead, we have to consider the k equations in (4.9) simultaneously. By rewriting (4.9) in a matrix form, the orthogonal projection's coefficient $\mathbf{x}^*$ is a solution to the following linear system

$$B\mathbf{x} = \mathbf{c},$$

where $B = (b_{ij})$ is a $k \times k$ symmetric matrix with $b_{ij} = \langle \mathbf{w}_i, \mathbf{w}_j\rangle$ and $\mathbf{c} = (c_i)$ is a vector with $c_i = \langle \mathbf{w}_i, \mathbf{v}\rangle$. Due to Lemma 4.1, B is invertible. Then, with $\mathbf{x}^* = B^{-1}\mathbf{c}$, we can compute the orthogonal projection $\mathbf{w}^*$, as

$$\mathbf{P}_{\mathbb{W}}(\mathbf{v}) = x_1^*\mathbf{w}_1 + \cdots + x_k^*\mathbf{w}_k \tag{4.10}$$

When the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is orthonormal, the basic vectors are linearly independent. Then, we can compute the projection using this method with $B = I_k$, that is, $\mathbf{x}^* = \mathbf{c}$. In this case of orthonormal basic vectors, it coincides with (4.7).

In the fact below, we learn a more specific way to express projection in the n-dimensional Euclidean space, which is popular in many applications.

Fact 4.10 *Let $\mathbf{P}_{\mathbb{W}}$ be projection of a vector onto a subspace $\mathbb{W}$ spanned by linearly independent vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} \subset \mathbb{R}^n$. Let P be the corresponding projection matrix with respect to the standard orthonormal basis of $\mathbb{R}^n$. If A is an $n \times k$ matrix whose columns are $\mathbf{w}_i$'s,*

$$P = A(A^\top A)^{-1}A^\top. \tag{4.11}$$

Furthermore, (4.6) and (4.8) are special cases of (4.11).

Proof: If we let $b_{ij} = \langle \mathbf{w}_i, \mathbf{w}_j\rangle = \mathbf{w}_i^\top\mathbf{w}_j$, $B = (b_{ij}) = A^\top A$ and B is invertible. For $\mathbf{v} \in \mathbb{R}^n$ to be projected, $\mathbf{c} = A^\top\mathbf{v}$, and $\mathbf{x}^* = B^{-1}\mathbf{c} = (A^\top A)^{-1}A^\top\mathbf{v}$. The projection is then

$$\mathbf{P}_{\mathbb{W}}(\mathbf{v}) = x_1^*\mathbf{w}_1 + \cdots + x_k^*\mathbf{w}_k = A\mathbf{x}^* = A(A^\top A)^{-1}A^\top\mathbf{v},$$

and, therefore, the projection matrix is

$$P = A(A^\top A)^{-1}A^\top.$$

By setting $A = \mathbf{e}_i$ and $B = I_k$ respectively, it is easy to see that (4.6) and (4.8) are special cases of (4.11). ■

Example 4.7 Let S be a 2-dimensional surface in $\mathbb{R}^3$. Consider a point on S, $\mathbf{v} = (1, 2, -1)^\top \in S$. When two tangential vectors of S at $\mathbf{v}$ are $(-1, 0, 1)^\top$ and $(1, 1, 0)^\top$, let us find the projection point of $(2, 3, 0)$ onto the tangential plane of S at $\mathbf{v}$. Assuming $\mathbf{v}$ as the origin, the tangential subspace (plane) is spanned by two tangential vectors. Set the 3×2 matrix A consisting of two tangential vectors: $A = \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$ and the projected point shifted to the new origin is $\mathbf{w} = (2, 3, 0)^\top - (1, 2, -1)^\top = (1, 1, 1)^\top$. Hence, the projection matrix is

$$\begin{aligned}
P &= A(A^\top A)^{-1}A^\top \\
&= \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \left(\begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \\
&= \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \\
&= \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \\
&= \frac{1}{3} \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 1 & 2 \\ 1 & 2 & 1 \end{bmatrix} \\
&= \frac{1}{3} \begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}
\end{aligned}$$

and the projection point $P\mathbf{w} = (\frac{2}{3}, \frac{4}{3}, \frac{2}{3})^\top$. ▮

4.4 Building Orthonormal Basis: Gram-Schmidt Procedure

As we have just learned, we can project a vector onto a subspace spanned by an orthonormal basis by summing the vectors separately projected onto each basic vector. We think in this section how to incrementally find orthonormal vectors using this projection method.

Let $\mathcal{B}_k = \{\mathbf{q}_1, \ldots, \mathbf{q}_k\} \subset \mathbb{W}$ be orthonormal. If $\mathbb{W}_k = \operatorname{span} \mathcal{B}_k$ is a proper subspace of $\mathbb{W}$, i.e., $\mathbb{W}_k \subsetneqq \mathbb{W}$, there exists at least one vector in $\mathbb{W}$ but not in $\mathbb{W}_k$, i.e., $\mathbf{w} \in \mathbb{W} \setminus \mathbb{W}_k$. Using (4.7), we can write the projection of $\mathbf{w}$ onto $\mathbb{W}_k$ as

$$\mathbf{P}_{\mathbb{W}_k}(\mathbf{w}) = \sum_{i=1}^{k} \langle \mathbf{w}, \mathbf{q}_i \rangle \mathbf{q}_i \quad \in \mathbb{W}_k.$$

Also, $\mathbf{w} - \mathbf{P}_{\mathbb{W}_k}(\mathbf{w}) \in \mathbb{W}_k^{\perp}$. Since $\mathbf{w} \notin \mathbb{W}_k$, $\mathbf{w} - \mathbf{P}_{\mathbb{W}_k}(\mathbf{w})$ cannot be $\mathbf{0}$. We now guess the next unit orthonormal vector $\mathbf{q}_{k+1}$ by

$$\mathbf{q}_{k+1} = \frac{1}{\left|\mathbf{w} - \mathbf{P}_{\mathbb{W}_k}(\mathbf{w})\right|} \big(\mathbf{w} - \mathbf{P}_{\mathbb{W}_k}(\mathbf{w})\big).$$

Because $\mathbf{q}_{k+1} \perp \mathcal{B}_k$, $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{\mathbf{q}_{k+1}\}$ is orthonormal, just like $\mathcal{B}_k$. That is, $\mathbb{W}_{k+1}$ spanned by $\mathcal{B}_{k+1}$ is a subspace of $\mathbb{W}$ however with one more dimension than $\mathbb{W}_k$. Because $\mathbb{W}$ is finite-dimensional, we repeat this procedure and get an orthonormal basis of $\mathbb{W}$. We call this iterative process the Gram-Schmidt procedure. We start the Gram-Schmidt procedure by choosing any non-zero vector in $\mathbb{W}$ and normalizing its norm to be 1. This tells us the following fact.

Fact 4.11 *Let $\mathbb{W}$ be a finite-dimensional subspace and $\mathcal{B}$ be a set of orthonormal vectors in $\mathbb{W}$. Then, there exists an orthonormal basis of $\mathbb{W}$ containing $\mathcal{B}$, and we can construct it explicitly.*

4.4.1 Gram-Schmidt Procedure for Given Vectors

Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ be a set of linearly independent vectors. The dimension of a vector space $\mathbb{W}$ spanned by these vectors is k. Just like before with orthonormal vectors, we can use the Gram-Schmidt procedure to find a basis with special properties. That is, we follow the above procedure by setting $\mathbb{W}_i = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_i\}$ and $\mathbf{a}_{i+1} \in \mathbb{W} \setminus \mathbb{W}_i$ for $i < k$.

Set $\mathbf{q}_1 = \frac{1}{|\mathbf{a}_1|}\mathbf{a}_1$, $i = 1$;

1. A set of i orthonormal vectors, $\mathcal{B}_i = \{\mathbf{q}_1, \ldots, \mathbf{q}_i\}$ with $i < k$,
2. Compute

$$\begin{aligned} \mathbf{v}_{i+1} &= \mathbf{a}_{i+1} - \mathbf{P}_{\operatorname{span}\mathcal{B}_i}(\mathbf{a}_{i+1}) \\ &= \mathbf{a}_{i+1} - \big(\langle \mathbf{a}_{i+1}, \mathbf{q}_1\rangle \mathbf{q}_1 + \cdots + \langle \mathbf{a}_{i+1}, \mathbf{q}_i\rangle \mathbf{q}_i\big) \quad (4.12) \\ \mathbf{q}_{i+1} &= \frac{1}{|\mathbf{v}_{i+1}|}\mathbf{v}_{i+1} \end{aligned}$$

3. Update $\mathcal{B}_{i+1} = \mathcal{B}_i \cup \{\mathbf{q}_{i+1}\}$.
4. Set $i \leftarrow i + 1$. Repeat while $i < k$.

$\mathbf{q}_i$'s produced by this procedure satisfy the following properties.

Fact 4.12 *Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ be linearly independent vectors. The Gram-Schmidt procedure above produces an orthonormal basis $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ of the subspace $\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ such that*

$$\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_j\} = \operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_j\} \quad \text{for } j = 1, \ldots, k\,.$$

Furthermore, each $\mathbf{q}_j$ explains some part of $\mathbf{a}_j$ not belonging to $\operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{j-1}\}$, that is,

$$\langle \mathbf{a}_j, \mathbf{q}_j\rangle \neq 0 \quad \text{for } j = 1, \ldots, k. \tag{4.13}$$

Proof: We use mathematical induction on k. It is trivially true with $k = 1$. Assume that the statement holds up to k, that is,

$$\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_j\} = \operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_j\} \quad \text{for all } j \leq k. \tag{4.14}$$

- In the second step of the Gram-Schmidt procedure, $\mathbf{v}_{k+1} = \mathbf{0}$ if $|\mathbf{v}_{k+1}| = 0$. This implies that $\mathbf{a}_{k+1} \in \operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ and that $\mathbf{a}_{k+1} \in \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ due to the assumption in (4.14), which contradicts the linear independence among $\mathbf{a}_j$'s. Therefore, $|\mathbf{v}_{k+1}| \neq 0$.
- If we rewrite the same second step by replacing $\mathbf{v}_{k+1}$ with $|\mathbf{v}_{k+1}|\mathbf{q}_{k+1}$,

$$\mathbf{a}_{k+1} = \langle \mathbf{a}_{k+1}, \mathbf{q}_1\rangle \mathbf{q}_1 + \cdots + \langle \mathbf{a}_{k+1}, \mathbf{q}_k\rangle \mathbf{q}_k + |\mathbf{v}_{k+1}|\mathbf{q}_{k+1}.$$

  Thus, $\mathbf{a}_{k+1} \in \operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\}$

- If we further divide (4.12) in the second step with $|\mathbf{v}_{k+1}|$ after replacing $\mathbf{v}_{k+1}$ with $|\mathbf{v}_{k+1}|\mathbf{q}_{k+1}$,

$$\mathbf{q}_{k+1} = \frac{1}{|\mathbf{v}_{k+1}|}\mathbf{a}_{k+1} - \frac{1}{|\mathbf{v}_{k+1}|}\big(\langle \mathbf{a}_{k+1}, \mathbf{q}_1\rangle \mathbf{q}_1 + \cdots + \langle \mathbf{a}_{k+1}, \mathbf{q}_k\rangle \mathbf{q}_k\big).$$

  In other words, $\mathbf{q}_{k+1}$ is a linear combination of $\mathbf{a}_{k+1}$ and $\mathbf{q}_1, \ldots, \mathbf{q}_k$. According to (4.14), $\mathbf{q}_{k+1} \in \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_{k+1}\}$.

Therefore, $\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_{k+1}\} = \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{k+1}\}$.

Let us show $\langle \mathbf{a}_{k+1}, \mathbf{q}_{k+1}\rangle \neq \mathbf{0}$. If we compute the inner products of both sides of (4.12) and $\mathbf{q}_{k+1}$, we get $\langle \mathbf{a}_{k+1}, \mathbf{q}_{k+1}\rangle = \langle \mathbf{v}_{k+1}, \mathbf{q}_{k+1}\rangle$, because $\mathbf{q}_{k+1} \perp \text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$. Since $\langle \mathbf{v}_{k+1}, \mathbf{q}_{k+1}\rangle = |\mathbf{v}_{k+1}|$, we get $\langle \mathbf{a}_{k+1}, \mathbf{q}_{k+1}\rangle = |\mathbf{v}_{k+1}| > 0$. ■

Example 4.8 Let $\mathbb{V}$ be a vector space consisting of polynomials of degree less than 3. That is, $\mathbb{V} = \{a_0 + a_1 x + a_2 x^2 : a_i \in \mathbb{R}\}$. For an inner product $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)dx$ for $f, g \in \mathbb{V}$, let us find an orthonormal basis starting with linearly independent monomials $\{1, x, x^2\}$.

1. $\mathbf{v}_1 = 1$, $\langle \mathbf{v}_1, \mathbf{v}_1\rangle = \int_{-1}^{1} 1dx = 2$,
   $\mathbf{q}_1 = \frac{1}{|\mathbf{v}_1|}\mathbf{v}_1 = \frac{1}{\sqrt{2}}$

2. $\langle x, \mathbf{q}_1\rangle = \frac{1}{\sqrt{2}}\int_{-1}^{1} x dx = 0$,
   $\mathbf{v}_2 = x - \langle x, \mathbf{q}_1\rangle \mathbf{q}_1 = x$, $\langle \mathbf{v}_2, \mathbf{v}_2\rangle = \int_{-1}^{1} x^2 dx = \frac{2}{3}$,
   $\mathbf{q}_2 = \frac{1}{|\mathbf{v}_2|}\mathbf{v}_2 = \sqrt{\frac{3}{2}}x$.

3. $\langle x^2, \mathbf{q}_1\rangle = \frac{1}{\sqrt{2}}\int_{-1}^{1} x^2 dx = \frac{\sqrt{2}}{3}$, $\langle x^2, \mathbf{q}_2\rangle = \sqrt{\frac{3}{2}}\int_{-1}^{1} x^3 dx = 0$,
   $\mathbf{v}_3 = x^2 - \langle x^2, \mathbf{q}_1\rangle \mathbf{q}_1 - \langle x^2, \mathbf{q}_2\rangle \mathbf{q}_2 = x^2 - \langle x^2, \mathbf{q}_1\rangle \mathbf{q}_1 = x^2 - \frac{1}{3}$, $\langle \mathbf{v}_3, \mathbf{v}_3\rangle = \int_{-1}^{1}\big(x^2 - \frac{1}{3}\big)^2 dx = \frac{1}{15}$,
   $\mathbf{q}_3 = \frac{1}{|\mathbf{v}_3|}\mathbf{v}_3 = \sqrt{15}\big(x^2 - \frac{1}{3}\big)$.

By applying the Gram-Schmidt procedure to $\{1, x, x^2\}$ as above, we obtain the following orthonormal basis;

$$\Big\{\frac{1}{\sqrt{2}},\ \sqrt{\frac{3}{2}}x,\ \sqrt{15}\Big(x^2 - \frac{1}{3}\Big)\Big\}.$$

▮

4.4.2 Projection as Distance Minimization

Consider a k-dimensional subspace $\mathbb{W}$, spanned by an orthonormal basis $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$, of a vector space $\mathbb{V}$. We can use for instance the Gram-Schmidt procedure above to find a basis of the full space $\mathbb{V}$ that includes the orthonormal basis of $\mathbb{W}$. Let $\{\mathbf{q}_1, \ldots, \mathbf{q}_k,\ \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ be the complete orthonormal basis of $\mathbb{V}$. For any vector $\mathbf{u}$ in $\mathbb{V}$, we can express $\mathbf{u}$ as follows, according to Fact 4.5:

$$\mathbf{u} = \langle \mathbf{u}, \mathbf{q}_1 \rangle \mathbf{q}_1 + \cdots + \langle \mathbf{u}, \mathbf{q}_k \rangle \mathbf{q}_k + \langle \mathbf{u}, \mathbf{v}_{k+1} \rangle \mathbf{v}_{k+1} + \cdots + \langle \mathbf{u}, \mathbf{v}_n \rangle \mathbf{v}_n \,.$$

With an arbitrary vector $\mathbf{w} = x_1 \mathbf{q}_1 + \cdots + x_k \mathbf{q}_k \in \mathbb{W}$,

$$\begin{aligned} |\mathbf{u} - \mathbf{w}|^2 &= \big(\langle \mathbf{u}, \mathbf{q}_1 \rangle - x_1\big)^2 + \cdots + \big(\langle \mathbf{u}, \mathbf{q}_k \rangle - x_k\big)^2 + \langle \mathbf{u}, \mathbf{v}_{k+1} \rangle^2 + \cdots + \langle \mathbf{u}, \mathbf{v}_n \rangle^2 \\ &\geq \langle \mathbf{u}, \mathbf{v}_{k+1} \rangle^2 + \cdots + \langle \mathbf{u}, \mathbf{v}_n \rangle^2, \end{aligned}$$

because

$$\mathbf{u} - \mathbf{w} = \big(\langle \mathbf{u}, \mathbf{q}_1 \rangle - x_1\big)\mathbf{q}_1 + \cdots + \big(\langle \mathbf{u}, \mathbf{q}_k \rangle - x_k\big)\mathbf{q}_k + \langle \mathbf{u}, \mathbf{v}_{k+1} \rangle \mathbf{v}_{k+1} + \cdots + \langle \mathbf{u}, \mathbf{v}_n \rangle \mathbf{v}_n.$$

We get the minimal distance $|\mathbf{u} - \mathbf{w}|^2$ when $x_i = \langle \mathbf{u}, \mathbf{q}_i \rangle$ for all i. In other words, $\mathbf{w} \in \mathbb{W}$ that minimizes $|\mathbf{u} - \mathbf{w}|^2$ is

$$\mathbf{w} = \langle \mathbf{u}, \mathbf{q}_1 \rangle \mathbf{q}_1 + \cdots + \langle \mathbf{u}, \mathbf{q}_k \rangle \mathbf{q}_k.$$

According to (4.7), this is precisely the projection of $\mathbf{u}$ onto $\mathbb{W}$, i.e., $\mathbf{P}_{\mathbb{W}}(\mathbf{u})$. Thus, the projection of $\mathbf{u}$ onto $\mathbb{W}$ is the vector in $\mathbb{W}$ closest to $\mathbf{u}$.

4.5 Decomposition into Orthogonal Complements

Given a subspace, we will show that a vector in a finite-dimensional vector space can be expressed as a sum of a vector in the subspace and a vector in the orthogonal complement. Since these two subspaces are orthogonal to each other, we can conclude that any finite-dimensional vector space can be expressed as a direct sum of a subspace and its orthogonal complement. With respect to the given subspace, this decomposition is unique, which makes it particularly useful.

Fact 4.13 *Let $\mathbb{V}$ be a finite-dimensional vector space and $\mathbb{W}$ be a subspace of $\mathbb{V}$. For any $\mathbf{v} \in \mathbb{V}$, there exist unique $\mathbf{w} \in \mathbb{W}$ and $\mathbf{z} \in \mathbb{W}^\perp$ such that $\mathbf{v} = \mathbf{w} + \mathbf{z}$. Therefore, the direct sum representation $\mathbb{V} = \mathbb{W} \oplus \mathbb{W}^\perp$ holds.*

Proof: By Fact 4.11, we can obtain an orthonormal basis of $\mathbb{W}$. Then, for any $\mathbf{v} \in \mathbb{V}$, we get a projection $\mathbf{w} = \mathbf{P}_{\mathbb{W}}(\mathbf{v})$ of $\mathbf{v}$ onto $\mathbb{W}$ through (4.7), and its orthogonal component $\mathbf{z} = \mathbf{v} - \mathbf{w}$ lies in $\mathbb{W}^{\perp}$. Then, the decomposition holds by Theorem 4.2. ■

Any subspace has its orthogonal complement, and thereby it is natural to consider the orthogonal complement of the orthogonal complement of a subspace. That is, $\left(\mathbb{W}^{\perp}\right)^{\perp}$. Consider a subspace $\mathbb{W} = \mathbb{R} \times \{0\}$ of $\mathbb{V} = \mathbb{R}^2 = (\mathbb{R} \times \{0\}) \oplus (\{0\} \times \mathbb{R})$. Then, $\mathbb{W}^{\perp} = \{0\} \times \mathbb{R}$ and $(\mathbb{W}^{\perp})^{\perp} = (\{0\} \times \mathbb{R})^{\perp} = \mathbb{R} \times \{0\} = \mathbb{W}$. We thus conjecture that $(\mathbb{W}^{\perp})^{\perp} = \mathbb{W}$. For a finite-dimensional vector space, in fact, the orthogonal complement of the orthogonal complement of a subspace is the original subspace.[6]

Fact 4.14 *Let $\mathbb{V}$ be a finite-dimensional vector space and $\mathbb{W}$ be a subspace of $\mathbb{V}$. Then, $(\mathbb{W}^{\perp})^{\perp} = \mathbb{W}$.*

Proof: Let $\mathbf{w} \in \mathbb{W}$. For any $\mathbf{z} \in \mathbb{W}^{\perp}$, $\langle \mathbf{w}, \mathbf{z} \rangle = 0$ by the definition of orthogonal complement. Therefore, $\mathbf{w} \in (\mathbb{W}^{\perp})^{\perp}$, that is, $\mathbb{W} \subset (\mathbb{W}^{\perp})^{\perp}$.

Conversely, assume $\mathbf{v} \in (\mathbb{W}^{\perp})^{\perp}$. By Fact 4.13, $\mathbf{v} = \mathbf{w} + \mathbf{z}$ for some $\mathbf{w} \in \mathbb{W}$ and $\mathbf{z} \in \mathbb{W}^{\perp}$, and $\langle \mathbf{w}, \mathbf{z} \rangle = 0$. Since $\mathbf{v} \in (\mathbb{W}^{\perp})^{\perp}$, we know $\mathbf{v} \perp \mathbf{z}$. Then, $0 = \langle \mathbf{v}, \mathbf{z} \rangle = \langle \mathbf{w} + \mathbf{z}, \mathbf{z} \rangle = \langle \mathbf{w}, \mathbf{z} \rangle + \langle \mathbf{z}, \mathbf{z} \rangle = |\mathbf{z}|^2$. Hence, $\mathbf{z} = \mathbf{0}$ and $\mathbf{v} = \mathbf{w} \in \mathbb{W}$, which implies $(\mathbb{W}^{\perp})^{\perp} \subset \mathbb{W}$. ■

4.6 Orthogonality of Fundamental Subspaces

In this section, we only consider a finite-dimensional Euclidean space and the standard inner product between two vectors

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \cdots + x_n y_n = \mathbf{x}^{\top} \mathbf{y}.$$

First, let us look at the orthogonality among four basic subspaces, starting from the definitions of the row and column spaces introduced in Section 3.1.2 and 3.5.

Fact 4.15 *Let A be an $m \times n$ matrix. The row space of A is orthogonal to the nullspace (in $\mathbb{R}^n$). The column space of A is orthogonal to the left nullspace (in $\mathbb{R}^m$).*

[6] This property does not necessarily hold in an infinite-dimensional vector space.

Proof: Let $\mathbf{y} \in \mathrm{Null}\,(A^\top)$ and $\mathbf{b} \in \mathrm{Col}\,(A)$. There exists $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{b} = A\mathbf{x}$. Since $\mathbf{y}^\top A = \mathbf{0}$, $\mathbf{y}^\top \mathbf{b} = \mathbf{y}^\top (A\mathbf{x}) = (\mathbf{y}^\top A)\mathbf{x} = \mathbf{0}^\top \mathbf{x} = 0$. Therefore, $\mathbf{y} \perp \mathbf{b}$, that is, $\mathrm{Null}\,(A^\top) \perp \mathrm{Col}\,(A)$. Considering $A^\top$ instead of A, we can similarly show the orthogonality between the nullspace of A and the row space of A. ■

Because orthogonal subspaces are not necessarily the orthogonal complement of each other, the fact above tells us only that $\mathrm{Row}\,(A) = \mathrm{Col}\,(A^\top) \perp \mathrm{Null}\,(A)$ but neither whether $\mathrm{Null}\,(A) = \mathrm{Col}\,(A^\top)^\perp$ nor whether $\mathrm{Null}\,(A)^\perp = \mathrm{Col}\,(A^\top)$.

4.6.1 Orthogonal Complements of Fundamental Subspaces

We now consider the relationship between subspaces of an $m \times n$ matrix A and their orthogonal complements.

- $\mathrm{Null}\,(A) = \mathrm{Col}\,(A^\top)^\perp$: Let $\mathbf{x} \in \mathrm{Null}\,(A)$, that is, $A\mathbf{x} = \mathbf{0}$. Because the i-th element of $A\mathbf{x}$ is the inner product between the i-th row of A and the vector $\mathbf{x}$, $\mathbf{x}$ is orthogonal to all row vectors of A. $\mathbf{x}$ is thereby orthogonal to all vectors in the row space of A, i.e., $\mathrm{Null}\,(A) \subset \mathrm{Col}\,(A^\top)^\perp$.

  On the other hand, assume $\mathbf{x} \in \mathrm{Col}\,(A^\top)^\perp$. Since all rows of A are included in $\mathrm{Col}\,(A^\top)$, the inner product between $\mathbf{x}$ and any of the rows of A is 0. In other words, because $A\mathbf{x} = \mathbf{0}$, we know that $\mathrm{Col}\,(A^\top)^\perp \subset \mathrm{Null}\,(A)$.

- $\mathrm{Null}\,(A)^\perp = \mathrm{Col}\,(A^\top)$: We derive $\mathrm{Null}\,(A)^\perp = \mathrm{Col}\,(A^\top)$ from $\mathrm{Null}\,(A) = \mathrm{Col}\,(A^\top)^\perp$ and $(\mathrm{Col}\,(A^\top)^\perp)^\perp = \mathrm{Col}\,(A^\top)$ by Fact 4.14.

From these orthogonality results, we conclude that "The nullspace contains everything orthogonal to the row space" and that "The row space contains everything orthogonal to the nullspace". If we swap A and $A^\top$, we end up with "The left nullspace is the orthogonal complement of the column space". We summarize these results as a lemma for reference.

Lemma 4.4 *For any matrix A,*

$$\mathrm{Null}\,(A) = \mathrm{Col}\,(A^\top)^\perp \quad \textit{and} \quad \mathrm{Null}\,(A)^\perp = \mathrm{Col}\,(A^\top)\,.$$

This result is a part of the fundamental theorem of linear algebra.

In Fact 4.9, we identified a matrix corresponding to a projection transformation. The projection matrix P in (4.8) is symmetric, and $P^2 = P$. According to

Lemma 4.4, we can show that these two conditions fully characterize a projection matrix.

Fact 4.16 *P is a matrix representing an orthogonal projection onto a subspace of the Euclidean vector space $\mathbb{R}^n$ if and only if P is symmetric and $P^2 = P$.*

Solution: Let a transformation orthogonally project a vector onto a subspace $\mathbb{W}$. Let A be a matrix whose columns are the linearly independent vectors, and $\mathbb{W}$ is the column space of A. Then, according to Fact 4.10, we can represent the projection onto $\text{Col}\,(A)$ with

$$P = A(A^\top A)^{-1}A^\top$$

which is symmetric and satisfies $P^2 = P$.

Conversely, assume that a symmetric matrix P satisfies $P^2 = P$. Let $\mathbb{W} = \text{Col}\,(P)$. Then, for an arbitrary vector $\mathbf{v}$,

$$(\mathbf{v} - P\mathbf{v}) \in \text{Null}\,(P) = \text{Null}\,(P^\top) = \text{Col}\,(P)^\perp = \mathbb{W}^\perp,$$

because

$$P(\mathbf{v} - P\mathbf{v}) = P\mathbf{v} - P^2\mathbf{v} = \mathbf{0}.$$

That is, $(\mathbf{v} - P\mathbf{v}) \perp \mathbb{W}$. In addition, $P\mathbf{v} \in \text{Col}\,(P) = \mathbb{W}$. Thus, the operator corresponding to P is a projection onto the subspace $\mathbb{W} = \text{Col}\,(P)$. ■

4.7 Orthogonal Matrices

Let the columns of an $m \times n$ matrix Q be orthonormal. Because orthonormal vectors are linearly independent, $m \geq n$. The (i, j)-th entry of $Q^\top Q$ results from the inner product between the column vectors, $\mathbf{q}_i$ and $\mathbf{q}_j$, which is 1 if $i = j$ and 0 otherwise ($i \neq j$). That is, $Q^\top Q = I_n$ and Q has a left inverse. With $m = n$, when all columns of the square matrix Q are linearly independent, Q is invertible and $Q^{-1} = Q^\top$, and we call Q an orthogonal matrix. Since $QQ^{-1} = QQ^\top = I$ in this case, both columns and rows of Q are respectively orthonormal. It is useful to know that a product of orthogonal matrices is also an orthogonal matrix.

Example 4.9 Both the rotation matrix in $\mathbb{R}^2$, $R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ and the permutation matrix in $\mathbb{R}^3$, $Q = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ are orthogonal matrices since

$$R^\top R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \sin^2\theta + \cos^2\theta \end{bmatrix} = I\,,$$

and $Q^\top Q = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I.$ ∎

In fact, (B.3) says that $QQ^\top = I$ for $n \times n$ permutation matrix Q. Therefore any permutation matrix is an orthogonal matrix.

Example 4.10 [Householder matrix] An example of an orthogonal matrix is a Householder matrix representing a reflection, which is created from a projection matrix $P = \frac{1}{\mathbf{v}^\top\mathbf{v}}\mathbf{v}\mathbf{v}^\top$ toward $\mathbf{v} \in \mathbb{R}^n$, defined as follows:

$$H = I - 2P = I - 2\frac{1}{\mathbf{v}^\top\mathbf{v}}\mathbf{v}\mathbf{v}^\top.$$

Since $HH^\top = (I-2P)(I-2P)^\top = (I-2P)^2 = I - 2P - 2P + 4P^2 = I - 2P - 2P + 4P = I$, the Household matrix is orthogonal.

Let us investigate H as a transformation matrix. For any vector $\mathbf{u}$, set $\mathbf{v} = \mathbf{u} + |\mathbf{u}|\,\mathbf{e}_1$. Since $\mathbf{v}^\top\mathbf{u} = \mathbf{u}^\top\mathbf{u} + |\mathbf{u}|\,u_1$,

$$\mathbf{v}^\top\mathbf{v} = \mathbf{u}^\top\mathbf{u} + 2u_1|\mathbf{u}| + |\mathbf{u}|^2 = 2(\mathbf{u}^\top\mathbf{u} + |\mathbf{u}|\,u_1) = 2\mathbf{v}^\top\mathbf{u}$$

holds. Therefore, the transformation corresponding to a Householder matrix H based on a direction $\mathbf{v}$ moves a vector $\mathbf{u}$ to

$$\begin{aligned} H\mathbf{u} &= \Big(I - 2\frac{1}{\mathbf{v}^\top\mathbf{v}}\mathbf{v}\mathbf{v}^\top\Big)\mathbf{u} = \mathbf{u} - 2\frac{1}{\mathbf{v}^\top\mathbf{v}}\mathbf{v}\mathbf{v}^\top\mathbf{u} = \mathbf{u} - \mathbf{v} \\ &= -|\mathbf{u}|\,\mathbf{e}_1 = (-|\mathbf{u}|, 0, \ldots, 0)^\top. \end{aligned}$$

That is, this reflection transformation keeps the norm of $\mathbf{u}$, but aligns the direction along the negative side of x_1 axis. Let us assume that $\mathbf{u}$ be the first column of an $n \times d$ matrix A, $n \geq d$. Since the first column of HA is $-|\mathbf{a}_1|\mathbf{e}_1$,

$$HA = \begin{bmatrix} -|\mathbf{a}_1| & \mathbf{b}_1^\top \\ \mathbf{0}_{n-1} & A_2 \end{bmatrix},$$

where $\mathbf{b}_1 \in \mathbb{R}^{d-1}$ and A_2 is $(n-1) \times (d-a)$. We can repeat this procedure for A_2. If the Household matrix for the first column of A_2, $\tilde{\mathbf{a}}_2 \in \mathbb{R}^{n-1}$ is $\tilde{H}_2$, set

$$H_2 = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{H}_2. \end{bmatrix}$$

If $|\tilde{\mathbf{a}}_2| = 0$, set $\tilde{H}_2 = I_{n-1}$ Then,

$$H_2 H A = \begin{bmatrix} -|\mathbf{a}_1| & \mathbf{b}_1^\top \\ \mathbf{0}_{n-1} & \tilde{H}_2 A_2 \end{bmatrix} = \begin{bmatrix} -|\mathbf{a}_1| & & \mathbf{b}_1^\top \\ & \begin{bmatrix} -|\tilde{\mathbf{a}}_2| & \mathbf{b}_2^\top \\ \mathbf{0}_{n-2} & A_3 \end{bmatrix} & \\ \mathbf{0}_{n-1} & & \end{bmatrix}.$$

It is easy to confirm that H_2 is also orthogonal. By repeating further, we arrive at

$$H_{d-1} H_{d-2} \cdots H_2 H A = R,$$

where R is an $n \times d$ upper triangular matrix. By setting $Q = H^\top H_2^\top \cdots H_{d-2}^\top H_{d-1}^\top$, we get $A = QR$ where Q is orthogonal. By choosing the first d columns of Q and the first d rows of R, we get the so-called QR-decomposition by Householder transformation. Refer to [8] for more details. We present another way of obtaining QR-decomposition by Gram-Schmidt procedure in Section 4.7.1. ▮

As we have seen so far, typical examples of orthogonal matrices correspond to geometric transformations, such as rotation, reflection and their combination. Another important example of orthogonal matrix is the permutation matrix, which is shown in Appendix B. They are however not exhaustive.

4.7.1 QR-Decomposition by Gram-Schmidt Procedure

Consider an $m \times n$ matrix $A = \begin{bmatrix} \mathbf{a}_1 | \mathbf{a}_2 | \ldots | \mathbf{a}_n \end{bmatrix}$ whose columns are linearly independent. Because the columns are linearly independent, $m \geq n$. Furthermore, let $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ be an orthonormal basis obtained from the linearly independent column vectors using the Gram-Schmidt procedure from Section 4.4.1. Say $Q = \begin{bmatrix} \mathbf{q}_1 \mid \mathbf{q}_2 \mid \cdots \mid \mathbf{q}_n \end{bmatrix}$. According to Fact 4.12, each column $\mathbf{a}_j$ is in $\text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_j\}$ for all $j = 1, \ldots, n$, and can be expressed as

$$\mathbf{a}_j = \langle \mathbf{a}_j, \mathbf{q}_1 \rangle \mathbf{q}_1 + \langle \mathbf{a}_j, \mathbf{q}_2 \rangle \mathbf{q}_2 + \cdots + \langle \mathbf{a}_j, \mathbf{q}_j \rangle \mathbf{q}_j.$$

These equations can be collectively rewritten as

$$A = \begin{bmatrix} \mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n \end{bmatrix} \begin{bmatrix} \langle \mathbf{a}_1, \mathbf{q}_1 \rangle & \langle \mathbf{a}_2, \mathbf{q}_1 \rangle & \cdots & \langle \mathbf{a}_n, \mathbf{q}_1 \rangle \\ 0 & \langle \mathbf{a}_2, \mathbf{q}_2 \rangle & \cdots & \langle \mathbf{a}_n, \mathbf{q}_2 \rangle \\ 0 & 0 & \cdots & \langle \mathbf{a}_n, \mathbf{q}_3 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \mathbf{a}_n, \mathbf{q}_n \rangle \end{bmatrix}.$$

Let us denote the last upper triangular matrix as R. Then, from these, we obtain so-called QR-decomposition of A as

$$A = QR\,.$$

Q and R are a matrix with orthonormal columns and an upper triangular matrix, respectively. It also holds that $\langle \mathbf{a}_j, \mathbf{q}_j \rangle \neq 0$ because of (4.13), and hence R has non-zero diagonal entries. Therefore, R is invertible, as it is upper triangular and does not have any zero in its diagonal. When there are negative entries on the diagonal of R, we can create another diagonal matrix D such that $d_{ii} = -1$ if $\langle \mathbf{a}_i, \mathbf{q}_i \rangle < 0$ and otherwise $d_{ii} = 1$. In this case, the following three properties hold; $D^2 = I$, DR is upper diagonal with positive diagonal entries, and the columns of QD continue to be orthonormal. We can therefore assume that all the diagonal entries of the upper triangular matrix are positive by decomposing A into $A = QR = (QD)(DR)$. Hereafter, we assume that the upper triangular matrix R of QR-decomposition has positive diagonals. In addition, if $m = n$, Q is further an orthogonal matrix.

4.7.2 Isometry induced by Orthogonal Matrix

Here we consider a linear transformation $T : \mathbf{x} \in \mathbb{R}^n \to T(\mathbf{x}) = Q\mathbf{x} \in \mathbb{R}^n$ with an orthogonal matrix Q. With $\langle \cdot, \cdot \rangle$ as the standard inner product in $\mathbb{R}^n$,

$$\big\langle T(\mathbf{x}), T(\mathbf{y}) \big\rangle = \langle Q\mathbf{x}, Q\mathbf{y} \rangle = \big(Q\mathbf{x}\big)^\top Q\mathbf{y} = \mathbf{x}^\top Q^\top Q\mathbf{y} = \mathbf{x}^\top \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle.$$

That is, for any pair of vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\big\langle T(\mathbf{x}), T(\mathbf{y}) \big\rangle = \langle \mathbf{x}, \mathbf{y} \rangle. \tag{4.15}$$

In other words, the standard inner product is preserved under orthogonal transformation. In the case of $\mathbf{x} = \mathbf{y}$, the norm induced by the standard inner product is also preserved, as

$$\big|T(\mathbf{x})\big|^2 = \big\langle T(\mathbf{x}), T(\mathbf{x}) \big\rangle = \langle \mathbf{x}, \mathbf{x} \rangle = |\mathbf{x}|^2.$$

We call such a transformation that preserves the norm isometry.

When a linear transformation ℓ preserves the norm induced by an inner product, i.e., $|\ell(\mathbf{x})| = |\mathbf{x}|$, the inner product is also preserved, i.e., $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \ell(\mathbf{x}), \ell(\mathbf{y}) \rangle$, because

$$\begin{aligned} & |\mathbf{x}|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + |\mathbf{y}|^2 \\ = \;& |\mathbf{x} + \mathbf{y}|^2 \\ = \;& \big|\ell(\mathbf{x} + \mathbf{y})\big|^2 \\ = \;& \big|\ell(\mathbf{x}) + \ell(\mathbf{y})\big|^2 \\ = \;& \big|\ell(\mathbf{x})\big|^2 + 2\big\langle \ell(\mathbf{x}), \ell(\mathbf{y}) \big\rangle + \big|\ell(\mathbf{y})\big|^2 \\ = \;& |\mathbf{x}|^2 + 2\big\langle \ell(\mathbf{x}), \ell(\mathbf{y}) \big\rangle + |\mathbf{y}|^2. \end{aligned}$$

In other words, preserving an inner product and preserving the norm induced by the inner product are equivalent under linear transformation.

Among the linear transformations from Section 3.8.2, following ones are isometries.

- Rotation: it is expected to be an isometry as a rotation preserves the norm and inner-product. Indeed, in $\mathbb{R}^2$, R_θ is orthogonal, because $R_\theta^\top R_\theta = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ with $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$.
- Reflection: We can expect a reflection matrix to be orthogonal, since reflection preserves an inner product and the associated norm. Indeed, $H^\top = H^{-1}$, because $H^\top = H$ and $H^2 = I$ for a reflection matrix H.

4.8 Matrix Norms

Let us start by defining the trace of an $n \times n$ matrix $A = (a_{ij})$ as the sum of its diagonal entries,

$$\operatorname{trace} A = a_{11} + \cdots + a_{nn}.$$

In the vector space of $m \times n$ matrices, we can show that

$$\langle A, B \rangle = \operatorname{trace}(A^\top B)$$

for two $m \times n$ matrices A and B is an inner product. Then, $\sqrt{\langle A, A \rangle} = \sqrt{\operatorname{trace}(A^\top A)}$ is the norm induced by this inner product. We call this matrix norm a Frobenius norm.

There are many other ways to define a norm for matrices. Another widely used one is a spectral norm. The spectral norm of a matrix measures how much a transformation induced by the matrix changes a vector. We do not derive it from an inner product but define the spectral norm directly below.

- Frobenius norm: This norm is a direct application of the Euclidean vector norm to a matrix, as it computes the sum of squared entries. That is,[7]

$$\|A\|_F = \sqrt{\text{trace}(AA^\top)} = \sqrt{\text{trace}(A^\top A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

  Based on this definition, we can see that $\|A\|_F = \|A^\top\|_F$. For a matrix V with orthogonal columns (that is, $V^\top V = I$) and a matrix A, the following holds:

$$\|AV^\top\|_F^2 = \text{trace}(AV^\top(AV^\top)^\top) = \text{trace}(AV^\top VA^\top) = \text{trace}(AA^\top) = \|A\|_F^2.$$

  Similarly, for a matrix U with orthogonal rows, $\|UA\|_F = \|A\|_F$.

- Spectral norm: This matrix norm measures how much unit vectors change through the linear transformation defined by the matrix. That is, we define the spectral norm of a matrix A by the following maximization problem

$$\max_{|\mathbf{x}|\leq 1} |A\mathbf{x}|.$$

  It is not apparent whether an optimal $\mathbf{x}^*$ exists and whether we can find such $\mathbf{x}^*$ in this optimization problem. The existence of such $\mathbf{x}^*$ and also that the norm of such $\mathbf{x}^*$ is 1 are shown in Lemma C.3. This result simplifies the optimization problem above and leads to the following definition of the spectral norm:

$$\|A\|_2 = \max_{|\mathbf{x}|\leq 1} |A\mathbf{x}| = \max_{|\mathbf{x}|=1} |A\mathbf{x}|. \tag{4.16}$$

  As examples, The spectral norms of rotation and reflection are both 1 since they induce isometries. Because $|A\big(\frac{1}{|\mathbf{x}|}\mathbf{x}\big)| \leq \|A\|_2$ for $\mathbf{x} \neq \mathbf{0}$, it holds for an arbitrary vector $\mathbf{x} \in \mathbb{R}^d$ that

$$|A\mathbf{x}| \;\leq\; \|A\|_2\, |\mathbf{x}|. \tag{4.17}$$

[7] If AB and BA are both defined for a pair of matrices A and B, their traces match, that is, $\text{trace}(AB) = \text{trace}(BA)$.

Here are some of interesting properties of matrix norms:

Fact 4.17 *For $n \times d$ matrix A and $d \times m$ matrix B,*

1. $\|AB\|_2 \le \|A\|_2 \|B\|_2$;
2. $\|AB\|_F \le \|A\|_F \|B\|_F$;
3. $\|A\|_2 \le \|A\|_F$;
4. *For an orthogonal matrix Q,* $\|QA\|_2 = \|A\|_2$ *and* $\|QA\|_F = \|A\|_F$;
5. *If $A = \mathbf{u}^\top$ or $A = \mathbf{u}$ for an n-dimensional vector $\mathbf{u}$,* $\|A\|_2 = \|A\|_F = |\mathbf{u}|$.

Proof:

1. By (4.17), $|AB\mathbf{x}| \le \|A\|_2 |B\mathbf{x}| \le \|A\|_2 \|B\|_2 |\mathbf{x}|$ for any $\mathbf{x}$.
2. Let $a_{i\bullet} = (a_{i1}, \ldots, a_{id})^\top$ and $b_{\bullet j} = (b_{1j}, \ldots, b_{dj})^\top$. Then,
$$\begin{aligned} \|AB\|_F^2 &= \sum_i \sum_j (a_{i\bullet}^\top b_{\bullet j})^2 \\ &\le \sum_i \sum_j |a_{i\bullet}|^2 |b_{\bullet j}|^2 = \sum_i |a_{i\bullet}|^2 \sum_j |b_{\bullet j}|^2 = \|A\|_F^2 \|B\|_F^2. \end{aligned}$$
3. For $|\mathbf{x}| \le 1$, $|A\mathbf{x}|^2 = \sum_{i=1}^n (a_{i\bullet}^\top \mathbf{x})^2 \le \sum_{i=1}^n |a_{i\bullet}|^2 |\mathbf{x}|^2 \le \sum_{i=1}^n |a_{i\bullet}|^2 = \|A\|_F^2$.
4. $|QA\mathbf{x}| = \sqrt{\mathbf{x}^\top A^\top Q^\top Q A \mathbf{x}} = \sqrt{\mathbf{x}^\top A^\top A \mathbf{x}} = |A\mathbf{x}|$, $\|QA\|_F^2 = \mathrm{trace}(A^\top Q^\top Q A)$ $= \mathrm{trace}(A^\top A) = \|A\|_F^2$.
5. It is clear that $\|A\|_F = |\mathbf{u}|$ from the definition of Frobenius norm. By Cauchy-Schwarz inequality, $|\mathbf{u}^\top \mathbf{x}| \le |\mathbf{u}|$ for any n-dimensional vector $\mathbf{x}$ with $|\mathbf{x}| = 1$. If we let $\mathbf{x} = \frac{1}{|\mathbf{u}|}\mathbf{u}$, then $A\mathbf{x} = \mathbf{u}^\top \mathbf{x} = \frac{1}{|\mathbf{u}|}\mathbf{u}^\top \mathbf{u} = |\mathbf{u}|$. Hence, $\|A\|_2 = |\mathbf{u}|$. A similar derivation works for $A = \mathbf{u}$.

■

Fact 4.18 *Every $m \times n$ rank-one matrix can be represented as $\mathbf{u}\mathbf{v}^\top$ for some $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$. An upper bound on both the Frobenius and spectral norms of a rank-one matrix $\mathbf{u}\mathbf{v}^\top$ is $|\mathbf{u}||\mathbf{v}|$.*

Proof: For any n-vector $\mathbf{x}$,

$$\begin{aligned} |(\mathbf{u}\mathbf{v}^\top)\mathbf{x}| &= |\mathbf{u}(\mathbf{v}^\top\mathbf{x})| \\ &\le \|\mathbf{u}\||\mathbf{v}^\top\mathbf{x}| && \text{(by regarding } \mathbf{u} \text{ as an } m\times 1 \text{ matrix)} \\ &\le \|\mathbf{u}\|\|\mathbf{v}^\top\||\mathbf{x}| && \text{(by regarding } \mathbf{v}^\top \text{ as an } 1\times n \text{ matrix)} \\ &= |\mathbf{u}||\mathbf{v}||\mathbf{x}| && \text{(by 5 of Fact 4.17),} \end{aligned}$$

regardless of the type of the norm $\|\cdot\|$. ■

4.9 Application: Least Square and Projection

Both in natural sciences and engineering, there are many cases in which data is expressed as the relationship between an explanatory/feature vector $\mathbf{z} = (z_1, \ldots, z_k)^\top \in \mathbb{R}^k$ and a dependent variable y. Consider a case where such a relationship is in the form of

$$y = \theta_1 f_1(\mathbf{z}) + \cdots + \theta_n f_n(\mathbf{z}),$$

with appropriate functions, $f_1, \ldots, f_n : \mathbb{R}^k \to \mathbb{R}$, and constants, θ_j. We assume each f_i is known and can be computed exactly. Then, instead of $\mathbf{z}$ from $(\mathbf{z}, y) \in \mathbb{R}^k \times \mathbb{R}$, we can use $\mathbf{x} = (x_1, \ldots, x_n)^\top = (f_1(\mathbf{z}), \ldots, f_n(\mathbf{z}))^\top \in \mathbb{R}^n$ and assume the following linear model of relationship:

$$y = \theta_1 x_1 + \cdots + \theta_n x_n = \boldsymbol{\theta}^\top \mathbf{x}\,,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. In reality, however, there often exists measurement error ε_i for each data pair $(\mathbf{a}_i, b_i)$ for $i = 1, \ldots, m$, and we assume that such measurement error is additive:

$$b_i = \boldsymbol{\theta}^\top \mathbf{a}_i + \varepsilon_i\,,\ i = 1, \ldots, m\,.$$

If we for now ignore measurement noise and express this linear relationship in terms of an $m \times n$ data matrix A, of which i-th row is $\mathbf{a}_i^\top$, and a vector $\mathbf{b} = (b_1, \ldots, b_m)^\top$, we obtain the following linear system:

$$A\boldsymbol{\theta} = \mathbf{b}.$$

The problem is then to find $\boldsymbol{\theta}$ that satisfies this linear system, although there may not be $\boldsymbol{\theta}$ that satisfies $b_i = \boldsymbol{\theta}^\top \mathbf{a}_i$ due to measurement noise ε_i. That is, it may be that $\mathbf{b} \notin \mathrm{Col}\,(A)$.

Alternatively, we can approach this problem of finding $\boldsymbol{\theta}$ that minimizes $\varepsilon_i = \boldsymbol{\theta}^\top \mathbf{a}_i - b_i$. That is,[8]

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^n} \sum_{i=1}^{m} \varepsilon_i^2 = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^n} \sum_{i=1}^{m} (\boldsymbol{\theta}^\top \mathbf{a}_i - b_i)^2 = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^n} \left| A\boldsymbol{\theta} - \mathbf{b} \right|^2. \tag{4.18}$$

4.9.1 Least Square as a Convex Quadratic Minimization

Let us expand the objective function $|A\boldsymbol{\theta} - \mathbf{b}|^2$ above:

$$\begin{aligned} |A\boldsymbol{\theta} - \mathbf{b}|^2 &= \langle A\boldsymbol{\theta} - \mathbf{b}, A\boldsymbol{\theta} - \mathbf{b} \rangle \\ &= \langle A\boldsymbol{\theta}, A\boldsymbol{\theta} \rangle - \langle A\boldsymbol{\theta}, \mathbf{b} \rangle - \langle \mathbf{b}, A\boldsymbol{\theta} \rangle + \langle \mathbf{b}, \mathbf{b} \rangle \\ &= (A\boldsymbol{\theta})^\top A\boldsymbol{\theta} - 2\mathbf{b}^\top A\boldsymbol{\theta} + |\mathbf{b}|^2 \\ &= \boldsymbol{\theta}^\top A^\top A\boldsymbol{\theta} - 2\mathbf{b}^\top A\boldsymbol{\theta} + |\mathbf{b}|^2. \end{aligned}$$

Because $|A\boldsymbol{\theta} - \mathbf{b}|^2$ is convex with respect to $\boldsymbol{\theta}$ due to Theorem A.1, the minimum is attained at the point where the gradient is zero. We thus compute the gradient with respect to $\boldsymbol{\theta}$, following Fact 4.19, and obtain

$$2A^\top A\boldsymbol{\theta} - 2A^\top \mathbf{b} = \mathbf{0} \quad \text{or} \quad A^\top A\boldsymbol{\theta} = A^\top \mathbf{b},$$

to which we refer as a normal equation.

Often, the number m of data points is significantly greater than the number n of parameters ($m \gg n$), and thereby the rank of A is n. Even if the rank of A is less than n, there is no issue in assuming that the rank is n, since we can always reduce the number of parameters by exploiting the linear dependence until the columns are linearly independent. If the rank of A is n, the rank of $A^\top A$ is also n due to Fact 3.8, and therefore the normal equation admits the following solution:

$$\hat{\boldsymbol{\theta}} = (A^\top A)^{-1} A^\top \mathbf{b}.$$

Fact 4.19 *Let Q be an $n \times n$ matrix, $\mathbf{b}$ an n-vector, and c a real number. A real-valued quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as*

$$f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c\,.$$

[8] We introduce a new notation argmin here. For a given function f, $\operatorname{argmin}_{\mathbf{x}\in A} f(\mathbf{x})$ refers to an element in A that minimizes f. If it is clear from the context, such as when $A = \mathbb{R}^n$, we often omit $\in A$. argmax is defined similarly.

Then, the gradient of f is given as

$$\nabla f(\mathbf{x}) = \big(Q + Q^\top\big)\mathbf{x} + \mathbf{b}\,.$$

If the matrix Q is symmetric, then $\nabla f(\mathbf{x}) = 2Q\mathbf{x} + \mathbf{b}$.

Proof: Since $f(\mathbf{x})$ can be written as

$$\begin{aligned} f(x_1, \ldots, x_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}x_i x_j + \sum_{i=1}^{n} b_i x_i + c \\ &= q_{kk}x_k^2 + \sum_{i\neq k}\sum_{j=1}^{n} q_{ij}x_i x_j + \sum_{j\neq k} q_{kj}x_k x_j + \sum_{i=1}^{n} b_i x_i + c\,, \end{aligned}$$

its partial derivative with respect to x_k is

$$\begin{aligned} \frac{\partial f}{\partial x_k}(x_1, \ldots, x_n) &= 2q_{kk}x_k + \sum_{i\neq k} q_{ik}x_i + \sum_{j\neq k} q_{kj}x_j + b_k \\ &= \sum_{i=1}^{n} q_{ik}x_i + \sum_{j=1}^{n} q_{kj}x_j + b_k = (Q^\top\mathbf{x})_k + (Q\mathbf{x})_k + b_k\,, \end{aligned}$$

which shows the desired representation. For a symmetric Q, the conclusion is straightforward. ■

4.9.2 Equivalence between Least Square and Projection

Because $\hat{\boldsymbol{\theta}}$ was the solution to the least-square problem in (4.18), we know that

$$A\hat{\boldsymbol{\theta}} = A(A^\top A)^{-1}A^\top\mathbf{b}$$

is the nearest vector to $\mathbf{b}$ among all the vectors in the column space of A, $\mathrm{Col}\,(A)$. Combining it with our earlier observation that that the projection of a vector onto a subspace is the nearest vector in the subspace to the original vector, we can reasonably guess that $A\hat{\theta}$ must be an orthogonal projection of $\mathbf{b}$ onto $\mathrm{Col}\,(A)$.

We confirm this guess of ours by checking whether the residual vector $\mathbf{b} - A\hat{\boldsymbol{\theta}}$ is orthogonal to $\mathrm{Col}\,(A)$. According to Lemma 4.4, we can alternatively check whether $\mathbf{b} - A\hat{\boldsymbol{\theta}} \in \mathrm{Null}\,(A^\top)$ holds, as $\mathrm{Col}\,(A)^\perp = \mathrm{Null}\,(A^\top)$. By multiplying $\mathbf{b} - A\hat{\boldsymbol{\theta}}$ with $A^\top$ from left,

$$A^\top(\mathbf{b} - A\hat{\boldsymbol{\theta}}) = A^\top(I - A(A^\top A)^{-1}A^\top)\mathbf{b} = (A^\top - A^\top A(A^\top A)^{-1}A^\top)\mathbf{b} = \mathbf{0}.$$

This residual vector is indeed orthogonal to the column space. In summary, the closest vector in $\mathrm{Col}\,(A)$ to an arbitrary vector is its own projection onto

$\mathrm{Col}\,(A)$, and we can find this nearest vector by multiplying the given vector with the following matrix

$$A(A^\top A)^{-1}A^\top.$$

Since this matrix is the projection matrix (4.11), we conclude that the solutions to least square problems and projections are equivalent.

Chapter 5

Singular Value Decomposition (SVD)

We often want to represent a bunch of high-dimensional vectors in a lower-dimensional vector space, for instance for the purpose of visualization. We can approach this problem as orthogonal projection, and a key question is which subspace we want to project these high-dimensional vectors, often data points. It turned out that the answer to this question depends on how we measure the error arising from this subspace projection. A natural choice of an error measure is a sum of squared norms of residual vectors connecting the original high-dimensional vectors and their projections on the subspace. With this choice, we can formulate the problem of finding the optimal subspace as that of searching for an orthonormal basis of the subspace onto which orthogonal projection minimizes this error measure. We will show that we can identify this optimal subspace by sequentially solving one-dimensional error minimization to find one orthonormal basic vector at a time, while satisfying increasingly more orthonormality constraints. We call these orthonormal basic vectors right-singular vectors, and the square root of the sum of the squared lengths of projected vectors from each one-dimensional error minimization problem a singular value. Together with left-singular vectors, we arrive at singular value decomposition (SVD). The left singular vectors can be treated as the optimal low-dimensional representations of the original higher-dimensional (data) vectors, assuming that we performed SVD on a data matrix consisting of data rows.

Due to our design of incremental optimizations searching for right-singular vectors, the singular value obtained at an earlier stage is bigger than the one at a later stage. The sum of the rank-one matrices built by the initially obtained k singular values and singular vectors turns out to be the best rank-k approximation of the data matrix. Furthermore, we can approximately invert any given data matrix, regardless of its invertibility, using singular vectors, which leads us to define pseudoinverse. This pseudoinverse projects data vectors onto a subspace spanned by linearly dependent vectors, corresponding to the least square solution, when the rank of the data matrix is not full. By translating this optimization formulation into the language of statistics, we arrive at principal components analysis (PCA) with the right-singular vectors correspond to the principal components in PCA.

5.1 A Variational Formulation for the Best-fit Subspaces

Let A be an $n \times d$ matrix whose rows correspond to d-dimensional (data) feature vectors. If $\mathbf{a}_i \in \mathbb{R}^d$ were the i-th row of A, A encodes the same information as $\{\mathbf{a}_i : i = 1, \ldots, n\}$. Here, we consider a problem of finding a k-dimensional subspace $\mathbb{W} \subset \mathbb{R}^d$ that represents the data set $\{\mathbf{a}_i : i = 1, \ldots, n\}$ well, assuming $k < d$. Among many possible criteria to measure how well such a subspace represents data, we use the square of a vector norm for singular vector decomposition (SVD). That is, we find the optimal k-dimensional subspace $\mathbb{W}^\star$ to minimize the "sum of squared residuals", as in

$$\mathbb{W}^\star = \underset{\mathbb{W}:\dim(\mathbb{W})\leq k}{\operatorname{argmin}} \sum_{i=1}^{n} \left|\mathbf{a}_i - \mathbf{P}_{\mathbb{W}}(\mathbf{a}_i)\right|^2. \tag{5.1}$$

Throughout this chapter, we use the standard inner product in $\mathbb{R}^d$. Up until now, we were mainly concerned about how to compute the projection, but here we are more concerned about how to determine the direction of projection.

Before going further, there are two questions that arise naturally here:

1. Why do we use the sum of squares? From the perspective of machine learning, the sum of squares is a special case of a decomposable loss function, $\sum_{i=1}^{n} \ell\big(\mathbf{a}_i, \mathbf{P}_{\mathbb{W}}(\mathbf{a}_i)\big)$. With the sum of squares, we can replace the sum of squared residuals with the sum of squared projection lengths using

the Pythagorean theorem, because

$$\left|\mathbf{a}_i\right|^2 = \left|\mathbf{P}_{\mathbb{W}}(\mathbf{a}_i)\right|^2 + \left|\mathbf{a}_i - \mathbf{P}_{\mathbb{W}}(\mathbf{a}_i)\right|^2$$

and $\sum_{i=1}^n |\mathbf{a}_i|^2$ is constant, we can rewrite (5.1) as

$$\mathbb{W}^\star = \operatorname*{argmax}_{\mathbb{W}:\dim(\mathbb{W})\le k} \sum_{i=1}^n \left|\mathbf{P}_{\mathbb{W}}(\mathbf{a}_i)\right|^2. \tag{5.2}$$

We use this form later when we derive SVD. Of course, this does not prevent us from using another loss function, but it must be determined for each loss function whether we can derive a simple solution.

2. How do we express a k-dimensional subspace? Although we can use k arbitrary, but linearly independent vectors to do so, we prefer to use k orthonormal vectors, as they collectively form a compact representation of orthogonal projection onto the k-dimensional subspace.

5.1.1 Best-fit 1-dimensional subspace

We start with the 1-dimensional subspace that best represents the data set $\{\mathbf{a}_i : 1 \le i \le n\}$. We constrain a basic vector to be a unit vector and use the inner product to measure the norm of the projection of a data vector onto the basic vector. If we let $\mathbf{v}$ be the basic vector of such a subspace, the projection of $\mathbf{a}_i$ onto this subspace is

$$\mathbf{P}_{\mathrm{span}\{\mathbf{v}\}}(\mathbf{a}_i) = \langle \mathbf{a}_i, \mathbf{v}\rangle \mathbf{v},$$

according to (4.5). The lengths of the data vectors after projection are then $\{|\langle \mathbf{a}_i, \mathbf{v}\rangle| : 1 \le i \le n\}$, and we can measure how well the overall data set is represented by

$$\sum_{i=1}^n \left|\mathbf{P}_{\mathrm{span}\{\mathbf{v}\}}(\mathbf{a}_i)\right|^2 = \sum_{i=1}^n \langle \mathbf{a}_i, \mathbf{v}\rangle^2 = \sum_{i=1}^n \left(\mathbf{a}_i{}^\top \mathbf{v}\right)^2.$$

Because $\mathbf{a}_i$ is the i-th row of A, this quantity is equivalent to $|A\mathbf{v}|^2$. We can thus find the best basic vector $\mathbf{v}_1$ to approximate the $n \times d$ matrix A by

$$\mathbf{v}_1 = \operatorname*{argmax}_{\mathbf{v}\in\mathbb{R}^d:|\mathbf{v}|=1} |A\mathbf{v}| = \operatorname*{argmax}_{\mathbf{v}\in\mathbb{R}^d:|\mathbf{v}|=1} \sum_{i=1}^n \langle \mathbf{a}_i, \mathbf{v}\rangle^2. \tag{5.3}$$

From Lemma C.3, we know that a solution exists to this problem, but there may not be a unique solution, as any unit vector would be a solution if A were for instance an identity matrix.

A Vector Orthogonal to a Set of Vectors

When looking for an optimal subspace of dimension greater than two, we must solve the problem of finding a vector, in a $(k+1)$-dimensional subspace, that is orthogonal to given k vectors, of course with $k < d$.

Lemma 5.1 *If $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbb{R}^d$ are orthonormal and $\mathbb{W}$ is a $(k+1)$-dimensional subspace, there exists at least one non-zero vector $\mathbf{w} \in \mathbb{W}$ that satisfies $\langle \mathbf{v}_1, \mathbf{w} \rangle = \cdots = \langle \mathbf{v}_k, \mathbf{w} \rangle = 0$. In other words, there exists a non-zero vector in $\mathbb{W}$ that is orthogonal to all the vectors in $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.*

Proof: Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_{k+1}\}$ be a basis of $\mathbb{W}$. Then, we can map a vector $\mathbf{w} \in \mathbb{W}$ to $\mathbf{x} = (x_1, \ldots, x_{k+1}) \in \mathbb{R}^{k+1}$ through $\mathbf{w} = x_1\mathbf{w}_1 + \cdots + x_{k+1}\mathbf{w}_{k+1}$. Considering this correspondence, a non-zero solution $\mathbf{x}$ to the following linear system corresponds to the desired vector $\mathbf{w}$:

$$\langle \mathbf{v}_i, \mathbf{w} \rangle = x_1 \langle \mathbf{v}_i, \mathbf{w}_1 \rangle + \cdots + x_{k+1} \langle \mathbf{v}_i, \mathbf{w}_{k+1} \rangle = 0, \quad i = 1, \ldots, k.$$

In other words, it is equivalent to finding a non-zero solution to $B\mathbf{x} = \mathbf{0}$, where

$$B = \big(\langle \mathbf{v}_i, \mathbf{w}_j \rangle\big)_{1 \le i \le k,\, 1 \le j \le k+1}.$$

Because there is one more variable than the number of equations, there must be a non-zero solution according to Lemma 3.2. ■

5.1.2 Best-fit 2-dimensional subspace

We can find the best-fit 2-dimensional subspace for the dataset $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ by solving the following optimization problem:

$$(\mathbf{w}_1^*, \mathbf{w}_2^*) = \underset{\substack{\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d: \\ |\mathbf{w}_1| = |\mathbf{w}_2| = 1, \\ \langle \mathbf{w}_1, \mathbf{w}_2 \rangle = 0}}{\operatorname{argmax}} \sum_{i=1}^{n} \left| \mathbf{P}_{\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2\}}(\mathbf{a}_i) \right|^2. \tag{5.4}$$

Solving this problem does not however shed light on the relationship between $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$. We instead plug in the best-fit one-dimensional subspace $\mathbf{v}_1$ of (5.3) in place of $\mathbf{w}_1$ and try to solve the problem (5.4) as another one-dimensional subspace problem. That is, we sequentially find $\mathbf{v}_1$ and $\mathbf{v}_2$ as follows:

1. $\mathbf{v}_1 = \underset{\mathbf{v} \in \mathbb{R}^d : |\mathbf{v}| = 1}{\operatorname{argmax}} |A\mathbf{v}|$;

2. $\mathbf{v}_2 = \underset{\substack{\mathbf{v}\in\mathbb{R}^d:|\mathbf{v}|=1,\\ \langle\mathbf{v},\mathbf{v}_1\rangle=0}}{\operatorname{argmax}} |A\mathbf{v}|.$

By Lemma C.4, we know the existence of a solution to one-dimensional problem 2 above. In the case of two dimensions, this sequential approach corresponds to finding the one-dimensional best-fit vector first and then subsequently finding the next one-dimensional best-fit vector orthogonal to the first best-fit vector.

We now check whether $\{\mathbf{v}_1, \mathbf{v}_2\}$, from this procedure, is as expressive as $\{\mathbf{w}_1^*, \mathbf{w}_2^*\}$, that is, whether

$$\sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{v}_1,\mathbf{v}_2\}}(\mathbf{a}_i)\right|^2 \geq \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{w}_1^*,\mathbf{w}_2^*\}}(\mathbf{a}_i)\right|^2$$

holds. First, we see that the original problem can be written down as

$$(\mathbf{w}_1^*, \mathbf{w}_2^*) = \underset{\substack{\mathbf{w}_1,\mathbf{w}_2\in\mathbb{R}^d:\\ |\mathbf{w}_1|=|\mathbf{w}_2|=1,\\ \langle\mathbf{w}_1,\mathbf{w}_2\rangle=0}}{\operatorname{argmax}} |A\mathbf{w}_1|^2 + |A\mathbf{w}_2|^2,$$

because

$$\begin{aligned}
\sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{w}_1,\mathbf{w}_2\}}(\mathbf{a}_i)\right|^2 &= \sum_{i=1}^{n} \left|\langle\mathbf{a}_i, \mathbf{w}_1\rangle\mathbf{w}_1 + \langle\mathbf{a}_i, \mathbf{w}_2\rangle\mathbf{w}_2\right|^2 \\
&= \sum_{i=1}^{n} \langle\mathbf{a}_i, \mathbf{w}_1\rangle^2 + \langle\mathbf{a}_i, \mathbf{w}_2\rangle^2 \qquad \text{since } \langle\mathbf{w}_1, \mathbf{w}_2\rangle = 0 \\
&= \left|A\mathbf{w}_1\right|^2 + \left|A\mathbf{w}_2\right|^2,
\end{aligned}$$

due to (4.7). Therefore, it is enough to show $|A\mathbf{w}_1^*|^2 + |A\mathbf{w}_2^*|^2 \leq |A\mathbf{v}_1|^2 + |A\mathbf{v}_2|^2$ if we let $\mathbb{W}^\star = \text{span}\{\mathbf{w}_1^*, \mathbf{w}_2^*\}$ be the optimal subspace found by solving the problem (5.4).

For the solution $\mathbf{v}_1$ of the first one-dimensional problem, according to Lemma 5.1, there exists a non-zero vector $\mathbf{w}$ in the 2-dimensional subspace $\mathbb{W}^\star$ such that $\langle\mathbf{v}_1, \mathbf{w}\rangle = 0$, to which we refer as $\hat{\mathbf{w}}_2$ after normalization. Once we get another unit vector $\hat{\mathbf{w}}_1 \in \mathbb{W}^\star$ orthogonal to $\hat{\mathbf{w}}_2$ using the Gram-Schmidt procedure, $\{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2\}$ is another basis of $\mathbb{W}^\star$. Because $\hat{\mathbf{w}}_2$ satisfies $\langle\mathbf{v}_1, \hat{\mathbf{w}}_2\rangle = 0$, it is a feasible solution to the second one-dimensional problem 2 and it must be that $|A\hat{\mathbf{w}}_2| \leq |A\mathbf{v}_2|$ since $\mathbf{v}_2$ is optimal to the problem 2. Furthermore, because $\hat{\mathbf{w}}_1$ satisfies the unit vector constraint of the first one-dimensional problem, it holds that $|A\hat{\mathbf{w}}_1| \leq |A\mathbf{v}_1|$. Since both $|A\mathbf{w}_1^*|^2 + |A\mathbf{w}_2^*|^2$ and $|A\hat{\mathbf{w}}_1|^2 + |A\hat{\mathbf{w}}_2|^2$

are squared sums of projections onto $\mathbb{W}^*$, they coincide to each other, that is, $|A\mathbf{w}_1^*|^2 + |A\mathbf{w}_2^*|^2 = |A\hat{\mathbf{w}}_1|^2 + |A\hat{\mathbf{w}}_2|^2$. Therefore, the sequential approach results in the optimal 2-dimensional subspace.

5.1.3 Best-fit k-dimensional subspace

Similarly to the earlier 2-dimensional case, we find the optimal k-dimensional subspace, $\mathbb{W}^* = \text{span}\{\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*\}$, for the data set $\{\mathbf{a}_i : 1 \le i \le n\}$ by solving

$$
\begin{aligned}
(\mathbf{w}_1^*, \ldots, \mathbf{w}_k^*) \quad = \quad & \underset{\mathbf{w}_1,\ldots,\mathbf{w}_k \in \mathbb{R}^d}{\text{argmax}} \quad \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{w}_1,\ldots,\mathbf{w}_k\}}(\mathbf{a}_i)\right|^2 \\
& \text{s.t.} \qquad |\mathbf{w}_i| = 1 \qquad \text{for } i = 1, \ldots, k \\
& \qquad\quad \langle \mathbf{w}_i, \mathbf{w}_j \rangle = 0 \quad \text{for } i \neq j
\end{aligned}
$$

Let us now consider a sequential approach to finding this subspace.

Sequential approach to find the best-fit k-dimensional subspace of an $n \times d$ matrix A

1. Set the first vector (by breaking ties arbitrarily) as

$$\mathbf{v}_1 \quad = \quad \underset{\mathbf{v} \in \mathbb{R}^d : |\mathbf{v}| = 1}{\text{argmax}} \quad |A\mathbf{v}| \; ;$$

2. For $j = 2, \ldots, k$: $\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$ is already known and set

$$\mathbf{v}_j \quad = \quad \underset{\substack{\mathbf{v} \in \mathbb{R}^d : |\mathbf{v}| = 1, \\ \langle \mathbf{v}, \mathbf{v}_1 \rangle = 0, \ldots, \langle \mathbf{v}, \mathbf{v}_{j-1} \rangle = 0}}{\text{argmax}} \quad |A\mathbf{v}| \, . \tag{5.5}$$

We use Lemma C.4 to show the existence of a solution to (5.5). Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis discovered by the sequential approach. Assume the sequential approach has successfully found the optimal $(k-1)$-dimensional subspace so far. Thanks to Lemma 5.1, there exists a unit vector $\hat{\mathbf{w}}_k$, in the k-dimensional $\mathbb{W}^*$, that is orthogonal to every vector in $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$, which allows us to find a basis of $\mathbb{W}^*$, $\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{k-1}, \hat{\mathbf{w}}_k\}$, that includes $\hat{\mathbf{w}}_k$. Because $\langle \mathbf{v}_1, \hat{\mathbf{w}}_k \rangle = 0, \ldots, \langle \mathbf{v}_{k-1}, \hat{\mathbf{w}}_k \rangle = 0$, $\hat{\mathbf{w}}_k$ is a feasible solution to the k-th optimization problem and hence $|A\hat{\mathbf{w}}_k| \le |A\mathbf{v}_k|$ holds. As an induction hypothesis, $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$

is an optimal $(k-1)$-dimensional subspace. Therefore, we get

$$\begin{aligned} |A\hat{\mathbf{w}}_1|^2 + \cdots + |A\hat{\mathbf{w}}_{k-1}|^2 &= \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\hat{\mathbf{w}}_1,\ldots,\hat{\mathbf{w}}_{k-1}\}}(\mathbf{a}_i)\right|^2 \\ \leq \quad \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}\}}(\mathbf{a}_i)\right|^2 &= |A\mathbf{v}_1|^2 + \cdots + |A\mathbf{v}_{k-1}|^2. \end{aligned}$$

By combining two inequalities, we conclude

$$\begin{aligned} \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\hat{\mathbf{w}}_1,\ldots,\hat{\mathbf{w}}_k\}}(\mathbf{a}_i)\right|^2 &= |A\hat{\mathbf{w}}_1|^2 + \cdots + |A\hat{\mathbf{w}}_{k-1}|^2 + |A\hat{\mathbf{w}}_k|^2 \\ &\leq |A\mathbf{v}_1|^2 + \cdots + |A\mathbf{v}_{k-1}|^2 + |A\mathbf{v}_k|^2 \end{aligned}$$

which implies, since $\mathbb{W}^* = \text{span}\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_k\}$,

$$\sum_{i=1}^{n} \left|\mathbf{P}_{\mathbb{W}^*}(\mathbf{a}_i)\right|^2 \leq |A\mathbf{v}_1|^2 + \cdots + |A\mathbf{v}_k|^2.$$

Therefore we can use the sequential approach above to find the best-fit k-dimensional subspace.

We call the unit vector $\mathbf{v}_i$, obtained by solving the i-th one-dimensional problem, the i-th right-singular vector and define the i-th singular value by

$$\sigma_i = |A\mathbf{v}_i|.$$

For right-singular vectors, we have a freedom of choosing the sign of each $\mathbf{v}_i$ since $-\mathbf{v}_i$ is also an optimal direction once $\mathbf{v}_i$ is optimal to the i-th optimization problem. Based on the derivation of the sequential approach above, we see that $\sigma_i \geq \sigma_{i+1}$ for all i and that this procedure of iteratively finding singular vectors terminates when $\sigma_{r+1} = 0$. That is,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1}.$$

We further define the i-th left-singular vector, $\mathbf{u}_i$, as

$$\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i,$$

which implies that

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \,.$$

In summary, we end up with the following theorem stating that the sequential approach above finds the optimal subspace.

Theorem 5.1 *Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be the right-singular vectors of an $n \times d$ matrix A. Then,* span$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$*, with $1 \leq k \leq r$, is the best-fit k-dimensional subspace of the rows of A (in terms of the sum of squared residuals.)*

If $\sigma_r > \sigma_{r+1} = 0$ for $r < d$, $|A\mathbf{v}_{r+1}| = 0$, meaning that every $\mathbf{a}_i$ is fully contained in a subspace spanned by $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. In other words, from $\mathbf{a}_i = \langle \mathbf{a}_i, \mathbf{v}_1 \rangle \mathbf{v}_1 + \cdots + \langle \mathbf{a}_i, \mathbf{v}_r \rangle \mathbf{v}_r$, $|\mathbf{a}_i|^2 = \langle \mathbf{a}_i, \mathbf{v}_1 \rangle^2 + \cdots + \langle \mathbf{a}_i, \mathbf{v}_r \rangle^2$ holds for any i, and thereby

$$\|A\|_F^2 = \sum_{i=1}^{n} |\mathbf{a}_i|^2 = |A\mathbf{v}_1|^2 + \cdots + |A\mathbf{v}_r|^2 = \sigma_1^2 + \cdots + \sigma_r^2. \tag{5.6}$$

What does it mean to explain the data set $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ using the best-fit k-dimensional subspace with k smaller than the minimal dimension r of a subspace that fully contain the data set? Although $|A\mathbf{v}_{k+1}| \neq 0$ in this case, we can perform a variety of analyses on the data set by treating each data point as a linear combination of k prototypes, as long as $|A\mathbf{v}_{k+1}|$ is small enough.

5.2 Orthogonality of left-singular Vectors

From the constraints of sequential one-dimensional optimizations, we impose the orthogonality on the right-singular vectors. A natural question is whether the left-singular vectors are also orthogonal to each other. The answer is yes even though it is not apparent since the left-singular vectors are defined through a matrix multiplication. Assume left-singular vectors, $\mathbf{u}_i$'s, are not orthogonal. Let $i < j$ be two indices of left-singular vectors not orthogonal to each other, that is, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle \neq 0$. By choosing the sign of $\mathbf{v}_j$ appropriately, we can further assume that $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta > 0$, without loss of generality. For a small positive constant $\epsilon > 0$, define a unit vector by

$$\mathbf{v}_i' = \frac{\mathbf{v}_i + \epsilon \mathbf{v}_j}{|\mathbf{v}_i + \epsilon \mathbf{v}_j|} = \frac{1}{\sqrt{1+\epsilon^2}}(\mathbf{v}_i + \epsilon \mathbf{v}_j).$$

By multiplying both sides with A, we get

$$A\mathbf{v}_i' = \frac{1}{\sqrt{1+\epsilon^2}}(\sigma_i \mathbf{u}_i + \epsilon \sigma_j \mathbf{u}_j).$$

The squared norm is then

$$|A\mathbf{v}_i'|^2 \quad = \quad \frac{1}{1+\epsilon^2}(\sigma_i^2 + 2\epsilon\delta\sigma_i\sigma_j + \epsilon^2\sigma_j^2)$$

$$\begin{aligned} &> (1-\epsilon^2)(\sigma_i^2 + 2\epsilon\delta\sigma_i\sigma_j + \epsilon^2\sigma_j^2) \\ &> (1-\epsilon^2)(\sigma_i^2 + 2\epsilon\delta\sigma_i\sigma_j) \\ &= \sigma_i^2 + \epsilon\big(-\epsilon\sigma_i^2 + 2(1-\epsilon^2)\delta\sigma_i\sigma_j\big) \\ &> \sigma_i^2, \end{aligned}$$

because $2(1-\epsilon^2)\delta\sigma_i\sigma_j > \epsilon\sigma_i^2$ for a sufficiently small ϵ. $\mathbf{v}_i$ and $\mathbf{v}_j$ are orthogonal to all $\mathbf{v}_\ell$ with $\ell < i$, and thus $\mathbf{v}_i'$ is feasible to the i-th optimization problem in the sequential approach. The objective value attained with $\mathbf{v}_i'$ is however greater than the optimal objective value $\sigma_i = |A\mathbf{v}_i|$, which is contradictory. Therefore, all $\mathbf{u}_i$'s are orthogonal.

Lemma 5.2 *The left-singular vectors $\mathbf{u}_i$'s of a matrix A defined as*

$$\mathbf{u}_i = \frac{1}{|A\mathbf{v}_i|} A\mathbf{v}_i,$$

for each right-singular vector $\mathbf{v}_i$, are orthonormal to each other.

5.3 Representing SVD in Various Forms

Let $(\sigma_1, \mathbf{v}_1, \mathbf{u}_1)$ be a singular triplet satisfying $A\mathbf{v}_1 = \sigma_1\mathbf{u}_1$. Can we come up with a simple matrix A_1 that represents the relationship encoded in this singular triplet? One way to measure the simplicity of a matrix is its rank. It turned out we can define a rank-one matrix A_1 from this singular triplet as

$$A_1 = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top,$$

which satisfies $A_1\mathbf{v}_1 = \sigma_1\mathbf{u}_1$. Let $(\sigma_2, \mathbf{v}_2, \mathbf{u}_2)$ be another singular triplet of the matrix A, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthonormal, and similarly $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthonormal. We build a corresponding rank-one matrix $\sigma_2\mathbf{u}_2\mathbf{v}_2^\top$. Let us add these two rank-one matrices to get

$$A_2 = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top + \sigma_2\mathbf{u}_2\mathbf{v}_2^\top.$$

From the orthonormality of right-singular vectors, we can easily see that A_2 satisfies both $A_2\mathbf{v}_1 = \sigma_1\mathbf{u}_1$ and $A_2\mathbf{v}_2 = \sigma_2\mathbf{u}_2$, and that A_2 is a rank-two matrix from Fact 5.4. From this observation, we can intuitively guess that we can obtain

$$A = \sum_{i=1}^{\operatorname{rank} A} \sigma_i\mathbf{u}_i\mathbf{v}_i^\top$$

using as many singular triplets as rank A.

Let $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = 0$ for an $n \times d$ matrix A. For this matrix, we have two very useful representations, sum of rank-one matrices and product of matrices with orthonormal columns and a diagonal matrix.

- $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$: Consider a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_d\}$ of $\mathbb{R}^d$ that includes $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. We may assume that the basis is orthonormal thanks to the Gram-Schmidt procedure. We can then represent an arbitrary vector in $\mathbb{R}^d$ as $\mathbf{x} = x_1\mathbf{v}_1 + \cdots + x_d\mathbf{v}_d$. Because $\sigma_{r+1} = 0$, $A\mathbf{v}_j = \mathbf{0}$ for $j = r+1, \ldots, d$. Furthermore, because $x_i = \mathbf{v}_i^\top \mathbf{x}$ for all i, we see that

$$\begin{aligned} A\mathbf{x} &= x_1 A\mathbf{v}_1 + \cdots + x_r A\mathbf{v}_r \\ &= \sum_{i=1}^{r} x_i(\sigma_i \mathbf{u}_i) = \sum_{i=1}^{r} (\sigma_i \mathbf{u}_i) x_i = \sum_{i=1}^{r} (\sigma_i \mathbf{u}_i)(\mathbf{v}_i^\top \mathbf{x}) = \Big(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \Big) \mathbf{x}, \end{aligned}$$

  for all $\mathbf{x} \in \mathbb{R}^d$. This allows us to express A as $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

- $A = UDV^\top$ or $AV = UD$: Here, $U^\top U = I_r = V^\top V$, and D is a positive diagonal matrix.

$$\underset{n \times d}{A} = \underset{n \times r}{U} \underset{r \times r}{D} \underset{r \times d}{V^\top}.$$

  We see this by applying Corollary 3.1 to U D, and V where the columns of U and V are left and right-singular vectors, respectively, and the diagonals of D are singular values. According to (3.8), then, $UDV^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Therefore, $A = UDV^\top$.

For $A^\top$, $A^\top \mathbf{u}_i = \big(\sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top \big)^\top \mathbf{u}_i = \big(\sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{u}_j^\top \big) \mathbf{u}_i = \sigma_i \mathbf{v}_i$. We derive and summarize the analogous properties of $A^\top$ below.

- $A^\top = \sum_{i=1}^r \sigma_i \mathbf{v}_i \mathbf{u}_i^\top$
- $A^\top = VDU^\top$ with $U^\top U = I_r$, $V^\top V = I_r$ and an $r \times r$ square matrix D with positive diagonals
- $A^\top \mathbf{u}_i = \sigma_i \mathbf{v}_i$

From the first one-dimensional optimization problem to compute the first singular value, we observe the following result.

Fact 5.1 *$\|A\|_2 = \sigma_1$ for any $n \times d$ matrix A.*

Proof: Because the optimization problem for finding the first singular value in (5.3) and the optimization problem for computing the spectral norm in (4.16) are equivalent, we get $\sigma_1 = |A\mathbf{v}_1| = \|A\|_2$. ■

5.4 Properties of Sum of Rank-one Matrices

Consider a matrix defined as

$$\sum_{i=1}^k \alpha_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{5.7}$$

where $\alpha_1 \geq \cdots \geq \alpha_k > 0$, and $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\} \subset \mathbb{R}^n$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset \mathbb{R}^d$ are orthonormal vectors, respectively. This $n \times d$ matrix may or may not have been constructed from SVD, and we still can derive the following results.

Fact 5.2 *$(\alpha_i, \mathbf{v}_i, \mathbf{u}_i)$ is the i-th singular triplet of $\sum_{i=1}^k \alpha_i \mathbf{u}_i \mathbf{v}_i^\top$.*

Proof: Let $A = \sum_{i=1}^k \alpha_i \mathbf{u}_i \mathbf{v}_i^\top$. We then extend $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ orthogonally to construct an orthonormal basis of $\mathbb{R}^d$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ via the Gram-Schmidt procedure. If we write a unit vector $\mathbf{v}$ in $\mathbb{R}^d$ as $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_d\mathbf{v}_d$, we can further write $A\mathbf{v}$ as

$$A\mathbf{v} = \Big(\sum_{i=1}^k \alpha_i \mathbf{u}_i \mathbf{v}_i^\top\Big)\mathbf{v} = \sum_{i=1}^k \alpha_i \mathbf{u}_i \mathbf{v}_i^\top \mathbf{v} = \sum_{i=1}^k \alpha_i x_i \mathbf{u}_i\,, \quad |A\mathbf{v}|^2 = \sum_{i=1}^k \alpha_i^2 x_i^2.$$

Under the unit-norm constraint ($|\mathbf{v}| = 1$), which translates to $x_1^2 + \cdots + x_d^2 = 1$, then $x_1 = 1, x_2 = \cdots = x_d = 0$ maximizes $|A\mathbf{v}|$. That is, $\mathbf{v} = \mathbf{v}_1$. The first right-singular vector of A is $\mathbf{v}_1$. Since $A\mathbf{v}_1 = \alpha_1\mathbf{u}_1$ and $|A\mathbf{v}_1| = \alpha_1$, α_1 is the first singular value, and $\mathbf{u}_1$ is the first left-singular vector. In order to get the second

singular vector, we add the extra constraint that $\langle \mathbf{v}, \mathbf{v}_1 \rangle = 0$ which corresponds to considering only $\mathbf{v}$ that can be expressed as $\mathbf{v} = x_2\mathbf{v}_2 + \cdots + x_d\mathbf{v}_d$. Then, α_2, $\mathbf{v}_2$ and $\mathbf{u}_2$ are the second singular value, left-singular vector, and right-singular vector, respectively. We can recover all the remaining singular triplets similarly. ■

When we have a singular triplet of A in hand, does it help us find SVD of $A^\top$? If we start by representing A as the sum of rank-one matrices, we can constructively answer this question. From

$$A^\top = \Big(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\Big)^\top = \sum_{i=1}^{r} \sigma_i \big(\mathbf{u}_i \mathbf{v}_i^\top\big)^\top = \sum_{i=1}^{r} \sigma_i \mathbf{v}_i \mathbf{u}_i^\top ,$$

we see that the role of two vectors in each rank-one summand are swapped. Thanks to Fact 5.2, the right (left) singular vectors of A are the left (right) singular vectors of $A^\top$. And, A and $A^\top$ share the same singular values.

For the sum of rank-one matrices, we can easily read the matrix norms from the coefficients of rank-one terms which are singular values.

Fact 5.3 $\big\| \sum_{i=1}^{k} \alpha_i \mathbf{u}_i \mathbf{v}_i^\top \big\|_2 = \alpha_1$ *and* $\big\| \sum_{i=1}^{k} \alpha_i \mathbf{u}_i \mathbf{v}_i^\top \big\|_F = \sqrt{\alpha_1^2 + \cdots + \alpha_k^2}$.

Proof: They are derived from Fact 5.1, Fact 5.2, and (5.6). ■

Another important question is on the rank of a sum of rank-one matrices.

Fact 5.4 $\operatorname{rank}\big(\sum_{i=1}^{k} \alpha_i \mathbf{u}_i \mathbf{v}_i^\top\big) = k$.

Proof: Let us characterize the null space of $A = \sum_{i=1}^{k} \alpha_i \mathbf{u}_i \mathbf{v}_i^\top$. We do so by looking for a vector $\mathbf{v} \in \mathbb{R}^d$ that satisfies $A\mathbf{v} = \mathbf{0}$. Start from orthonormal vectors, $\mathbf{v}_1, \ldots, \mathbf{v}_k$, and extend them to construct a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_d\}$ of $\mathbb{R}^d$. When we write $\mathbf{v} \in \mathbb{R}^d$ as $\mathbf{v} = \sum_{i=1}^{d} x_i \mathbf{v}_i$ with $(x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, the necessary and sufficient condition for $A\mathbf{v} = \mathbf{0}$ is $x_1 = \cdots = x_k = 0$, since

$$A\mathbf{v} = \sum_{i=1}^{d} x_i A\mathbf{v}_i = \sum_{i=1}^{d} x_i \Big(\sum_{j=1}^{k} \alpha_i \mathbf{u}_j \mathbf{v}_j^\top\Big)\mathbf{v}_i = \sum_{j=1}^{k}\sum_{i=1}^{d} x_i \alpha_i \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i = \sum_{i=1}^{k} x_i \alpha_i \mathbf{u}_i .$$

In other words, $\operatorname{Null}(A) = \operatorname{span}\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_d\}$, and thereby, $\dim \operatorname{Null}(A) = d - k$. Then, according to the rank-nullity theorem (3.5),

$$\operatorname{rank} A = d - \dim \operatorname{Null}(A) = k.$$

■

To see the usefulness of these facts, let us consider the following matrix given as a sum of rank-one matrices.

Example 5.1 Let a 4×5 matrix A is given as

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

After normalization of vectors, we get

$$\begin{aligned} A &= \sqrt{2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \end{bmatrix} - \sqrt{6} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix} \\ &\quad - \sqrt{10} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \end{bmatrix}. \end{aligned}$$

We re-arrange the terms with modification the leading signs and get

$$\begin{aligned} A &= \sqrt{10} \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \end{bmatrix} + \sqrt{6} \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 \end{bmatrix} \\ &\quad + \sqrt{2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \end{bmatrix} \\ &= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^\top \end{aligned}$$

Observe that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ are orthonormal. Fact 5.2 tells us that $(\sigma_i, \mathbf{v}_i, \mathbf{u}_i)$ is the i-th singular triplet of matrix A, and thereby, $\|A\|_2 = \sqrt{10}$ and $\|A\|_F = 3\sqrt{2}$ according to Fact 5.3. We also read out of the rank of A, $\operatorname{rank} A = 3$, using Fact 5.4. ∎

With the result on the rank of the sum of rank-one matrices, we now state the singular value decomposition.

Singular Value Decomposition (SVD) Any $n \times d$ matrix A (with $r = \text{rank}\, A$) can be represented as

$$A = UDV^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{5.8}$$

where $n \times r$ matrix U and $d \times r$ matrix V are matrices with orthonormal columns, respectively, and $D = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$. $\mathbf{u}_i$ and $\mathbf{v}_i$ are i-th column vectors of U and V, respectively. Observe that $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \subset \mathbb{R}^n$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\} \subset \mathbb{R}^d$ are orthonormal sets of vectors, respectively.

We can express the SVD representation (5.8) in many different ways. For any $r \leq k \leq n$ and $r \leq \ell \leq d$, we can extend the left-singular vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ to k orthonormal vectors of $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_k\}$ and right-singular vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ to ℓ orthonormal vectors of $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_\ell\}$, with the Gram-Schmidt procedure. Let U and V be $n \times k$ and $d \times \ell$ matrices whose columns are $\mathbf{u}_i$'s and $\mathbf{v}_i$'s, respectively. Further define a $k \times \ell$ matrix D whose r leading diagonal entries are the singular values and all others zero, i.e., $(D)_{ii} = \sigma_i$ for $i = 1, \ldots, r$ and $(D)_{ij} = 0$ for $i \neq j$. Then,

$$A = UDV^\top.$$

If $k = \ell$, we can represent this matrix product as

$$A = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

by setting $\sigma_{r+1} = \cdots = \sigma_k = 0$. Because of the minimality of representation, we sometimes call (5.8) compact singular value decomposition, or compact SVD.

Based on the definition of SVD, we can relate the spectral and Frobenius norms with each other.

Fact 5.5 *For any matrix A,* $\|A\|_2 \leq \|A\|_F$. *If* $\text{rank}(A) = r$, $\|A\|_F \leq \sqrt{r}\|A\|_2$. *For rank-one matrices, two norms coincide.*

Proof: Although we have already proven in Fact 4.17 that $\|A\|_2 \leq \|A\|_F$, we consider another proof based on SVD here. Let us express an $n \times d$ matrix A as $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ using SVD. Then, due to Fact 5.3,

$$\|A\|_2 = \sigma_1 \leq \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} = \|A\|_F.$$

Furthermore,

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} \le \sqrt{r\sigma_1^2} = \sqrt{r}\sigma_1 = \sqrt{r}\|A\|_2.$$

For the case of rank $A = 1$, $\|A\|_2 \le \|A\|_F \le \sqrt{1}\|A\|_2$ implies $\|A\|_2 = \|A\|_F$. ■

Once we obtain an SVD of an invertible matrix, we can write the inverse of the matrix using the singular triplets.

Fact 5.6 *Let A be an $n \times n$ invertible matrix. Assume that an SVD of A is given as*

$$U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top .$$

Then, all $\sigma_i > 0$, and

$$A^{-1} = V\Sigma^{-1}U^\top = \sum_{j=1}^{n} \sigma_j^{-1} \mathbf{v}_j \mathbf{u}_j^\top . \tag{5.9}$$

Proof: The invertibility of A implies rank $A = n$, and due to Fact 5.4, all σ_i's are positive. Then, Σ^{-1} and $\sum_{j=1}^{n} \sigma_j^{-1} \mathbf{v}_j \mathbf{u}_j^\top$ are well-defined. Furthermore, U and V are orthogonal matrices. So,

$$A^{-1} = (U\Sigma V^\top)^{-1} = (V^\top)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^\top = \sum_{j=1}^{n} \sigma_j^{-1} \mathbf{v}_j \mathbf{u}_j^\top .$$

■

We will introduce the notion of pseudoinverse for non-rectangular and/or non-invertible matrices in Section 5.8. The pseudo-inverse can be expressed similarly to (5.9).

With SVD, we can readily compute $\|A^{-1}\|_2\|A\|_2$ which is called the *condition number* of a matrix A and denoted by $\kappa(A)$. This is a very important concept in numerical linear algebra.

Example 5.2 Let A be an $n \times n$ invertible matrix. If an SVD of A is given as $\sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\sigma_1 \ge \cdots \ge \sigma_n > 0$. Let us find $\|A^{-1}\|_2\|A\|_2$. Facts 5.3 and 5.6 imply $\|A\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = \sigma_n^{-1}$. Therefore, $\|A^{-1}\|_2\|A\|_2 = \sigma_1\sigma_n^{-1}$. ▮

5.4.1 SVD Interpretation of Transformation

Let A be an $m \times n$ matrix representing a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$. To understand the action of the transformation T in geometric terms, we consider an SVD of A in a form of $U\Sigma V^\top$ where U and V are orthogonal matrices of appropriate sizes and Σ is an $m \times n$ diagonal matrix with potential zeros as diagonals. As we discussed in Section 4.7, well-known transformations described by an orthogonal matrix includes rotation, reflection, and their combination.

In other words, we can view any linear transformations T as a composition of three transformations; first, orthogonal transformation in $\mathbb{R}^n$ by an orthogonal matrix $V^\top$, then scaling, which is interpretable, by a diagonal matrix Σ, and finally another orthogonal transformation by an orthogonal matrix U in $\mathbb{R}^m$. An asymmetric diagonal matrix Σ not only scales some components of vectors, but may also reduce the dimension of vectors if $m < n$ or inflate the number of nominal dimensions if $m > n$.

5.5 Spectral Decomposition of Symmetric Matrix via SVD

Consider a symmetric, rank-r, $n \times n$ matrix A with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1}$. Because it is symmetric, both right and left-singular vectors are $\mathbb{R}^n$ vectors. Let $(\sigma_1, \mathbf{v}_1, \mathbf{u}_1), \ldots, (\sigma_{j-1}, \mathbf{v}_{j-1}, \mathbf{u}_{j-1})$ be the first $(j-1)$ singular triplets. Let us assume either $\mathbf{u}_i = \mathbf{v}_i$ or $\mathbf{u}_i = -\mathbf{v}_i$ holds for all of these known first $(j-1)$ singular triplets. If we denote by $(\sigma_j, \mathbf{v}, \mathbf{u})$ the j-th singular triplet obtained by solving the j-th optimization problem in the sequential SVD procedure (5.5), we know that

- $\mathbf{v} \perp \{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$ according to (5.5);
- $\mathbf{u} \perp \{\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}\}$ according to Lemma 5.2.

Because of the assumption that $\mathbf{u}_i = \pm\mathbf{v}_i$ for $i \leq j-1$, we obtain

$$(\mathbf{v} + \mathbf{u}) \perp \{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}. \tag{5.10}$$

Since A is symmetric, both $A\mathbf{v} = \sigma_j\mathbf{u}$ and $A^\top\mathbf{u} = A\mathbf{u} = \sigma_j\mathbf{v}$ hold, which implies

$$A(\mathbf{v} + \mathbf{u}) = \sigma_j(\mathbf{v} + \mathbf{u}).$$

One of the following two cases holds:

- $\mathbf{v}+\mathbf{u} \neq \mathbf{0}$: Set $\mathbf{v}_j = \frac{1}{|\mathbf{v}+\mathbf{u}|}(\mathbf{v}+\mathbf{u})$, and we get $A\mathbf{v}_j = \sigma_j \mathbf{v}_j$. $\mathbf{v}_j$ is a feasible solution to (5.5) thanks to (5.10). Furthermore, $|A\mathbf{v}_j| = \sigma_j$ implies that $\mathbf{v}_j$ is an optimal solution to (5.5). Hence, $\mathbf{v}_j$ is eligible for the j-th right-singular vector and also the j-th left-singular vector;

- $\mathbf{v}+\mathbf{u} = \mathbf{0}$: Set $\mathbf{v}_j = \mathbf{v}$ and $\mathbf{u}_j = \mathbf{u}$. Since $A\mathbf{v} = \sigma_j \mathbf{u}$, $(\mathbf{v}_j, \mathbf{u}_j)$ is j-th pair of singular vectors with the property $\mathbf{v}_j = -\mathbf{u}_j$.

That is, either way, we get a j-th singular triplet $(\sigma_j, \mathbf{v}_j, \mathbf{u}_j)$ with $\mathbf{u}_j = \pm\mathbf{v}_j$. By repeating this procedure for r times, we obtain r singular triplets where each pair of right and left-singular vectors are parallel to each other.

If we set λ_i as σ_i when $\mathbf{v}_i = \mathbf{u}_i$ or $-\sigma_i$ when $\mathbf{v}_i = -\mathbf{u}_i$, then $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ holds for all i, resulting in r scalar-vector pairs, $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_r, \mathbf{v}_r)$. This simple relationship for the pairs are named in the following definition.

Definition 5.1 *For a square matrix A, a scalar λ and a non-zero vector $\mathbf{v}$ are eigenvalue and eigenvector of A, respectively, if they satisfy*

$$A\mathbf{v} = \lambda\mathbf{v}. \tag{5.11}$$

The pair $(\lambda, \mathbf{v})$ is called an eigenpair.[a] *Similarly, $\mathbf{u}$ is called a left-eigenvector if it satisfies $\mathbf{u}^\top A = \lambda\mathbf{u}^\top$. By eigen-decomposition of A, we refer to finding the eigenpairs of A.*

[a] We study much more in-depth eigenvalues and eigenvectors in Chapter 9. Until then, all we need is the definition of eigenvalues and eigenvectors satisfying (5.11).

After obtaining the eigenpairs $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_r, \mathbf{v}_r)$ of symmetric A for the case where left and right-singular vectors are parallel, we can express the matrix A as the sum of rank-one matrices, as follows:

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Starting from these r eigenvectors, we can extend to n orthonormal vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ using the Gram-Schmidt procedure. By letting $\lambda_{r+1} = \cdots = \lambda_n = 0$, we end up with n eigenpairs, $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_n, \mathbf{v}_n)$. These eigenvalues are all real, as they are either singular values themselves or their negations. We call such decomposition of a matrix spectral decomposition.

Theorem 5.2 (Real Spectral Decomposition) *Let A be a real symmetric matrix of rank r. Then, A can be represented as*

$$A = V\Lambda V^\top = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top , \tag{5.12}$$

where V is an orthogonal matrix with orthonormal columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, $|\mathbf{v}_i| = 1$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. λ_i's and $\mathbf{v}_i$'s are the eigenvalues and eigenvectors of A, respectively. There are exactly r non-zero eigenvalues. Note that the summands of (5.12) can be rearranged in any order.

Real spectral decomposition is the most popular form of so-called eigendecomposition, which we will learn more later in Chapter 9. Later in Appendix F, we provide another proof of the real spectral decomposition that does not rely on SVD. We highly recommend you take a look at the alternative proof together. We demonstrate the consequences of this theorem using the following symmetric matrix given as the sum of rank-one matrices.

Example 5.3 Let a 4×4 symmetric matrix A be given as

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

1. Observe that $[0\ 0\ 1\ 0]^\top$, $[1\ 1\ 0\ 0]^\top$, $[1\ -1\ 0\ 0]^\top$ are mutually orthogonal. Let $\mathbf{v}_1 = [0\ 0\ 1\ 0]^\top$, $\mathbf{v}_2 = \frac{1}{\sqrt{2}}[1\ 1\ 0\ 0]^\top$, $\mathbf{v}_3 = \frac{1}{\sqrt{2}}[1\ -1\ 0\ 0]^\top$. Then $\mathbf{v}_i$'s are orthonormal and

   $$A = 2\mathbf{v}_1\mathbf{v}_1^\top - 6\mathbf{v}_2\mathbf{v}_2^\top + 4\mathbf{v}_3\mathbf{v}_3^\top .$$

   If we multiply A with $\mathbf{v}_1$ from right, we obtain $A\mathbf{v}_1 = 2\mathbf{v}_1$, because $\mathbf{v}_i$'s are orthonormal. Repeating it for all $\mathbf{v}_i$'s, we arrive at the eigenpairs, $(2, \mathbf{v}_1)$, $(-6, \mathbf{v}_2)$ and $(4, \mathbf{v}_3)$. Moreover, if we multiply A with a vector perpendicular to $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$, such as $\mathbf{v}_4 = [0\ 0\ 0\ 1]^\top$, we get $A\mathbf{v}_4 = \mathbf{0}$. In other words, $(0, \mathbf{v}_4)$ is also an eigenpair.

2. We can slightly modify decomposition in 1 by letting $\mathbf{u}_1 = \mathbf{v}_1$, $\mathbf{u}_2 = -\mathbf{v}_2$ and $\mathbf{u}_3 = \mathbf{v}_3$. Then

   $$A = 2\mathbf{u}_1\mathbf{v}_1^\top + 6\mathbf{u}_2\mathbf{v}_2^\top + 4\mathbf{u}_3\mathbf{v}_3^\top .$$

We effortlessly get a (compact) SVD from this decomposition:

$$A = UDV^\top, \quad U = [\mathbf{u}_2|\mathbf{u}_3|\mathbf{u}_1],\ V = [\mathbf{v}_2|\mathbf{v}_3|\mathbf{v}_1],\ D = \mathrm{diag}(6,4,2).$$

The singular triplets are then

$$(6, \mathbf{u}_2, \mathbf{v}_2),\ (4, \mathbf{u}_3, \mathbf{v}_3),\ (2, \mathbf{u}_1, \mathbf{v}_1).$$

$\mathbf{u}_i$'s are left-singular vectors, and $\mathbf{v}_i$'s are right-singular vectors.

∎

For a real symmetric matrix, we can derive another interesting representation in terms of projections thanks to the real spectral decomposition. Assume a real symmetric matrix A and its spectral decomposition as in (5.12). Suppose further that $\lambda_1 = \cdots = \lambda_k = \lambda$ and other eigenvalues are different from λ. Let $\mathbf{P}_{A,\lambda}$ be a projection transformation onto $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. That is,[1]

$$\mathbf{P}_{A,\lambda} = \sum_{j=1}^{k} \mathbf{v}_j \mathbf{v}_j^\top = \mathbf{P}_{\mathrm{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}}. \tag{5.13}$$

Observe that $\mathbf{P}_{A,\lambda} = \mathbf{P}_{A,\lambda_1} = \cdots = \mathbf{P}_{A,\lambda_k}$. If $\lambda_i \neq \lambda_j$, $\mathbf{P}_{A,\lambda_i}\mathbf{P}_{A,\lambda_j} = \mathbf{0}$ holds since the eigenvectors constituting $\mathbf{P}_{A,\lambda_i}$ and $\mathbf{P}_{A,\lambda_j}$ are non-overlapped. In addition, because $\mathbf{P}_{A,\lambda_i}$ is a projection, the following hold:

$$\mathbf{P}_{A,\lambda_i}{}^2 = \mathbf{P}_{A,\lambda_i} \quad \text{and} \quad \mathbf{P}_{A,\lambda_i}{}^\top = \mathbf{P}_{A,\lambda_i}\,.$$

When there are r distinct eigenvalues $\lambda_1, \ldots, \lambda_r$, we can re-write real spectral decomposition (5.12) succinctly as

$$A = \sum_{i=1}^{r} \lambda_i \mathbf{P}_{A,\lambda_i}. \tag{5.14}$$

This representation will help us prove the uniqueness of the square root of a positive definite matrix in Chapter 7.

Consider an example of such projection-based decomposition.

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

[1] We treat the linear operator and its corresponding matrix interchangeably here.

$$
\begin{aligned}
&= 2\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} + 2\begin{bmatrix}0&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} + 3\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix} + 3\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix} \\
&= 2\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}^\top + 2\begin{bmatrix}0\\1\\0\\0\end{bmatrix}\begin{bmatrix}0\\1\\0\\0\end{bmatrix}^\top + 3\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}^\top + 3\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\\1\end{bmatrix}^\top \\
&= 2\left(\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}^\top + \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\begin{bmatrix}0\\1\\0\\0\end{bmatrix}^\top\right) + 3\left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}^\top + \begin{bmatrix}0\\0\\0\\1\end{bmatrix}\begin{bmatrix}0\\0\\0\\1\end{bmatrix}^\top\right) \\
&= 2\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} + 3\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} \\
&= 2\mathbf{P}_{A,2} + 3\mathbf{P}_{A,3}.
\end{aligned}
$$

Even if we choose to use different orthonormal vectors in the third step above, we end up with the same projection-based expression:

$$
\begin{aligned}
A &= 2\begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\\0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\\0\end{bmatrix}^\top + 2\begin{bmatrix}-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\\0\end{bmatrix}\begin{bmatrix}-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\\0\end{bmatrix}^\top + 3\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}^\top + 3\begin{bmatrix}0\\0\\0\\-1\end{bmatrix}\begin{bmatrix}0\\0\\0\\-1\end{bmatrix}^\top \\
&= 2\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0&0\\\frac{1}{2}&\frac{1}{2}&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} + 2\begin{bmatrix}\frac{1}{2}&-\frac{1}{2}&0&0\\-\frac{1}{2}&\frac{1}{2}&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} + 3\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix} \\
&\quad +3\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= 2\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + 3\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= 2\,\mathbf{P}_{A,2} + 3\,\mathbf{P}_{A,3}.
\end{aligned}
$$

Based on many ideas we have explored so far, we can derive the following maximum principle of Ky Fan for the eigenvalues of a symmetric matrix:

Theorem 5.3 (Ky Fan) *Let A be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and the corresponding orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then*

$$\lambda_1 + \cdots + \lambda_k = \max_{\substack{X\,:\,n\times k \\ X^\top X = I_k}} \operatorname{trace}\left(X^\top A X\right).$$

An optimal X is given by $[\mathbf{v}_1|\cdots|\mathbf{v}_k]Q$ with Q an arbitrary orthogonal matrix.

Proof: For $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = \sum_{i=1}^n (\mathbf{x}^\top \mathbf{v}_i)\mathbf{v}_i$ with $|\mathbf{x}|^2 = \sum_{i=1}^n (\mathbf{x}^\top \mathbf{v}_i)^2$ and

$$A\mathbf{x} = \sum_{i=1}^n (\mathbf{x}^\top \mathbf{v}_i) A\mathbf{v}_i = \sum_{i=1}^n (\mathbf{x}^\top \mathbf{v}_i)\lambda_i \mathbf{v}_i\,.$$

Then the quadratic form $\mathbf{x}^\top A\mathbf{x}$ expands and is bounded above as

$$
\begin{aligned}
\mathbf{x}^\top A\mathbf{x} &= \sum_{i=1}^n \lambda_i (\mathbf{x}^\top \mathbf{v}_i)^2 \\
&= \sum_{i=1}^n (\lambda_k + \lambda_i - \lambda_k)(\mathbf{x}^\top \mathbf{v}_i)^2 \\
&= \lambda_k \sum_{i=1}^n (\mathbf{x}^\top \mathbf{v}_i)^2 + \sum_{i=1}^k (\lambda_i - \lambda_k)(\mathbf{x}^\top \mathbf{v}_i)^2 + \sum_{i=k+1}^n (\lambda_i - \lambda_k)(\mathbf{x}^\top \mathbf{v}_i)^2 \\
&\leq \lambda_k |\mathbf{x}|^2 + \sum_{i=1}^k (\lambda_i - \lambda_k)(\mathbf{x}^\top \mathbf{v}_i)^2.
\end{aligned}
$$

Let us plug in $\mathbf{x}_j$ with $|\mathbf{x}_j| = 1$ into $|\mathbf{x}|$ to obtain

$$\mathbf{x}_j^\top A\mathbf{x}_j \leq \lambda_k + \sum_{i=1}^k (\lambda_i - \lambda_k)(\mathbf{x}_j^\top \mathbf{v}_i)^2.$$

Then, by summing after some rearrangement, we get

$$\sum_{j=1}^k \left(\lambda_j - \mathbf{x}_j^\top A\mathbf{x}_j\right) \geq \sum_{j=1}^k (\lambda_j - \lambda_k) - \sum_{j=1}^k \sum_{i=1}^k (\lambda_i - \lambda_k)(\mathbf{x}_j^\top \mathbf{v}_i)^2$$

$$= \sum_{i=1}^{k}(\lambda_i - \lambda_k) - \sum_{i=1}^{k}(\lambda_i - \lambda_k)\sum_{j=1}^{k}(\mathbf{x}_j^\top \mathbf{v}_i)^2$$
$$= \sum_{i=1}^{k}(\lambda_i - \lambda_k)\Big(1 - \sum_{j=1}^{k}(\mathbf{x}_j^\top \mathbf{v}_i)^2\Big).$$

If, in addition, $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are orthogonal to each other, we also observe

$$1 = |\mathbf{v}_i|^2 \geq \sum_{j=1}^{k}(\mathbf{x}_j^\top \mathbf{v}_i)^2 \text{ for each } i$$

through the Gram-Schmidt procedure on $\mathbf{x}_j$'s and get

$$\sum_{j=1}^{k}\lambda_j \geq \sum_{j=1}^{k}\mathbf{x}_j^\top A\mathbf{x}_j.$$

Furthermore, $\mathbf{v}_j^\top A\mathbf{v}_j = \lambda_j$ implies

$$\lambda_1 + \cdots + \lambda_k = \sum_{j=1}^{k}\mathbf{v}_j^\top A\mathbf{v}_j = \max\Big\{\sum_{j=1}^{k}\mathbf{x}_j^\top A\mathbf{x}_j : \mathbf{x}_j\text{'s are orthonormal}\Big\}.$$

If $\mathbf{x}_j$ is the j-th column of X, trace $\left(X^\top AX\right) = \sum_{j=1}^{k}\mathbf{x}_j^\top A\mathbf{x}_j$ holds and the first part of the theorem also holds. Since

$$\text{trace}\left(Q^\top X^\top AXQ\right) = \text{trace}\left(X^\top AXQQ^\top\right) = \text{trace}\left(X^\top AX\right)$$

for any orthogonal matrix Q, the second part also holds. ■

5.6 Relationship between Singular Values and Eigenvalues

As clear in Definition 5.1, eigenpair is defined only for square matrices. Even for real square matrices, eigenvalues and eigenvectors are not guaranteed to be real if they are not symmetric. There might be fewer independent eigenvectors than the number of columns. The real spectral decomposition however guarantees the existence of real eigenvalues and enough orthonormal eigenvectors for symmetric matrices, whereas SVD always results in as many singular triplets as the rank of a matrix of any size. As a further difference, singular values are always positive, but eigenvalues of the corresponding matrix may be non-positive. In short, SVD and eigendecomposition are not equivalent.

For a symmetric matrix, we can however correspond each singular triplet with the eigenpair of the symmetric matrix by modifying the sequential approach (5.5) for SVD as in Section 5.5. Conversely, we can obtain an SVD from eigenvalues and orthonormal eigenvectors of a symmetric matrix. Assume we know all eigenpairs $(\lambda_i, \mathbf{v}_i)$, $i = 1, \ldots, n$ of an $n \times n$ matrix A with orthonormal eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Because eigenvectors are orthonormal, we can write A as $A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Let $\operatorname{rank} A = r$ or, equivalently, there be r non-zero eigenvalues, i.e., $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0 = \lambda_{r+1} = \cdots = \lambda_n$. Then, by letting $\sigma_i = |\lambda_i|$ and $\mathbf{u}_i = \operatorname{sign}(\lambda_i)\mathbf{v}_i$ for $i = 1, \ldots, r$, we get an SVD of A, as $A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$.

Let us translate the matrix norms of a symmetric matrix from the terms of singular values into the terms of eigenvalues, which is parallel to Fact 5.3.

Fact 5.7 *If a symmetric matrix A has eigenvalues $\lambda_1, \ldots, \lambda_n$, then*

$$\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i| = \max_{|\mathbf{x}|=1} \left|\mathbf{x}^\top A \mathbf{x}\right| \quad \textit{and} \quad \|A\|_F = \sqrt{\lambda_1^2 + \cdots + \lambda_n^2}\,.$$

Proof: Let $A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, and assume $|\lambda_i| \geq |\lambda_{i+1}|$ for convenience. As above, set $\sigma_i = |\lambda_i|$. We also set $\mathbf{u}_i = -\mathbf{v}_i$ if $\lambda_i < 0$ and $\mathbf{u}_i = \lambda_i$ otherwise. Then, we have an SVD representation of $A = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, and the first and the third equalities hold by Fact 5.3. Note that $\lambda_{r+1} = \cdots = \lambda_n = 0$ if $\operatorname{rank} A = r < n$. For $|\mathbf{x}| = 1$, $|\mathbf{x}^\top A \mathbf{x}| \leq |\mathbf{x}||A\mathbf{x}| = |A\mathbf{x}| \leq \|A\|_2$ and $\max_{|\mathbf{x}|=1} \left|\mathbf{x}^\top A \mathbf{x}\right| \leq \|A\|_2 = |\lambda_1|$. Because $\max_{|\mathbf{x}|=1} \left|\mathbf{x}^\top A \mathbf{x}\right| \geq |\mathbf{v}_1^\top A \mathbf{v}_1| = |\lambda_1|$, the second equality holds. ■

Singular triplets of A and eigenpairs of $A^\top A$

Unlike asymmetric or non-square matrices, symmetric matrices admit complete eigendecomposition thanks to the real spectral decomposition theorem. Therefore, it is common to consider eigenpairs of $A^\top A$ or $AA^\top$ instead of A. Then, how are the eigenpairs of $A^\top A$ or $AA^\top$ connected to the singular triplets of A, when A is not square nor symmetric?

Let A be an $n \times d$ matrix of rank r. Assume that $(\sigma, \mathbf{v}, \mathbf{u})$ is a singular triplet of A such that $A\mathbf{v} = \sigma\mathbf{u}$ and $A^\top \mathbf{u} = \sigma\mathbf{v}$. Then, $A^\top A\mathbf{v} = \sigma A^\top \mathbf{u} = \sigma^2 \mathbf{v}$ and $AA^\top \mathbf{u} = \sigma A\mathbf{v} = \sigma^2 \mathbf{u}$ which implies that $(\sigma^2, \mathbf{v})$ and $(\sigma^2, \mathbf{u})$ are eigenpairs of $A^\top A$ and $AA^\top$, respectively.

Conversely, consider an eigenpair $(\lambda, \mathbf{v})$ of $A^\top A$, where $\lambda \neq 0$ and $|\mathbf{v}| = 1$. Since $A^\top A\mathbf{v} = \lambda\mathbf{v}$, which implies that $|A\mathbf{v}|^2 = \lambda|\mathbf{v}|^2 = \lambda$ after multiplying both sides with $\mathbf{v}^\top$, $\lambda > 0$ holds if $\lambda \neq 0$. If we set $\mathbf{u} = \frac{1}{\sqrt{\lambda}} A\mathbf{v}$, we also see that $|\mathbf{u}| = 1$

and $A^\top \mathbf{u} = \sqrt{\lambda}\mathbf{v}$. That is, $(\sqrt{\lambda}, \mathbf{v}, \mathbf{u})$ is a singular triplet of A. Since $A^\top A$ is symmetric, Theorem 5.2 allows us to assume that all eigenvectors of $A^\top A$ are orthonormal. Let $(\hat{\lambda}, \hat{\mathbf{v}})$ be another eigenpair of $A^\top A$ such that $\hat{\mathbf{v}}^\top \mathbf{v} = 0$. By letting $\hat{\mathbf{u}} = \frac{1}{\sqrt{\hat{\lambda}}} A\hat{\mathbf{v}}$, similarly with $\mathbf{u}$, $\mathbf{u}$ and $\hat{\mathbf{u}}$ are also orthonormal since

$$\hat{\mathbf{u}}^\top \mathbf{u} = \frac{1}{\sqrt{\hat{\lambda}\lambda}} \hat{\mathbf{v}}^\top A^\top A \mathbf{v} = \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \hat{\mathbf{v}}^\top \mathbf{v} = 0\,.$$

Therefore, we get r triplets, $(\sqrt{\lambda_i}, \mathbf{v}_i, \mathbf{u}_i)$, with orthonormal $\mathbf{v}_i$ and $\mathbf{u}_i$ and positive λ_i. The sum of rank-one matrices induced by these triplets, $\sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top$, results in the same vector as the matrix A when multiplied from right by an arbitrary vector. Thus, they coincide, i.e.,

$$A = \sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^\top,$$

and these r triplets are the singular triplets of A due to Fact 5.2. Many commercial software packages compute singular values by computing eigenvalues of $AA^\top$ or $A^\top A$. Though, such decomposition is often not unique, since the correspondence between singular values and eigenvalues is not unique.

Lemma 5.3 *Assume that A is a real matrix of arbitrary size. Then the square of any singular value and right-singular vector of A is an eigenpair of $A^\top A$, and the left-singular vector is an eigenvector of $AA^\top$. Conversely, if $A^\top A$ admits eigenvalues and orthonormal eigenvectors, then the square roots of eigenvalues are singular values of A, and the eigenvectors form right-singular vectors.*

Symmetrization

We define a symmetrization $s(A)$ of an $m \times n$ matrix A as

$$s(A) = \begin{bmatrix} \mathbf{0} & A \\ A^\top & \mathbf{0} \end{bmatrix}. \tag{5.15}$$

As the name suggests, the resulting $(m+n) \times (m+n)$ matrix is symmetric. Symmetrization is linear, since $s(A+\alpha B) = s(A) + \alpha s(B)$. Consider $\mathbb{R}^m$-vector $\mathbf{u}$ and $\mathbb{R}^n$-vector $\mathbf{v}$, as well as the following vectors of size $(m+n)$ to work with symmetrization: $\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$ and $\hat{\mathbf{w}} = \begin{bmatrix} -\mathbf{u} \\ \mathbf{v} \end{bmatrix}$. Then,

- If $(\sigma, \mathbf{v}, \mathbf{u})$ is a singular triplet of A with $\sigma > 0$, $|\mathbf{u}| = 1$, and $|\mathbf{v}| = 1$,

$$s(A)\mathbf{w} = \begin{bmatrix} A\mathbf{v} \\ A^\top \mathbf{u} \end{bmatrix} = \begin{bmatrix} \sigma\mathbf{u} \\ \sigma\mathbf{v} \end{bmatrix} = \sigma\mathbf{w}.$$

  Therefore, $(\sigma, \mathbf{w})$ is an eigenpair of $s(A)$.

- Conversely, let $(\lambda, \mathbf{w})$ an eigenpair of $s(A)$ with $\lambda \neq 0$. Then, both $A\mathbf{v} = \lambda\mathbf{u}$ and $A^\top\mathbf{u} = \lambda\mathbf{v}$ hold, because

$$s(A)\mathbf{w} = \begin{bmatrix} A\mathbf{v} \\ A^\top \mathbf{u} \end{bmatrix} = \lambda\mathbf{w} = \begin{bmatrix} \lambda\mathbf{u} \\ \lambda\mathbf{v} \end{bmatrix}.$$

  Neither $\mathbf{v}$ nor $\mathbf{u}$, which constitute the eigenvector $\mathbf{w}$, can be zero vectors. If one of them is zero, the other has to be also zero from the singular relations. This implies $\mathbf{w} = \mathbf{0}$, but the eigenvector $\mathbf{w}$ is not a zero vector. Therefore, $\mathbf{v}$ is an eigenvector of $A^\top A$ corresponding to an eigenvalue λ^2, and $|\lambda|$ is a singular value of A, according to Lemma 5.3.

- In addition, if $(\lambda, \mathbf{w})$ were an eigenpair of $s(A)$,

$$s(A)\hat{\mathbf{w}} = \begin{bmatrix} A\mathbf{v} \\ A^\top(-\mathbf{u}) \end{bmatrix} = \begin{bmatrix} (-\lambda)(-\mathbf{u}) \\ (-\lambda)\mathbf{v} \end{bmatrix} = (-\lambda)\hat{\mathbf{w}},$$

  because $A\mathbf{v} = \lambda\mathbf{u}$ and $A^\top\mathbf{u} = \lambda\mathbf{v}$. Therefore, $(-\lambda, \hat{\mathbf{w}})$ is also an eigenpair of $s(A)$.

We obtain the following result summarizing these observations.

Lemma 5.4 *The symmetrization $s(A)$ has its eigenvalues in both signs; that is, if λ is an eigenvalue of $s(A)$, then both $\pm|\lambda|$ are eigenvalues of $s(A)$. There exists a one-to-one correspondence between the singular values of A and the positive eigenvalues of $s(A)$.*

This lemma is useful when we convert any result on the eigenvalues of a symmetric matrix into that on the singular values of a matrix of arbitrary size.

5.7 Low Rank Approximation and Eckart–Young–Mirsky Theorem

Suppose that an SVD of a matrix is given as

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \quad \sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1}.$$

Let A_k be a matrix resulting from taking the summands that correspond to the largest k singular values, as follows

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \tag{5.16}$$

Then, $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Based on Fact 5.4, the ranks of A, A_k and $A - A_k$ are r, k, and $r - k$, respectively.

Low Rank Approximation: Spectral Norm

Lemma 5.5 *For any matrix B of rank at most k, $\|A - A_k\|_2 \leq \|A - B\|_2$.*

Proof: Suppose that B is an arbitrary matrix of rank at most k.

- Because $A - A_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, $\sigma_{k+1} \geq \cdots \geq \sigma_r > 0$,

  $$\|A - A_k\|_2 = \sigma_{k+1},$$

  according to Fact 5.3, for $k \leq r$.

- The rank of the null space of B is at least $d - k$, since the rank of B is at most k. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ be the right-singular vectors of A that correspond to the $k+1$ largest singular values. The dimension of Null $(B) \cap$ span$\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ is at least $(d-k)+(k+1)-d = 1$ due to Fact 3.6, and the intersection includes a non-zero unit-vector $\mathbf{z}$. Note that $B\mathbf{z} = \mathbf{0}$ from $\mathbf{z} \in$ Null (B) and $\mathbf{z} = \sum_{j=1}^{k+1} \langle \mathbf{z}, \mathbf{v}_j \rangle \mathbf{v}_j$ from $\mathbf{z} \in$ span$\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$. Then,

  $$A\mathbf{z} = \Big(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \Big) \Big(\sum_{j=1}^{k+1} \langle \mathbf{z}, \mathbf{v}_j \rangle \mathbf{v}_j \Big) = \sum_{i=1}^{k+1} \sigma_i \langle \mathbf{z}, \mathbf{v}_i \rangle \mathbf{u}_i.$$

  Thus,

  $$\|A - B\|_2 \quad \geq \quad |(A - B)\mathbf{z}| \qquad \text{(by the definition of spectral norm)}$$

$$
\begin{aligned}
&= |A\mathbf{z}| && (\text{by } B\mathbf{z} = \mathbf{0}) \\
&= \Big(\sum_{i=1}^{k+1} \sigma_i^2 \langle \mathbf{z}, \mathbf{v}_i\rangle^2\Big)^{1/2} && \Big(\text{by } A\mathbf{z} = \sum_{i=1}^{k+1} \sigma_i \langle \mathbf{z}, \mathbf{v}_i\rangle \mathbf{u}_i\Big) \\
&\geq \sigma_{k+1}\Big(\sum_{i=1}^{k+1} \langle \mathbf{z}, \mathbf{v}_i\rangle^2\Big)^{1/2} && (\text{by } \sigma_1 \geq \cdots \geq \sigma_k \geq \sigma_{k+1}) \\
&= \sigma_{k+1} && \Big(\text{by } \sum_{i=1}^{k+1} \langle \mathbf{z}, \mathbf{v}_i\rangle^2 = |\mathbf{z}|^2 = 1\Big) \\
&= \|A - A_k\|_2 \,.
\end{aligned}
$$

Therefore, $\|A - A_k\|_2 \leq \|A - B\|_2$ holds for any matrix B of rank at most k. ■

Low Rank Approximation: Frobenius Norm

Lemma 5.6 *For any matrix B of rank at most k, $\|A - A_k\|_F \leq \|A - B\|_F$.*

Proof: Assume that $\|A - B\|_F$ is minimized by an $n \times d$ matrix B of rank at most k. Let $\mathbf{b}_1, \ldots, \mathbf{b}_n$ be the rows of this matrix B. We further assume that the projection of the i-th row $\mathbf{a}_i$ of A onto $\text{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is not $\mathbf{b}_i$. That is, $\mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i) \neq \mathbf{b}_i$. We create B' from B by replacing $\mathbf{b}_i$ with $\mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)$. Then, $\|A - B\|_F > \|A - B'\|_F$, since $|\mathbf{a}_i - \mathbf{b}_i| > \big|\mathbf{a}_i - \mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)\big|$ (due to the distance minimizing property of orthogonal projection) and $\|A - B\|_F^2 = \sum_{i=1}^n |\mathbf{a}_i - \mathbf{b}_i|^2$. The rank of B' is however not greater than that of B, since $\mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)$ is a linear combination of the rows of B. This is contradictory with the assumption of minimal Frobenius norm, and thus $\mathbf{b}_i = \mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)$. That is,

$$\min_{B:\text{rank}(B)\leq k} \|A - B\|_F^2 = \min_{B:\text{rank}(B)\leq k} \sum_{i=1}^n \big|\mathbf{a}_i - \mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)\big|^2.$$

Finding $\text{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ among matrices of rank at most k is equivalent to finding a subspace of dimension up to k, i.e.,

$$\min_{B:\text{rank}(B)\leq k} \sum_{i=1}^n \big|\mathbf{a}_i - \mathbf{P}_{\text{span}\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}}(\mathbf{a}_i)\big|^2 = \min_{\mathbb{B}:\dim(\mathbb{B})\leq k} \sum_{i=1}^n \big|\mathbf{a}_i - \mathbf{P}_{\mathbb{B}}(\mathbf{a}_i)\big|^2.$$

Furthermore, because we can represent a subspace with a basis,

$$\min_{\mathbb{B}:\dim(\mathbb{B})\le k} \sum_{i=1}^{n} \left|\mathbf{a}_i - \mathbf{P}_{\mathbb{B}}(\mathbf{a}_i)\right|^2 = \min_{\substack{\mathbf{v}_1,\dots,\mathbf{v}_k:\\ \text{orthonormal}}} \sum_{i=1}^{n} \left|\mathbf{a}_i - \mathbf{P}_{\text{span}\{\mathbf{v}_1,\dots,\mathbf{v}_k\}}(\mathbf{a}_i)\right|^2.$$

By putting all these equations together, we get

$$\begin{aligned}
\min_{B:\text{rank}(B)\le k} \|A-B\|_F^2 &= \min_{\substack{\mathbf{v}_1,\dots,\mathbf{v}_k:\\ \text{orthonormal}}} \sum_{i=1}^{n} \left|\mathbf{a}_i - \mathbf{P}_{\text{span}\{\mathbf{v}_1,\dots,\mathbf{v}_k\}}(\mathbf{a}_i)\right|^2 \\
&= \min_{\substack{\mathbf{v}_1,\dots,\mathbf{v}_k:\\ \text{orthonormal}}} \sum_{i=1}^{n} \left|\mathbf{a}_i\right|^2 - \left|\mathbf{P}_{\text{span}\{\mathbf{v}_1,\dots,\mathbf{v}_k\}}(\mathbf{a}_i)\right|^2 \\
&= \sum_{i=1}^{n} \left|\mathbf{a}_i\right|^2 - \max_{\substack{\mathbf{v}_1,\dots,\mathbf{v}_k:\\ \text{orthonormal}}} \sum_{i=1}^{n} \left|\mathbf{P}_{\text{span}\{\mathbf{v}_1,\dots,\mathbf{v}_k\}}(\mathbf{a}_i)\right|^2 \\
&= \|A\|_F^2 - \max_{\substack{\mathbf{v}_1,\dots,\mathbf{v}_k:\\ \text{orthonormal}}} |A\mathbf{v}_1|^2 + \cdots + |A\mathbf{v}_k|^2 \\
&= (\sigma_1^2 + \cdots + \sigma_r^2) - (\sigma_1^2 + \cdots + \sigma_k^2) \qquad \text{by (5.6)} \\
&= \sigma_{k+1}^2 + \cdots + \sigma_r^2 \\
&= \|A - A_k\|_F^2 \qquad \text{by (5.6)}
\end{aligned}$$

Therefore, $\|A - A_k\|_F \le \|A - B\|_F$ for any matrix B of rank at most k. ■

Consider $n \times d$ matrices A and A_k in (5.16). By Lemma 5.5,

$$\|A - A_k\|_2 \le \|A - B\|_2$$

holds for any matrix B of rank at most k. In addition, thanks to Fact 5.3,

$$\|A - A_k\|_2 = \Big\| \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \Big\|_2 = \sigma_{k+1}$$

holds. We combine these results with $\operatorname{rank} A_k = k$ to arrive at the following Eckart–Young–Mirsky Theorem.

Theorem 5.4 (Eckart–Young–Mirsky Theorem) *For any matrix A and its i-th singular value σ_i, the following holds for any k:*

$$\min_{B:\text{rank}\, B\le k} \|A - B\|_2 = \sigma_{k+1}\,. \tag{5.17}$$

5.8 Pseudoinverse

We used SVD to express the inverse of a square invertible matrix in Fact 5.6. In this section, we define pseudoinverse, or Moore-Penrose generalized inverse, in a similar form, for general matrices including non-square matrices or non-invertible matrices.

Definition 5.2 (Pseudoinverse) *Let A be an $n \times d$ matrix of rank r and $A = U\Sigma V^\top$ be a singular value decomposition of A. Then, we call the following $d \times n$ matrix as the pseudoinverse of A:*

$$A^+ = V\Sigma^{-1}U^\top, \tag{5.18}$$

where $\Sigma^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}\right)$ with singular values $\sigma_1, \ldots, \sigma_r$ of A.

V and U in (5.18) are not uniquely determined for an arbitrary matrix A, because its SVD is not unique in general. It is therefore unclear whether the pseudoinverse, constructed from V and U above, is unique. It is thus surprising to find that the pseudoinverse of any matrix is unique.

Both U and V consist of orthonormal columns, but they may not be orthogonal. In other words, we can rely only on the fact that $U^\top U = I_r = V^\top V$ since we do not know what $UU^\top$ and $VV^\top$ are. From these equalities, we know that AA^+ and A^+A are symmetric matrices, since

$$AA^+ = U\Sigma V^\top V\Sigma^{-1}U^\top = UU^\top \quad \text{and} \quad A^+A = V\Sigma^{-1}U^\top U\Sigma V^\top = VV^\top.$$

Furthermore, we can also show that $AA^+A = UU^\top U\Sigma V^\top = U\Sigma V^\top = A$ and $A^+AA^+ = VV^\top V\Sigma^{-1}U^\top = V\Sigma^{-1}U^\top = A^+$. We call these properties collectively as the Penrose identities.

Definition 5.3 (Penrose identities) *An $n \times d$ matrix A and a $d \times n$ matrix B satisfy the Penrose identities if A and B satisfy the following identities:*

(*a*) $(AB)^\top = AB$ *and* $(BA)^\top = BA$;

(*b*) $ABA = A$;

(*c*) $BAB = B$.

Surprisingly, for any matrix A, there is a unique matrix B that satisfies the Penrose identities.

Fact 5.8 *There is a unique matrix satisfying the Penrose identities with any matrix.*

Proof: Assume there are two matrices, B and C, that satisfy the Penrose identities with a given A. First, we list three identities in Definition 5.3 for B and C as follows:

$$(t1)\ (AB)^\top = AB, \quad (t2)\ (BA)^\top = BA;$$

$$(t1')\ (AC)^\top = AC, \quad (t2')\ (CA)^\top = CA;$$

$$(a)\ ABA = A, \qquad (a')\ ACA = A;$$

$$(b)\ BAB = B, \qquad (b')\ CAC = C.$$

From these identities,

$$\begin{aligned} B &\overset{(b)}{=} BAB \overset{(t1)}{=} B(AB)^\top = BB^\top A^\top \overset{(a')}{=} BB^\top (ACA)^\top = BB^\top A^\top (AC)^\top \\ &= B(AB)^\top (AC)^\top \overset{(t1),(t1')}{=} BABAC = (BAB)AC \overset{(b)}{=} BAC \\ &\overset{(t2)}{=} (BA)^\top C = A^\top B^\top C \overset{(a')}{=} (ACA)^\top B^\top C = (CA)^\top A^\top B^\top C \\ &= (CA)^\top (BA)^\top C \overset{(t2'),(t2)}{=} CABAC = C(ABA)C \overset{(a)}{=} CAC \\ &\overset{(b')}{=} C\,. \end{aligned}$$

That is, there is only one matrix that satisfies all three Penrose identities. ■

From this property, we can show that the Penrose identities are equivalent to the definition of pseudoinverse.

Fact 5.9 *A $d \times n$ matrix is a pseudoinverse of A if and only if the matrix satisfies the Penrose identities with A.*

Proof: We already showed above "only if", which tells us that A^+ satisfies the Penrose identities. According to Fact 5.8, there is only one matrix that satisfies the Penrose identities with A, and thereby this matrix is the pseudoinverse. ■

By combining above two Facts, we conclude that a pseudoinverse is unique.

Theorem 5.5 *A pseudoinverse of a matrix is unique.*

Proof: We can show this based on Facts 5.8 and 5.9. ■

This uniqueness result with the Penrose identities in Fact 5.9 allows us to say that a matrix is the pseudoinverse of A by simply checking if it satisfies the Penrose identities with A. We do not need to perform SVD on A.

As an example, let us compute the pseudoinverse as well as a low-rank approximation of the following matrix given as a sum of rank-one matrices.

Example 5.4 [Example 5.3 revisited] Consider the following 4×4 matrix A:

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

1. We already computed the compact SVD of this matrix in Example 5.3:

$$\begin{aligned} A &= 6 \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} + 4 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} \\ &\quad + 2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \\ &= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^\top . \end{aligned}$$

Therefore, its pseudoinverse is

$$\begin{aligned} A^+ &= \frac{1}{\sigma_1} \mathbf{u}_1 \mathbf{v}_1^\top + \frac{1}{\sigma_2} \mathbf{u}_2 \mathbf{v}_2^\top + \frac{1}{\sigma_3} \mathbf{u}_3 \mathbf{v}_3^\top \\ &= \frac{1}{6} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{-\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} \\ &\quad + \frac{1}{2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&= VD^{-1}U^\top \\
&= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/6 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.
\end{aligned}
$$

2. By Lemma 5.5,

$$
\begin{aligned}
B &= A_2 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top \\
&= 6 \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} + 4 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix}
\end{aligned}
$$

minimizes $\|A - B\|_2$ among 4×4 matrices of rank 2.

∎

Another example of pseudoinverse is a left- or right-inverse of a non-square full-rank matrix.

Fact 5.10 *Let the rank of an $n \times d$ matrix A be d. Then, $A^+ = (A^\top A)^{-1} A^\top$, and A^+ is a left-inverse of A. If $n > d$, then A^+ is not a right-inverse of A. For the case of $n = d$, A^+ is the inverse of A.*

Proof: $(A^\top A)^{-1} A^\top$ is the pseudoinverse of A by Fact 5.9 since it satisfies the Penrose identities with A. Since $A^+ A = (A^\top A)^{-1} A^\top A = I_d$, A^+ is a left-inverse of A. However, A has no inverse if $n > d$, and A^+ is not a right-inverse of A in this case. If $n = d$, then A is invertible since the rank $A = n = d$, and $A^+ = (A^\top A)^{-1} A^\top = A^{-1} (A^\top)^{-1} A^\top = A^{-1}$.

∎

Because the following holds for pseudoinverse,

$$A^\top (AA^+ - I) = V \Sigma U^\top (UU^\top - I) = \mathbf{0},$$

we can show for an arbitrary pair of vectors, $\mathbf{x}$ and $\mathbf{b}$, that

$$
\begin{aligned}
|A\mathbf{x} - \mathbf{b}|^2 &= |A\mathbf{x} - AA^+\mathbf{b} + AA^+\mathbf{b} - \mathbf{b}|^2 \\
&= |A(\mathbf{x} - A^+\mathbf{b}) + (AA^+ - I)\mathbf{b}|^2 \\
&= \left|A(\mathbf{x} - A^+\mathbf{b})\right|^2 + 2(\mathbf{x} - A^+\mathbf{b})^\top \underbrace{A^\top (AA^+ - I)}_{=\mathbf{0}} \mathbf{b} + \left|(AA^+ - I)\mathbf{b}\right|^2
\end{aligned}
$$

$$\begin{aligned} &= \left|A(\mathbf{x} - A^+\mathbf{b})\right|^2 + \left|AA^+\mathbf{b} - \mathbf{b}\right|^2 \\ &\geq \left|AA^+\mathbf{b} - \mathbf{b}\right|^2. \end{aligned}$$

In short, for an arbitrary matrix A and an arbitrary vector $\mathbf{b}$,

$$|A\mathbf{x} - \mathbf{b}| \geq \left|AA^+\mathbf{b} - \mathbf{b}\right| \quad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{5.19}$$

That is, when a linear system $A\mathbf{x} = \mathbf{b}$ is not solvable, $A^+\mathbf{b}$ is an approximate solution that minimizes the error measured in the Euclidean length.

Similarly, with $A^- = (A^+)^\top$,

$$A(A^\top A^- - I) = U\Sigma V^\top (VV^\top - I) = \mathbf{0},$$

because $A^\top A^- = (A^+A)^\top = VV^\top$. From this, we obtain the following inequality, transposed version of (5.19):

$$\begin{aligned} |A^\top\mathbf{y} - \mathbf{c}|^2 &= |A^\top(\mathbf{y} - A^-\mathbf{c}) + (A^\top A^-\mathbf{c} - \mathbf{c})|^2 \\ &= \left|A^\top(\mathbf{y} - A^-\mathbf{c})\right|^2 + 2(\mathbf{y} - A^-\mathbf{c})^\top \underbrace{A(A^\top A^- - I))}_{=\mathbf{0}}\mathbf{c} + \left|A^\top A^-\mathbf{c} - \mathbf{c}\right|^2 \\ &= \left|A^\top(\mathbf{y} - A^-\mathbf{c})\right|^2 + \left|A^\top A^-\mathbf{c} - \mathbf{c}\right|^2 \\ &\geq \left|A^\top A^-\mathbf{c} - \mathbf{c}\right|^2, \end{aligned}$$

given appropriately-sized $\mathbf{y}$ and $\mathbf{c}$. Therefore, for any given matrix A and vector $\mathbf{c}$,

$$\left|\mathbf{y}^\top A - \mathbf{c}^\top\right| \geq \left|\mathbf{c}^\top A^+A - \mathbf{c}^\top\right| \quad \text{for all } \mathbf{y} \in \mathbb{R}^m. \tag{5.20}$$

Using these two inequalities, we arrive at the following characterization of pseudoinverse.

Theorem 5.6 *The pseudoinverse of A is a matrix X that minimizes $\|AX - I_n\|_F$, that is,*

$$A^+ = \operatorname*{argmin}_{X\,:\,d\times n \text{ matrix}} \|AX - I_n\|_F = \operatorname*{argmin}_{Y\,:\,d\times n \text{ matrix}} \|YA - I_d\|_F \tag{5.21}$$

Proof: Let $\mathbf{e}_j$ be the j-th standard basic vector in $\mathbb{R}^n$. Note that $\|B\|_F^2 = \sum_{j=1}^n \left|B\mathbf{e}_j\right|^2$ for any matrix B with n columns. If replace $\mathbf{x}$ in (5.19) with the

j-row of an arbitrary matrix X, i.e., $X\mathbf{e}_j$ and set $\mathbf{b}$ to be $\mathbf{e}_j$, we get

$$|AX\mathbf{e}_j - \mathbf{e}_j| \geq |AA^+\mathbf{e}_j - \mathbf{e}_j|.$$

From this, we can establish the lower bound of $\|AX - I_n\|_F^2$ as

$$\|AX - I_n\|_F^2 = \sum_{j=1}^{n} |AX\mathbf{e}_j - \mathbf{e}_j|^2 \geq \sum_{j=1}^{n} |AA^+\mathbf{e}_j - \mathbf{e}_j|^2 = \|AA^+ - I_n\|_F^2.$$

Let $\mathbf{e}_i$ be the i-th standard basic vector in $\mathbb{R}^d$. Note that $\|B\|_F^2 = \sum_{i=1}^{d} \left|\mathbf{e}_i^\top B\right|^2$ for any matrix B with d rows. Similarly, from (5.20), we get

$$|\mathbf{e}_i^\top YA - \mathbf{e}_i^\top| \geq |\mathbf{e}_i^\top A^+A - \mathbf{e}_i^\top|,$$

and subsequently

$$\|YA - I_d\|_F^2 = \sum_{i=1}^{d} |\mathbf{e}_i^\top YA - \mathbf{e}_i^\top|^2 \geq \sum_{i=1}^{d} |\mathbf{e}_i^\top A^+A - \mathbf{e}_i^\top|^2 = \|A^+A - I_d\|_F^2 .$$

Since A^+ is a $d \times n$ matrix, A^+ is a minimizer of (5.21). ■

5.8.1 Generalized Projection and Least Squares

Let us consider a projection or a least square problem in a more general setup. In this setup, the matrix $A = [\,\mathbf{a}_1 \,|\, \cdots \,|\, \mathbf{a}_d\,]$ may not be a full-rank matrix. In order to approximately solve $A\mathbf{x} = \mathbf{b}$, we need to either find $\mathbf{x}$ that minimizes $|A\mathbf{x} - \mathbf{b}|$, or project $\mathbf{b}$ onto the column space $\mathrm{Col}\,(A) = \mathrm{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$. From the inequality in (5.19), we know that $\hat{\mathbf{x}} = A^+\mathbf{b}$ minimizes $|A\mathbf{x} - \mathbf{b}|$. To connect this length minimizing solution $\hat{\mathbf{x}}$ to a projection, let us investigate the orthogonality of $\mathbf{b} - A\hat{\mathbf{x}}$ and $A\hat{\mathbf{x}}$ through the Penrose identities.

$$\begin{aligned}(\mathbf{b} - A\hat{\mathbf{x}})^\top A\hat{\mathbf{x}} &= (\mathbf{b} - AA^+\mathbf{b})^\top AA^+\mathbf{b} = \mathbf{b}^\top(I - AA^+)AA^+\mathbf{b} \\ &= \mathbf{b}^\top(AA^+ - AA^+AA^+)\mathbf{b} = \mathbf{b}^\top(AA^+ - AA^+)\mathbf{b} = \mathbf{0}\end{aligned}$$

implies that $\mathbf{b} - A\hat{\mathbf{x}}$ and $A\hat{\mathbf{x}}$ are orthogonal to each other. Since $A\hat{\mathbf{x}} \in \mathrm{Col}\,(A)$, $A\hat{\mathbf{x}}$ is the projection of $\mathbf{b}$ onto $\mathrm{Col}\,(A)$. In summary, $A^+\mathbf{b}$ is the solution to the least squares problem, and at the same time, we can express the orthogonal projection of $\mathbf{b}$ onto the column space, $\mathrm{Col}\,(A)$, as

$$\mathbf{P}_{\mathrm{Col}\,(A)}(\mathbf{b}) = AA^+\mathbf{b}, \tag{5.22}$$

where AA^+ is a matrix corresponding to the orthogonal projection onto $\mathrm{Col}\,(A)$.

If the rank of A was d, the least squares solution is $(A^\top A)^{-1}A^\top \mathbf{b}$, and the orthogonal projection is $A(A^\top A)^{-1}A^\top \mathbf{b}$. These coincide with the earlier results, as implied by Fact 5.10. If A is a low-rank matrix, the pseudoinverse of A may not be expressed in the same way.

Let us summarize various orthogonal projections derived for an $n \times d$ matrix $A = [\,\mathbf{a}_1 \,|\, \cdots \,|\, \mathbf{a}_d\,]$:

- Without any particular constraint on A,

$$\mathbf{P}_{\mathrm{Col}\,(A)}(\mathbf{x}) = AA^+\mathbf{x} \qquad \text{and} \qquad A^+\,;$$

- If the columns of A are linearly independent, i.e. $\operatorname{rank} A = d$,

$$\mathbf{P}_{\mathrm{Col}\,(A)}(\mathbf{x}) = A(A^\top A)^{-1}A^\top \mathbf{x} \qquad \text{and} \qquad A^+ = (A^\top A)^{-1}A^\top\,;$$

- If the columns of A are orthonormal,

$$\mathbf{P}_{\mathrm{Col}\,(A)}(\mathbf{x}) = AA^\top \mathbf{x} \qquad \text{and} \qquad A^+ = A^\top\,;$$

- If there is only one column, $\mathbf{a}$, in A,

$$\mathbf{P}_{\mathrm{Col}\,(A)}(\mathbf{x}) = \frac{1}{\mathbf{a}^\top\mathbf{a}}\mathbf{a}\mathbf{a}^\top\mathbf{x} \qquad \text{and} \qquad A^+ = \frac{1}{\mathbf{a}^\top\mathbf{a}}\mathbf{a}^\top\,.$$

Each of these projections is a special case of the one above.

5.9 Numerically Stable QR-Decomposition

Now that we know how to compute the rank of a sum of rank-one matrices, according to Fact 5.4, we now re-visit the Gram-Schmidt procedure (4.12) from the perspective of numerical stability.

Using the standard inner product in $\mathbb{R}^n$, we can re-write the orthogonal projection (4.12) onto known basic vectors in Gram-Schmidt procedure as

$$\begin{aligned}
\mathbf{v}_{i+1} &= \mathbf{a}_{i+1} - \big(\mathbf{q}_1^\top \mathbf{a}_{i+1}\big)\mathbf{q}_1 - \big(\mathbf{q}_2^\top \mathbf{a}_{i+1}\big)\mathbf{q}_2 - \cdots - \big(\mathbf{q}_i^\top \mathbf{a}_{i+1}\big)\mathbf{q}_i \\
&= \mathbf{a}_{i+1} - \mathbf{q}_1\big(\mathbf{q}_1^\top \mathbf{a}_{i+1}\big) - \mathbf{q}_2\big(\mathbf{q}_2^\top \mathbf{a}_{i+1}\big) - \cdots - \mathbf{q}_i\big(\mathbf{q}_i^\top \mathbf{a}_{i+1}\big) \\
&= \big(I - \mathbf{q}_1\mathbf{q}_1^\top - \mathbf{q}_2\mathbf{q}_2^\top - \cdots - \mathbf{q}_i\mathbf{q}_i^\top\big)\mathbf{a}_{i+1}\,.
\end{aligned}$$

Let us focus on the $n \times n$ matrix

$$\hat{P} = I - \mathbf{q}_1\mathbf{q}_1^\top - \mathbf{q}_2\mathbf{q}_2^\top - \cdots - \mathbf{q}_i\mathbf{q}_i^\top.$$

For any orthonormal basis $\{\mathbf{q}_1, \ldots, \mathbf{q}_i, \mathbf{q}_{i+1}, \ldots \mathbf{q}_n\}$ extended from $\{\mathbf{q}_1, \ldots, \mathbf{q}_i\}$, it holds that $\hat{P}\mathbf{q}_j = \mathbf{0}$ for $1 \le j \le i$ and $\hat{P}\mathbf{q}_j = \mathbf{q}_j$ for $i+1 \le j \le n$. This implies that

$$\hat{P} = \mathbf{q}_{i+1}\mathbf{q}_{i+1}^\top + \cdots \mathbf{q}_n\mathbf{q}_n^\top$$

and $\operatorname{rank}\hat{P} = n - i$ by Fact 5.4. Therefore, the step (4.12) in the original Gram-Schmidt procedure multiplies a matrix of rank $n - i$.

There is another way to represent $\hat{P}$ as the product of higher-rank matrices. Because of the orthonormality of $\mathbf{q}_j$ and $\mathbf{q}_k$ for $j \neq k$,

$$\big(I - \mathbf{q}_j\mathbf{q}_j^\top\big)\big(I - \mathbf{q}_k\mathbf{q}_k^\top\big) = I - \mathbf{q}_j\mathbf{q}_j^\top - \mathbf{q}_k\mathbf{q}_k^\top + \mathbf{q}_j\mathbf{q}_j^\top\mathbf{q}_k\mathbf{q}_k^\top = I - \mathbf{q}_j\mathbf{q}_j^\top - \mathbf{q}_k\mathbf{q}_k^\top.$$

This implies furthermore that

$$\big(I - \mathbf{q}_j\mathbf{q}_j^\top\big)\big(I - \mathbf{q}_k\mathbf{q}_k^\top\big) = \big(I - \mathbf{q}_k\mathbf{q}_k^\top\big)\big(I - \mathbf{q}_j\mathbf{q}_j^\top\big),$$

and

$$\big(I - \mathbf{q}_j\mathbf{q}_j^\top\big)\big(I - \mathbf{q}_k\mathbf{q}_k^\top\big)\big(I - \mathbf{q}_\ell\mathbf{q}_\ell^\top\big) = I - \mathbf{q}_j\mathbf{q}_j^\top - \mathbf{q}_k\mathbf{q}_k^\top - \mathbf{q}_\ell\mathbf{q}_\ell^\top.$$

Repeating this further, we get

$$\begin{aligned}\hat{P} &= I - \mathbf{q}_1\mathbf{q}_1^\top - \mathbf{q}_2\mathbf{q}_2^\top - \cdots - \mathbf{q}_i\mathbf{q}_i^\top \\ &= \big(I - \mathbf{q}_1\mathbf{q}_1^\top\big)\big(I - \mathbf{q}_2\mathbf{q}_2^\top\big)\cdots\big(I - \mathbf{q}_i\mathbf{q}_i^\top\big) \\ &= \big(I - \mathbf{q}_i\mathbf{q}_i^\top\big)\big(I - \mathbf{q}_{i-1}\mathbf{q}_{i-1}^\top\big)\cdots\big(I - \mathbf{q}_1\mathbf{q}_1^\top\big). \qquad (5.23)\end{aligned}$$

In this way, we obtain $\hat{P}$ by multiplying matrices of rank $n - 1$ rather than (4.12). We call this approach a modified Gram-Schmidt procedure, and this is a popular way to implement the Gram-Schmidt procedure in practice.

We can illustrate the difference between the classical (4.12) and modified (5.23) versions with a sample example. Pick a small ϵ such that $\epsilon^2 = 0$ due to the finite numerical precision. Then, consider the following ill-conditioned matrix, due to near-identical columns:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{bmatrix} = [\mathbf{a}_1\, \mathbf{a}_2\, \mathbf{a}_3].$$

Because $|\mathbf{a}_1| = |\mathbf{a}_2| = |\mathbf{a}_3| = 1$ numerically, both versions result in the same first and second orthonormal vectors $\mathbf{q}_1 = \mathbf{a}_1^\top$ and $\mathbf{q}_2 = \left(0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)^\top$.

The third orthonormal vector however requires a more careful treatment. In the case of the classical procedure, the third vector ends up with

$$\mathbf{q}_3 = \left(0, -\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)^\top,$$

as

$$(I - \mathbf{q}_1\mathbf{q}_1^\top - \mathbf{q}_2\mathbf{q}_2^\top)\mathbf{a}_3 = \mathbf{a}_3 - (\mathbf{a}_1^\top\mathbf{a}_3)\mathbf{a}_1 - (\mathbf{q}_2^\top\mathbf{a}_3)\mathbf{q}_2 = \mathbf{a}_3 - 1\cdot\mathbf{a}_1 - 0\cdot\mathbf{q}_2 = \mathbf{a}_3 - \mathbf{a}_1 = \begin{bmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{bmatrix}.$$

On the other hand, the modified procedure results in

$$\begin{aligned}(I - \mathbf{q}_2\mathbf{q}_2^\top)(I - \mathbf{q}_1\mathbf{q}_1^\top)\mathbf{a}_3 &= (I - \mathbf{q}_2\mathbf{q}_2^\top)(\mathbf{a}_3 - (\mathbf{a}_1^\top\mathbf{a}_3)\mathbf{a}_1) \\ = \quad (I - \mathbf{q}_2\mathbf{q}_2^\top) \begin{bmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{bmatrix} &= \begin{bmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{bmatrix} - \frac{\epsilon}{\sqrt{2}}\mathbf{q}_2 = \begin{bmatrix} 0 \\ -\epsilon/2 \\ -\epsilon/2 \\ \epsilon \end{bmatrix}.\end{aligned}$$

This difference manifests itself when we check the orthogonality of the resulting Q. The classical procedure leaves us with a wrong orthogonal matrix

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ \epsilon & -0.7071 & -0.7071 \\ 0 & 0.7071 & 0 \\ 0 & 0 & 0.7071 \end{bmatrix} \text{ with } Q^\top Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}.$$

The modified one however produces an orthogonal matrix

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ \epsilon & -0.7071 & -0.4082 \\ 0 & 0.7071 & -0.4082 \\ 0 & 0 & 0.8165 \end{bmatrix} \text{ with } Q^\top Q = I,$$

numerically. Despite the mathematical equivalence between these two procedures, they produce different result numerically, emphasizing the importance of practical implementations of these algorithms with finite-precision arithmetic. See [8] for more details.

5.10 How to Obtain SVD Meaningfully

When a matrix is provided us in a form similar to (5.7), we can obtain singular values and vectors by Fact 5.2. For instance, if the rank of the matrix is one, we can readily use Fact 5.2 to obtain a singular triplet. If this is not the case, it is usual to solve the symmetric eigenvalue problem on a smaller-sized one of $AA^\top$ and $A^\top A$. The eigenvectors then serve as right or left-singular vectors, and the square root of the eigenvalues are singular values. We can compute the remaining singular vectors (either left or right) using Lemma 5.2.

Centering Data

When we look for the optimal k-dimensional subspace to represent data, we must decide first whether we are looking for a k-dimensional subspace or a k-dimensional affine space.[2] In the latter case, we first subtract the mean vector from each data point so that the centroid of the data set is located at the origin and then perform SVD.

Fact 5.11 *The k-dimensional affine space that minimizes the sum of squared perpendicular distances to the data points must pass through the centroid of the points.*

Proof: Set the centroid $\boldsymbol{\mu}$ of data points $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subset \mathbb{R}^d$ as $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i$. Consider modified data points $\{\mathbf{a}_1 - \boldsymbol{\mu}, \ldots, \mathbf{a}_n - \boldsymbol{\mu}\}$ which satisfies $\sum_{i=1}^n (\mathbf{a}_i - \boldsymbol{\mu}) = \mathbf{0}$. We desire to show that the best-fit affine space of the modified data points $\{\mathbf{a}_1 - \boldsymbol{\mu}, \ldots, \mathbf{a}_n - \boldsymbol{\mu}\}$ is actually a subspace passing the origin. For simplicity of presentation, we assume that the original data set $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ satisfies $\sum_{i=1}^n \mathbf{a}_i = \mathbf{0}$.

Let $\mathbb{A} = \{\mathbf{v}_0 + \sum_{j=1}^k c_j \mathbf{v}_j : c_1, \ldots, c_k \in \mathbb{R}\}$ be a k-dimensional best-fit affine space of $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, where $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthonormal vectors. Let $\mathbf{v}_0$ be a point in $\mathbb{A}$ that is closest to the origin $\mathbf{0}$. The $\mathbf{v}_0$ must be orthogonal to all $\mathbf{v}_j$'s. The closest point to $\mathbf{a}_i$ in $\mathbb{A}$ is the projection of $\mathbf{a}_i$ onto $\mathbb{A}$, and we represent it as $\mathbf{v}_0 + \sum_{j=1}^k c_j^* \mathbf{v}_j$. The vector that represents the difference between $\mathbf{a}_i$ and this projected vector is orthogonal to $\mathbb{A}$. That is, $\langle \mathbf{a}_i - \mathbf{v}_0 - \sum_{j=1}^k c_{ij}^* \mathbf{v}_j, \mathbf{v}_\ell \rangle = 0$

[2] An affine space is constructed by translating a linear space off the origin and is often expressed as $\{\mathbf{v}_0 + x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k : (x_1, \ldots, x_k) \in \mathbb{R}^k\}$. Linear transformation is not preserved within an affine space but affine combination is. That is, if $\mathbf{v}_1$ and $\mathbf{v}_2$ are included in an affine space, $\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2$ is also within the same space for an arbitrary real-valued λ.

for all $\ell = 1, \ldots, k$. By re-arranging terms, we get $c^*_{ij} = \langle \mathbf{a}_i - \mathbf{v}_0, \mathbf{v}_j \rangle = \langle \mathbf{a}_i, \mathbf{v}_j \rangle$. The sum of squared distances from the data points to $\mathbb{A}$ is then

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{dist}(\mathbf{a}_i, \mathbb{A})^2 &= \sum_{i=1}^{n} \Big| \mathbf{a}_i - \mathbf{v}_0 - \sum_{j=1}^{k} c^*_{ij} \mathbf{v}_j \Big|^2 \\
&= \sum_{i=1}^{n} \Big\{ \big|\mathbf{a}_i - \mathbf{v}_0\big|^2 - 2\Big\langle \mathbf{a}_i - \mathbf{v}_0, \sum_{j=1}^{k} c^*_{ij} \mathbf{v}_j \Big\rangle + \Big| \sum_{j=1}^{k} c^*_{ij} \mathbf{v}_j \Big|^2 \Big\} \\
&= \sum_{i=1}^{n} \Big\{ \big|\mathbf{a}_i - \mathbf{v}_0\big|^2 - 2 \sum_{j=1}^{k} c^*_{ij} \langle \mathbf{a}_i - \mathbf{v}_0, \mathbf{v}_j \rangle + \sum_{j=1}^{k} (c^*_{ij})^2 \Big\} \\
&= \sum_{i=1}^{n} \Big\{ \big|\mathbf{a}_i - \mathbf{v}_0\big|^2 - \sum_{j=1}^{k} (c^*_{ij})^2 \Big\} \\
&= \sum_{i=1}^{n} \Big\{ \big|\mathbf{a}_i\big|^2 - 2\langle \mathbf{v}_0, \mathbf{a}_i \rangle + \big|\mathbf{v}_0\big|^2 - \sum_{j=1}^{k} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2 \Big\} \\
&= \sum_{i=1}^{n} |\mathbf{a}_i|^2 - 2\Big\langle \mathbf{v}_0, \underbrace{\sum_{i=1}^{n} \mathbf{a}_i}_{=\mathbf{0}} \Big\rangle + n\big|\mathbf{v}_0\big|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2 \\
&= \sum_{i=1}^{n} |\mathbf{a}_i|^2 + n\big|\mathbf{v}_0\big|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2 \\
&\geq \sum_{i=1}^{n} |\mathbf{a}_i|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2 \\
&= \sum_{i=1}^{n} \mathrm{dist}(\mathbf{a}_i, \mathbb{A}')^2
\end{aligned}$$

where $\mathbb{A}' = \{\sum_{j=1}^{k} c_j \mathbf{v}_j : c_1, \ldots, c_k \in \mathbb{R}\}$ is a k-dimensional subspace, a special case of $\mathbb{A}$ with $\mathbf{v}_0 = \mathbf{0}$. Therefore, it must be $\mathbf{v}_0 = \mathbf{0}$, and the k-dimensional affine space best-fitted to the modified data points thus passes through the origin. Therefore, the k-dimensional affine space best-fitted to the original data points passes the centroid of the data points. ■

5.11 Connecting SVD and Principal Component Analysis (PCA)

A principal component of a random vector[3] is a linear combination of random variables which are components of the random vector. A principal component is, of course, a random variable and explains the variability of the random vector. Therefore, a better principal component explains more variability of the random vector. For a given random vector $\mathbf{X}$, a principal component is characterized by a $\mathbf{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, a deterministic coefficient vector of the linear combination, and then, can be represented as $\mathbf{v}^\top \mathbf{X} = \sum_{j=1}^{d} v_j X_j$. We adopt the variance of a random variable as the measure of its variability. Then, finding the best principal component is equivalent to finding $\mathbf{v}$ that maximizes the variance of $\mathbf{v}^\top \mathbf{X}$ among unit vectors, i.e., $|\mathbf{v}| = 1$. We call this best principal component or the first principal component.

Assume the mean of $\mathbf{X}$ is $\mathbf{0}$ without loss of generality.[4] Then, the mean of $\mathbf{v}^\top \mathbf{X}$ is zero, and its variance is

$$\begin{aligned} \mathbb{V}\mathsf{ar}(\mathbf{v}^\top \mathbf{X}) &= \mathbb{E}\big[(\mathbf{v}^\top \mathbf{X})^2\big] = \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}(\mathbf{v}^\top \mathbf{X})^\top\big] = \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{v}\big] = \mathbf{v}^\top \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]\mathbf{v} \\ &= \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}. \end{aligned}$$

- We can obtain the first principal component by solving

$$\operatorname*{argmax}_{|\mathbf{v}|=1} \mathbb{V}\mathsf{ar}(\mathbf{v}^\top \mathbf{X}) = \operatorname*{argmax}_{|\mathbf{v}|=1} \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}.$$

  Since the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]$ is not known in practice, but we are only given a set of iid[5] observations of $\mathbf{X}$, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we use a sample covariance $\hat{\boldsymbol{\Sigma}}$ estimated from these observations. Let A be the data matrix of which the i-th row corresponds to the i-th observation $\mathbf{x}_i$. Then, due to Fact 3.5, we can write an estimate of the sample covariance as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n-1} A^\top A \,.$$

  Since $\mathbf{v}^\top A^\top A \mathbf{v} = |A\mathbf{v}|^2$, $\mathbf{v}^\top \hat{\boldsymbol{\Sigma}} \mathbf{v} = \frac{1}{n-1} |A\mathbf{v}|^2$ holds, and we can rewrite the

[3] We explain random variables and vectors in Appendix D.

[4] This can be satisfied easily in practice by subtracting the sample mean from all the data points.

[5] Independent and identically distributed.

optimization problem to find the first principal component into

$$\operatorname*{argmax}_{|\mathbf{v}|=1} \mathbf{v}^\top \hat{\mathbf{\Sigma}} \mathbf{v} \;=\; \operatorname*{argmax}_{|\mathbf{v}|=1} |A\mathbf{v}|.$$

In other words, the first principal component coincides with the first right-singular vector $\mathbf{v}_1$ of the observation matrix A. Furthermore, the variance explained by the first principal component is proportional to the square of the first singular value, $\frac{\sigma_1^2}{n-1}$.

- We define the second principal component $\mathbf{v}_*$ as a principal component that is not correlated[6] with the first principal component $\mathbf{v}_1^\top \mathbf{X}$, and that maximizes the variance $\mathbb{V}\mathsf{ar}(\mathbf{v}^\top \mathbf{X})$. In statistical terminology, we are solving

$$\mathbf{v}_* = \operatorname*{argmax}_{\substack{|\mathbf{v}|=1,\\ \mathbb{C}\mathsf{ov}(\mathbf{v}^\top \mathbf{X}, \mathbf{v}_1^\top \mathbf{X})=0}} \mathbb{V}\mathsf{ar}(\mathbf{v}^\top \mathbf{X}).$$

The covariance constraint can be rewritten as

$$\begin{aligned}
\mathbb{C}\mathsf{ov}(\mathbf{v}^\top \mathbf{X}, \mathbf{v}_1^\top \mathbf{X}) &= \mathbb{E}\big[(\mathbf{v}^\top \mathbf{X})(\mathbf{v}_1^\top \mathbf{X})\big] - \underbrace{\mathbb{E}\big[\mathbf{v}^\top \mathbf{X}\big]}_{=\mathbf{0}} \underbrace{\mathbb{E}\big[\mathbf{v}_1^\top \mathbf{X}\big]}_{=0} \\
&= \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}(\mathbf{v}_1^\top \mathbf{X})^\top\big] = \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{v}_1\big] \\
&= \mathbf{v}^\top \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]\mathbf{v}_1 = \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}_1 \\
&= 0.
\end{aligned}$$

Using the sample covariance, we further rewrite it into

$$\mathbf{v}^\top \hat{\mathbf{\Sigma}} \mathbf{v}_1 = 0 \quad \Leftrightarrow \quad \mathbf{v}^\top A^\top A \mathbf{v}_1 = 0.$$

Since, due to Lemma 5.3, $A^\top A \mathbf{v}_1 = \sigma_1^2 \mathbf{v}_1$ for the right-singular vector $\mathbf{v}_1$, $\mathbf{v}^\top A^\top A \mathbf{v}_1 = 0$ is equivalent to $\mathbf{v}^\top \mathbf{v}_1 = 0$. In other words, the orthogonality between $\mathbf{v}$ and $\mathbf{v}_1$, is the necessary and sufficient condition for uncorrelatedness of $\mathbf{v}_1^\top \mathbf{X}$ and $\mathbf{v}^\top \mathbf{X}$. Therefore,

$$\mathbf{v}^\top \hat{\mathbf{\Sigma}} \mathbf{v}_1 = 0 \quad \Leftrightarrow \quad \mathbf{v}^\top \mathbf{v}_1 = 0,$$

and we can obtain the second principal component by solving

$$\mathbf{v}_* = \operatorname*{argmax}_{\substack{|\mathbf{v}|=1,\\ \mathbf{v}^\top \mathbf{v}_1=0}} |A\mathbf{v}|.$$

[6] For random variables X_1 and X_2, we say that they are uncorrelated if their covariance is zero, that is, $\mathbb{C}\mathsf{ov}(X_1, X_2) = 0$.

That is, the second right-singular vector $\mathbf{v}_2$ is the second principal component, and $\frac{\sigma_2^2}{n-1}$ is the variance explained by the second principal component.

- We can repeat this procedure for $\operatorname{rank} A = r$ times to compute the r principal components. We could find the $(r+1)$-th principal component, but this component explains nothing since the explained variance is zero.
- The total variance explained by $k \leq r$ principal components is $\frac{1}{n-1}(\sigma_1^2 + \cdots + \sigma_k^2)$, which is the maximum variance that can be explained by k right-singular vectors due to the best-fit character of SVD. Given a target proportion $0 < \alpha < 1$, we choose the smallest k that satisfies

$$\frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \cdots \sigma_k^2 + \sigma_{k+1}^2 \cdots + \sigma_r^2} \geq \alpha.$$

We then call $\mathbf{v}_1, \ldots, \mathbf{v}_k$ the principal components that explain $100\alpha\%$ of the total variance.

These results can be summarized into the following theorem in Statistics.

Theorem 5.7 *For a random vector* $\mathbf{X}$*, assume that* $\mathbb{E}[\mathbf{X}] = \mathbf{0}$*. Let* $\mathbb{C}\mathsf{ov}(\mathbf{X}, \mathbf{X}) = \mathbf{\Sigma}$ *have eigenvectors* $\mathbf{v}_1, \ldots, \mathbf{v}_r$ *with corresponding eigenvalues* $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$*. Then:*

(*i*) *The* j*-th PC (principal component) is* $\mathbf{v}_j^\top \mathbf{X} = v_{j,1} X_1 + \cdots + v_{j,d} X_d$ *for* $j = 1, \ldots, r$*.*

(*ii*) *The variance of* j*-th PC is* $\mathbb{V}\mathsf{ar}(\mathbf{v}_j^\top \mathbf{X}) = \mathbf{v}_j^\top \mathbf{\Sigma} \mathbf{v}_j = \sigma_j^2$*.*

(*iii*) *The covariance between two PCs is uncorrelated, i.e.* $\mathbb{C}\mathsf{ov}(\mathbf{v}_j^\top \mathbf{X}, \mathbf{v}_k^\top \mathbf{X}) = \mathbf{v}_j^\top \mathbf{\Sigma} \mathbf{v}_k = 0$ *for* $j \neq k$*.*

The proof of this theorem is straightforward if we notice that $\mathbf{\Sigma} = A^\top A$ where $A = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$.

Chapter 6

SVD in Practice

Singular value decomposition (SVD) is widely used in practice. In this chapter, we consider three practical use cases of SVD and its variant. First, SVD is used to compress a large matrix, representing an image, into two smaller matrices, without compromising its perceptual quality. This is done by taking only the top-K singular triplets, after performing SVD on the full matrix. Second, we show how left-singular vectors can be used to visualize high-dimensional data, for convenient analysis. As evident from our variational formulation of SVD, these left-singular vectors are the optimal representations of data in terms of reconstruction. We furthermore demonstrate how this approach of using left-singular vectors for visualization can be extended nonlinearly, with a variational autoencoder. Finally, we show how the right-singular vectors of financial time-series data automatically capture major underlying factors, by analyzing historical yield curves using SVD. These examples are only three out of increasingly many applications of SVD in data science and artificial intelligence.

6.1 Single-Image Compression via SVD

A representative example of using SVD in practice is image compression. An image is often represented as a collection of pixels on a 2-dimension grid, and each pixel is represented using three numbers corresponding to three color channels (r, g, b). An $n \times d$ image can thus be represented as a collection of three $n \times d$ matrices. Let $A^{(1)}, A^{(2)}, A^{(3)}$ be these three matrices, respectively, and r_1, r_2, r_3 be their respective ranks. We start by assuming that the column sums

of each of these matrices are all 0.

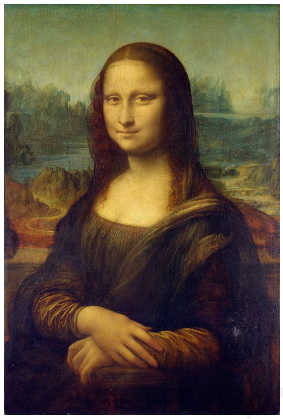


Figure 6.1: Mona Lisa in 1024×687 pixels.

SVD allows us to represent A as the sum of rank-one matrices:

$$A^{(j)} = \sum_{i=1}^{r_j} \sigma_i^{(j)} \mathbf{u}_i^{(j)} {\mathbf{v}_i^{(j)}}^\top,$$

with positive $\sigma_i^{(j)}$'s and vectors $\mathbf{u}_i^{(j)}$'s and $\mathbf{v}_i^{(j)}$'s. If $k \leq \min\{r_1, r_2, r_3\}$, rank k approximation of $A^{(j)}$ is

$${A^{(j)}}_k = \sum_{i=1}^{k} \sigma_i^{(j)} \mathbf{u}_i^{(j)} {\mathbf{v}_i^{(j)}}^\top.$$

We show in Figure 6.2 five approximated images of the original Mona Lisa in Figure 6.1 with the ranks $k = 3, 8, 18, 23$ and 34, respectively. Even when we use only 8.3% of the original pixels (see Figure 6.2f), it is difficult to discern this approximated (compressed) version from the original.

When the column means are not zeros, we simply subtract the column mean from each column, perform SVD, compute low-rank approximation and add back the column mean to each column. In other words, we perform SVD on

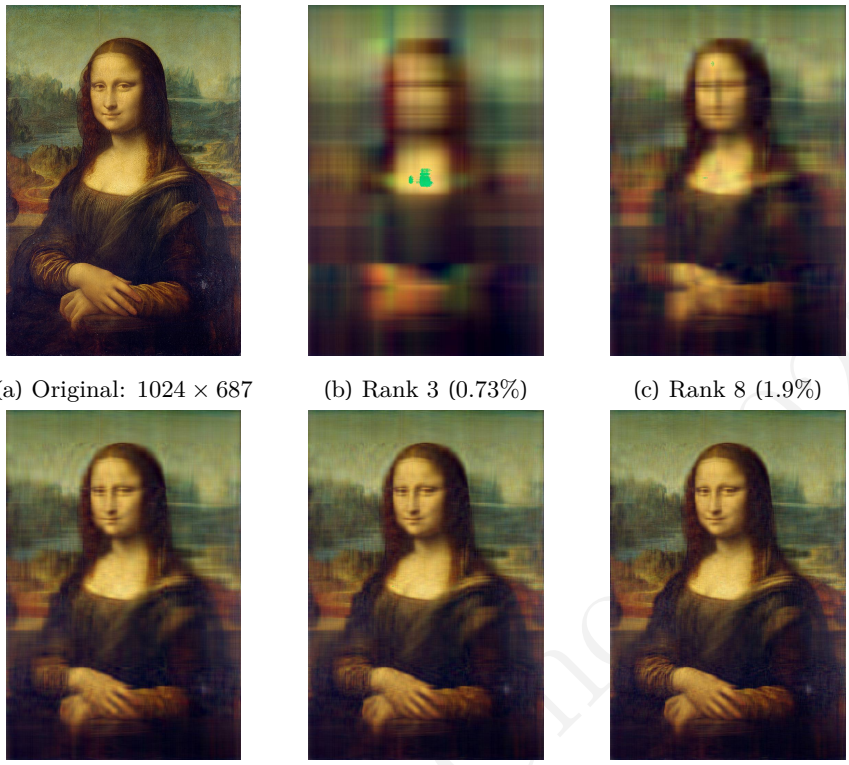
$$A^{(j)} - \mathbf{1}_n \boldsymbol{\mu}_j^\top,$$

where

$$\boldsymbol{\mu}_j = \frac{1}{n} \mathbf{1}_n^\top A^{(j)}.$$

This column average is added back to low-rank approximation;

$$A^{(j)} \approx \mathbf{1}_n \boldsymbol{\mu}_j^\top + {A^{(j)}}_k = \mathbf{1}_n \boldsymbol{\mu}_j^\top + \sum_{i=1}^{k} \sigma_i^{(j)} \mathbf{u}_i^{(j)} {\mathbf{v}_i^{(j)}}^\top,$$



(a) Original: 1024×687 (b) Rank 3 (0.73%) (c) Rank 8 (1.9%)

(d) Rank 18 (4.4%) (e) Rank 23 (5.6%) (f) Rank 34 (8.3%)

Figure 6.2: Low-rank approximations of the Mona Lisa (The percentage is the portion of data in use.)

where low-rank approximation was done on

$$A^{(j)} - \mathbf{1}_n \boldsymbol{\mu}_j^\top \approx {A^{(j)}}_k = \sum_{i=1}^{k} \sigma_i^{(j)} \mathbf{u}_i^{(j)} {\mathbf{v}_i^{(j)}}^\top.$$

Refer to Section 5.10 to see why we need to subtract and add back the column means.

6.1.1 Singular values reveal the amount of information in low rank approximation

We can quantify how well the original image is represented by the compressed image by

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{687} \sigma_i^2},$$

using the k-largest singular values. We can plot this proportion for each color channel over k, to visually inspect and determine the right balance between the compression and fidelity.

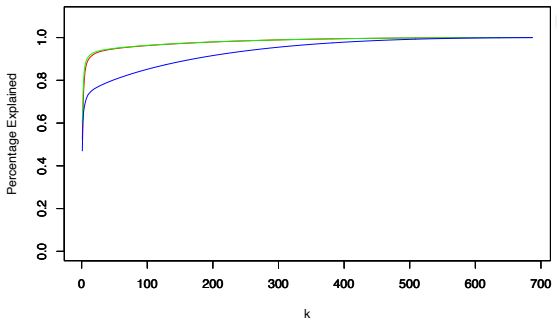


Figure 6.3: Ratio explained by singular triplets

From Figure 6.3, we observe that red and green colors are more easily captured with a fewer singular triplets, compared to blue. Even when the ratio of blue was less than 0.8 with $k = 34$, we were not able to visually discern a compressed image from the original image. This may be explained by the fact that the proportion of cone cells, which are photoreceptor cells in the retinas, that respond to blue (known to be only about 2%) is much lower than those of cone cells responding to red and green.

6.2 Visualizing High-Dimensional Data via SVD

6.2.1 Left-singular Vectors as the Coordinates of Embedding Vectors in the Latent Space

We are interested in finding a k-dimensional subspace that approximates well n data points in a d-dimensional vector space, $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subset \mathbb{R}^d$. We can formulate a problem of identifying a k-dimensional subspace as a problem of

finding k d-dimensional basic vectors that form a basis of the subspace. In other words, we can phrase this problem as estimating an unknown $d \times k$ matrix V that consists of these basic vectors as its columns to maximize $\sum_{i=1}^{n} |V^\top \mathbf{a}_i|^2$. That is,

$$V^* = \underset{\substack{V:\, d\times k \\ V^\top V = I_k}}{\operatorname{argmax}} \sum_{i=1}^{n} \left|V^\top \mathbf{a}_i\right|^2. \tag{6.1}$$

Using the orthonormality of columns, $V^\top V = I_k$, we see that

$$\begin{aligned}
\left|\mathbf{a}_i - VV^\top \mathbf{a}_i\right|^2 &= \left(\mathbf{a}_i - VV^\top \mathbf{a}_i\right)^\top \left(\mathbf{a}_i - VV^\top \mathbf{a}_i\right) \\
&= \left(\mathbf{a}_i^\top - \mathbf{a}_i^\top VV^\top\right)\left(\mathbf{a}_i - VV^\top \mathbf{a}_i\right) \\
&= \mathbf{a}_i^\top \mathbf{a}_i - 2\mathbf{a}_i^\top VV^\top \mathbf{a}_i + \mathbf{a}_i VV^\top VV^\top \mathbf{a}_i \\
&= |\mathbf{a}_i|^2 - \mathbf{a}_i^\top VV^\top \mathbf{a}_i \\
&= |\mathbf{a}_i|^2 - \left|V^\top \mathbf{a}_i\right|^2.
\end{aligned}$$

Because $\sum_{i=1}^{n} |\mathbf{a}_i|^2$ is a constant with respect to V, finding V that maximizes $\sum_{i=1}^{n} |V^\top \mathbf{a}_i|^2$ is equivalent to finding V that minimizes $\sum_{i=1}^{n} |\mathbf{a}_i - VV^\top \mathbf{a}_i|^2$. In other words, the original problem (6.1) can be rephrased as

$$V^* = \underset{\substack{V:\, d\times k \\ V^\top V = I_k}}{\operatorname{argmax}} \sum_{i=1}^{n} |\mathbf{a}_i - VV^\top \mathbf{a}_i|^2. \tag{6.2}$$

Under this formulation, you can view $VV^\top \mathbf{a}_i$ as reconstructing the original vector $\mathbf{a}_i$ back from the k-dimensional compression $V^\top \mathbf{a}_i$ by multiplying it with V. Then, this problem can be thought of as minimizing the reconstruction error.[1][2] We often refer to this k-dimensional subspace in which we transformed and embedded the data points as a latent space.

Let A be an $n \times d$ data matrix with $\mathbf{a}_i$'s as its rows and V be this unknown $d \times k$ matrix with $\mathbf{v}_j$'s as its columns. Then, the subspace spanned by the column vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, obtained as a solution to (6.1), has the same effect as dimension reduction as the subspace spanned by the singular vectors of A that correspond to the k largest singular values, in terms of the squared residual

[1] This is a special case of an autoencoder in machine learning, where $V^\top$ is an encoder and V is a decoder. That is, this corresponds to a linear autoencoder with a tied weight.

[2] It is natural to consider $VV^\top$ as the matrix form of projection onto a subspace spanned by orthonormal vectors, as in (4.8). It is however non-trivial to go from projection to reconstruction via $V^\top$.

distance, because

$$\sum_{i=1}^{n} \left|V^\top \mathbf{a}_i\right|^2 = \|AV\|_F^2 = \sum_{j=1}^{k} |A\mathbf{v}_j|^2.$$

The column vectors of V however may not be ordered according to the singular values. That is, $|A\mathbf{v}_i| \geq |A\mathbf{v}_{i+1}|$ may not hold for some i.

Let $\mathbf{b}_i$ be the image of $\mathbf{a}_i$ in the latent space. It is usual for us to call $\mathbf{b}_i$ the embedding vector of $\mathbf{a}_i$. The i-th element of the j-th left-singular vector of A, $(\mathbf{u}_j)_i$, is a scalar multiple of the j-th element of the embedding vector $\mathbf{b}_i$, because $\sigma_j(\mathbf{u}_j)_i = (A\mathbf{v}_j)_i = \mathbf{a}_i^\top \mathbf{v}_j = (V^\top \mathbf{a}_i)_j = (\mathbf{b}_i)_j$. In other words, the j-th left-singular vector multiplied by its singular value $\sigma_j \mathbf{u}_j$, which is n-dimensional, is a collection of the j-th coordinate values of the n data points in the latent space. That is, the i-th row of $U\Sigma = [\,\sigma_1\mathbf{u}_1|\sigma_2\mathbf{u}_2|\ldots|\,\sigma_k\mathbf{u}_k\,]$ is $\mathbf{b}_i$. With an appropriate choice of k, we can gain insights into the data point $\mathbf{a}_i$ by analyzing its embedding vector $\mathbf{b}_i$.

6.2.2 Geometry of MNIST Images According to SVD

The MNIST Dataset consists of handwritten digits and has been used extensively to train and evaluate various image processing systems as well as machine learning algorithms. It contains 60,000 training images and 10,000 test images of handwritten digits (0-9). See Figure 6.4 for 160 randomly selected images from MNIST.

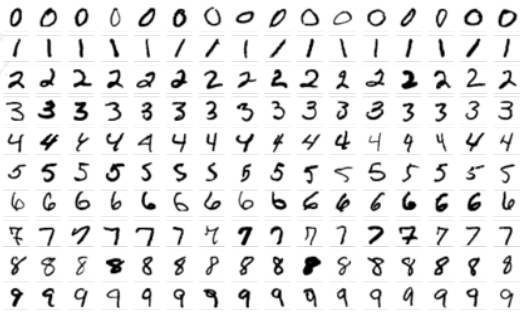
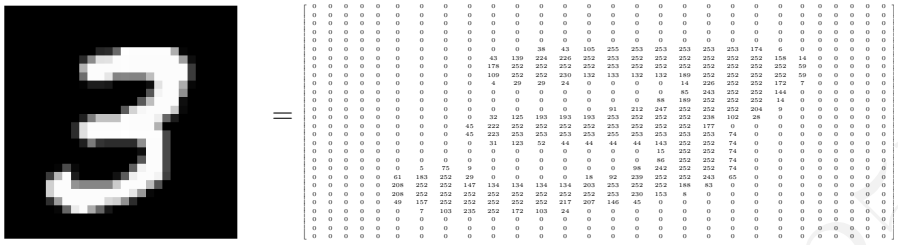


Figure 6.4: Examples of images in MNIST data set

Each handwritten digit in MNIST is represented as a 28×28 grayscale image. Each pixel can be one of 256 intensity levels (0-255). In Figure 6.5, we demonstrate how the 8-th training image from MNIST, that corresponds

to a handwritten three, can be plotted in two different ways; one as an actual grayscale image and the other as a 28×28 matrix.



(a) An image of 3 (b) 28×28 matrix representation of the left image

Figure 6.5: An Example of MNIST Image and its Matrix Representation

When analyzing such a dataset, it is often more convenient to analyze it as a collection of vectors rather than as images, although the latter tend to be more familiar to us. In the case of MNIST, we can reshape the matrix of each handwritten digit into a 784-dimensional vector. Once numerical analysis is completed, we can visualize these data points as well as intermediate quantities as images rather than vectors.

Once we create a $70,000 \times 784$ data matrix A of MNIST by vertically stacking 784-dimensional row vectors, we first compute the mean of the rows (column means) by $\frac{1}{n}\mathbf{1}_n^\top A$ where $n = 70,000$, which we visualize in Figure 6.6. We can check other properties of this data matrix, such as its rank by using `numpy` package of Python, which results in rank $A = 713$.

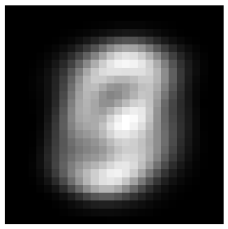


Figure 6.6: The Mean Image of MNIST

We now subtract this mean from each row of the data matrix A by

$$A - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top A,$$

after which we perform SVD on A. We first visualize the top-64 right-singular vectors, according to their associated singular values, in Figure 6.7. Although it is not easy to interpret these right-singular vectors intuitively, we notice that the spatial frequency increases as the singular values decrease. This is evident from the increasingly more frequent flips between black and white contiguous regions.

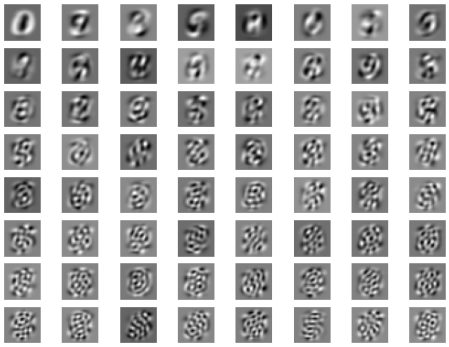


Figure 6.7: 64 Leading right-singular Vectors of A

These right-singular vectors form a basis of a subspace. In Figure 6.8, we plot all rows of the data matrix on the subspace spanned by the top-2 leading singular values, by putting each row using the first two elements of the corresponding left-singular vector. We can visually confirm that digits of a similar shape are placed nearby and sometimes even overlap with each other. As could have been guessed from the first right-singular vector, which depicted a 0-like shape, zeros are clustered in a region with large x.

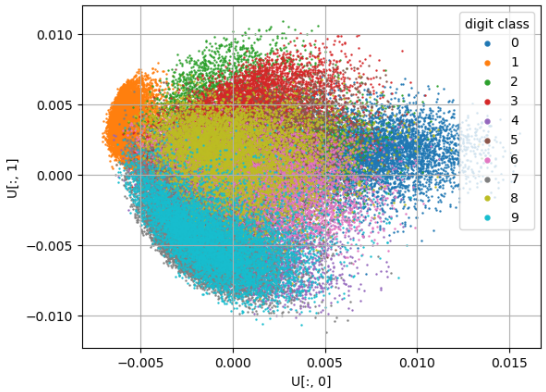


Figure 6.8: Two Leading left-singular Vectors of A

Next, we visually inspect the quality of low-rank approximation, while varying the number of the right-singular vectors; 2, 22, 42, 62 and 82. For each handwritten digit, we plot its reconstructed versions from the corresponding low-dimensional subspaces.

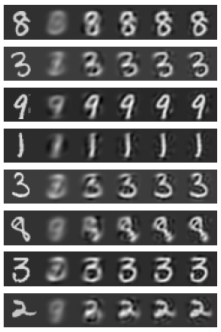


Figure 6.9: More Singular Vectors, More Accurate Approximation

The first column in Figure 6.9 is the original image, and the subsequent columns correspond to the reconstructed images based on small numbers of right-singular vectors. Already with 42 right-singular vectors alone, out of 700 or so, reconstructed images are almost as good as the original ones. We can quantify the quality of approximation with $\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{784} \sigma_i^2}$, as in the following Figure 6.10.

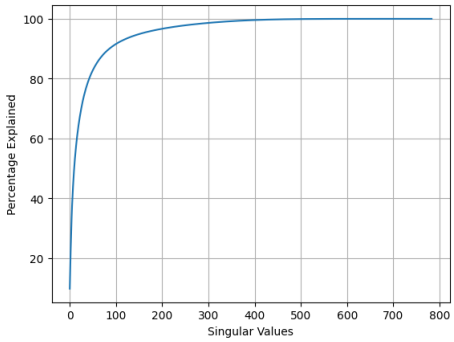


Figure 6.10: Ratio explained by singular triplets

6.2.3 Geometry of MNIST Images in the Latent Space of a Variational Auto-Encoder

Handwritten digits in MNIST still have significant overlaps across digit classes, when we visualize them using SVD, as shown in Figure 6.8. This clutters our analysis effort, but is also difficult to overcome due to the linearity of SVD. It is thus a common practice to project data nonlinearly onto a lower-dimensional subspace for analyzing data more in-depth. A variational autoencoder (VAE) is one representative example of such an approach. A VAE consists of two neural networks, described earlier in Section 3.10.1, called the encoder and decoder. The output layer of the encoder has d nodes, and the output from the encoder is fed to the decoder as the input. The decoder outputs a 784-dimensional vector, corresponding to the dimensionality of the original MNIST image. The VAE is trained to minimize the reconstruction error while being regularized to prevent overfitting. The space in which the output from the encoder resides is often called a latent space, and the output from the encoder a latent representation. With $d = 2$, we can visualize the MNIST images readily, as in Figure 6.11.

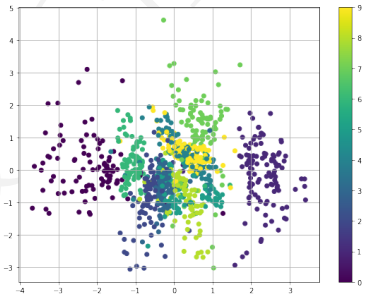


Figure 6.11: Latent space of VAE

6.3 Approximation of Financial Time-Series via SVD

In finance, the yield to maturity of a fixed-interest security (typically, bonds) is an estimate of the rate of total return to be earned by its owner who buys it at a market price, holds it to maturity, and receives all interest payments and the capital redemption on schedule. The yield curve is a graph which depicts how the yields to maturity vary as a function of their years remaining to

maturity. The graph's horizontal axis is a time line of months or years remaining to maturity. The vertical axis depicts the annualized yield to maturity. As a demonstration, Figure 6.12 shows 416 yield curves observed weekly from 2010 to 2017. These curves linearly connect yields corresponding to 11 remaining maturities – 3 months, 6 months, 9 months, one year, one and a half year, two years, two and a half years, 3 years, 5 years, 10 year, and 20 years.

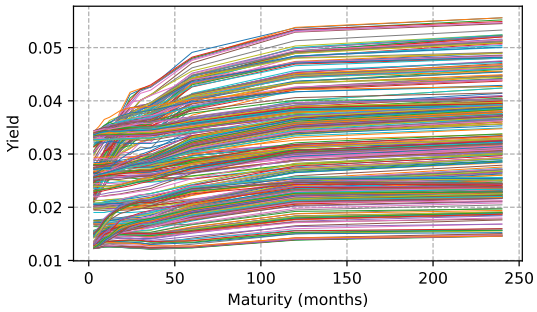


Figure 6.12: Yield curves observed weekly from 2010 to 2017.

To see whether any underlying patterns exist in the yield curves, we apply SVD to a matrix whose rows represent the curves. That is, let us regard the 11 yields observed at each week as a row of a matrix and call the matrix A. Note that A is a 416×11 matrix. By subtracting the column sum of A, $\boldsymbol{\mu} = \frac{1}{416}\mathbf{1}_{416}{}^{\top} A \in \mathbb{R}^{11}$, from each row, we get

$$\hat{A} = A - \mathbf{1}_{416}\,\boldsymbol{\mu}^{\top}$$

whose columns sums vanish. Denote the SVD of $\hat{A}$ as

$$\hat{A} = UDV^{\top} = \sum_{k=1}^{11} \sigma_k \mathbf{u}_k \mathbf{v}_k$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{11}$ and $|\mathbf{v}_k| = 1$, $k = 1, \ldots, 11$. The i-th row of $\hat{A}$ corresponding to the curve at i-the week is

$$\sum_{k=1}^{11} u_{ik}\sigma_k \mathbf{v}_k$$

where u_{ik} is the i-th entry of $\mathbf{u}_k$. For $\hat{A}$, the first right-singular vector explains 93.74% of the total variation and the top three right-singular vectors explain

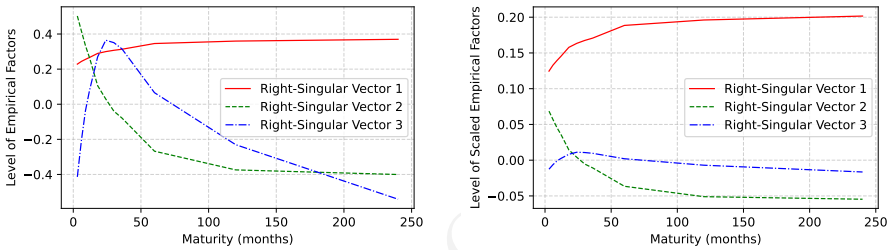
99.9% of the total variation. The ratios $100 \times \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{11} \sigma_i^2}$ for $k = 1, \ldots, 11$ are

$$93.74,\ 99.61,\ 99.90,\ 99.94,\ 99.97,\ 99.98,\ 99.98,\ 99.99,\ 99.99,\ 99.99,\ 100.00\,,$$

respectively. Since the third ratio is close to 100, the truncated sum of three leading terms approximates the i-th row of $\hat{A}$ very well. That is, for all $i = 1, \ldots, 416$,

$$\sum_{k=1}^{11} u_{ik}\sigma_k \mathbf{v}_k \approx u_{i1}\sigma_1\mathbf{v}_1 + u_{i2}\sigma_2\mathbf{v}_2 + u_{i3}\sigma_3\mathbf{v}_3\,. \tag{6.3}$$

Let us look at these right-singular vectors.



(a) right-singular vectors, $\mathbf{v}_k$ (b) Scaled right-singular vectors, $\sigma_k\mathbf{v}_k$

Figure 6.13: Right-singular vectors of leading three singular values

Three right-singular vectors having the three largest singular values are depicted in the left panel of Figure 6.13a. The right panel shows the right-singular vectors multiplied by their singular values. The red curve is the first right-singular vector in Figure 6.14. It determines the overall levels of yields and is called a *level factor* in econometrics. The green one affects initial slopes of curves and is called a *slope factor*. The blues one represents the curvature of yield curves near at maturity of 4 years and is called a *curvature factor*. This empirical findings through SVD accompanies an analytical modeling of the three factor curves as

$$1, \quad \frac{1 - e^{-\lambda\tau}}{\lambda\tau}, \quad \frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}$$

whose graphs are depicted in Figure 6.14.

The estimation of the parameter λ is crucial in empirical finance. A natural next step would be reconstructing the yield curves by these three analytic curves. Denote the yield at t of a bond with its maturity date τ as $y_t(\tau)$. With well-chosen L_t, S_t, C_t and λ,

$$Y_t(\tau) = L_t \times 1 + S_t \times \frac{1 - e^{-\lambda\tau}}{\lambda\tau} + C_t \times \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{6.4}$$

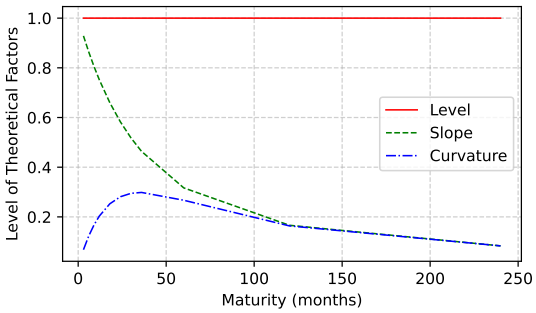


Figure 6.14: Three factor curves of DNS for $\lambda = 0.05$

may approximate $y_t(\tau)$. This idea is called (*Dynamic*) *Nelson-Siegel* method in econometrics. If λ vary through t, the method is called *Nelson-Siegel* method. If λ is fixed for all t, it is called *Dynamic Nelson-Siegel* (*DNS*) method. If we compare two representations (6.3) and (6.4), we may regard the three right-singular vectors in (6.3) as proxies of the three analytic factor functions in (6.4). Then, the leading three left-singular vectors correspond to the sequence of (L_t, S_t, C_t) in (6.4). $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ is depicted in Figure 6.15 as a sequence of $\mathbb{R}^3$-vectors. $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ correspond to red, green, and blue curves, respectively. As a finding from the curves, the red curve representing overall level of yields explains that the yields from bonds decreases through the period.[3] Further empirical findings can be observed in financial view point, but we do not dive into.

Let us compare the original yield curves and the approximated ones in Figure 6.16.

They seem similar, but the approximated curves in the right panel lose some curved features of original yields requiring more right-singular vectors to capture. However, insights from the simple expression might be more important for financial decisions than keeping local details of yield curves.

[3]Most of yields are explained by $\sigma_1 \mathbf{v}_1$ whose entries are all positive. The yields are approximated as

$$\boldsymbol{\mu} + u_{t1}\sigma_1\mathbf{v}_1 + u_{t2}\sigma_2\mathbf{v}_2 + u_{t3}\sigma_3\mathbf{v}_3$$

and usually stay at positive regime even if the coefficient u_{t1} is negative because of the first mean vector term $\boldsymbol{\mu}$. u_{t1} has negative values in the second half of the period as we can see in Figure 6.15.

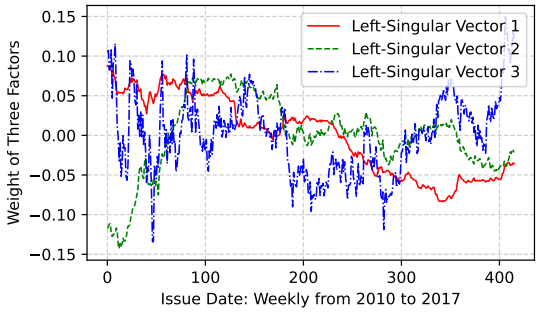
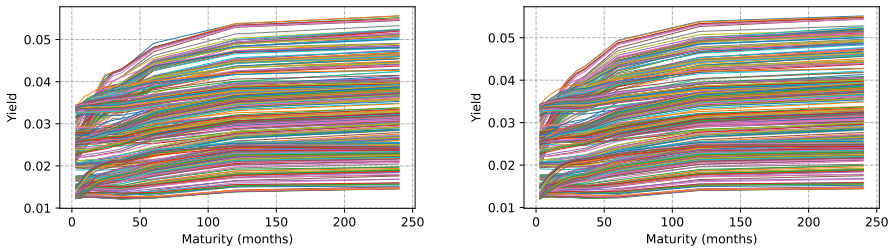


Figure 6.15: $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ plot



(a) Original yields curves (b) Yield curves represented in 3 factors

Figure 6.16: Illustration of dimension reduction in yield curve representation

Chapter 7

Positive Definite Matrices

Positive definite matrices are everywhere around us. A positive definite matrix called a covariance captures the relationship among scattered observations of multiple random variables. Many problems in engineering and data science are formulated as parameter optimizations. When the second-order derivative of such an optimization problem is positive definite, we can often find an efficient and stable procedure to solve the problem. Positive definite matrices furthermore have an elegant connection to a convexity through high-dimensional ellipsoids. We can view positive definite matrices similarly to real numbers from the arithmetic perspective. They are both additive and multiplicative and result in a positive value when inserted in a quadratic form for any non-zero vector and real number, respectively. Furthermore, the square root of a positive definite matrix is unique and positive definite, just like the square root of a positive real number is unique and positive.

7.1 Positive (Semi-)Definite Matrices

Positive (semi-)definiteness is widely used to describe the positiveness of quadratic forms induced by special matrices, such as the covariance matrix[1] of a random vector and the Hessian matrix of a convex[2] multivariate function. Furthermore, there is a beautiful one-to-one correspondence between a positive definite matrix and an inner product of a vector space (see Theorem 4.1.) We already

[1] Refer Appendix D.
[2] Refer Appendix A.

introduced positive definiteness in Definition 4.2. Similarly, we define a positive semi-definite matrix by relaxing the positiveness by a non-negativity.

Definition 7.1 *A square matrix A is positive semi-definite if $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all $\mathbf{x}$.*

The following conditions are equivalent to positive semi-definiteness.

Fact 7.1 *For a symmetric matrix A, the followings are equivalent:*

1. *For all $\mathbf{x}$, $\mathbf{x}^\top A\mathbf{x} \geq 0$;*
2. *All eigenvalues of A are nonnegative;*
3. *$A = B^\top B$ for some matrix B.*

Proof: According to the real spectral theorem, a symmetric matrix A can be expressed as

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top = V\Lambda V^\top,$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with eigenvalues of A as its diagonal entries, and $V = [\,\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\,]$ is an orthogonal matrix whose columns consist of orthonormal eigenvectors of A.

- **(1) $\Rightarrow$ (2)**: for all j, $\lambda_j = \mathbf{v}_j^\top A\mathbf{v}_j \geq 0$.
- **(2) $\Rightarrow$ (3)**: Because all diagonal entries of Λ are greater than or equal to 0, $D = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ is well-defined such that $\Lambda = D^2$ and $D^\top = D$. Thus, $A = V\Lambda V^\top = VDD^\top V^\top = B^\top B$.
- **(3) $\Rightarrow$ (1)**: $\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top B^\top B\mathbf{x} = |B\mathbf{x}|^2 \geq 0$.

■

We can derive a similar set of equivalences for a positive definite matrix as well.

Fact 7.2 *For a $d \times d$ symmetric matrix A, the followings are equivalent:*

1. *For all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^\top A\mathbf{x} > 0$;*
2. *All eigenvalues are positive;*

3. $A = B^\top B$ *for some invertible matrix* B.

4. $A = B^\top B$ *for some matrix* B *of rank* d.

Proof: As in the proof of Fact 7.1, we have $A = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top = V\Lambda V^\top$ in hand due to the spectral decomposition theorem.

- **(1)** $\Rightarrow$ **(2)**: for all j, $\lambda_j = \mathbf{v}_j^\top A \mathbf{v}_j > 0$ since $\mathbf{v}_j \neq \mathbf{0}$;
- **(2)** $\Rightarrow$ **(3)**: Since diagonal entries of Λ are positive eigenvalues, $D = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ is well-defined such that $\Lambda = D^2$ and $D^\top = D$. Therefore, $A = V\Lambda V^\top = VDD^\top V^\top = B^\top B$ where B consists of linearly independent rows since they are positively scaled columns of the orthogonal matrix V. Therefore, B is invertible;
- **(3)** $\Rightarrow$ **(4)**: An invertible matrix has full-column rank and $\mathrm{rank}\, B = d$.
- **(4)** $\Rightarrow$ **(1)**: $\mathbf{x}^\top A \mathbf{x} = \mathbf{x}^\top B^\top B \mathbf{x} = |B\mathbf{x}|^2 = 0$ if and only if $B\mathbf{x} = \mathbf{0}$. Since B has d columns and $\mathrm{rank}\, B = d$, $\dim \mathrm{Null}\, B = d - \mathrm{rank}\, B = 0$. Therefore, $B\mathbf{x} \neq \mathbf{0}$ if $\mathbf{x} \neq \mathbf{0}$.

■

We can further derive the following result from Fact 7.2.

Fact 7.3 *If* A *is symmetric and positive definite, then* A *is invertible, and* A^{-1} *is also positive definite.*

Proof: According to Fact 7.2, any positive definite matrix can be expressed as $B^\top B$ for some invertible matrix B. Then, the product itself is also invertible. Furthermore, its inverse $B^{-1}\left(B^{-1}\right)^\top$ is again a positive definite matrix by Fact 7.2. ■

We then can list up some necessary conditions that must be satisfied by a positive definite matrix.

Example 7.1 For a positive (semi-)definite matrix $A = (a_{ij})$,

1. all diagonal entries are positive, i.e., $a_{ii} > (\geq) 0$ for all i since $\mathbf{x} = \mathbf{e}_i$ leads to $\mathbf{x}^\top A \mathbf{x} = a_{ii}$ which has to be greater than (or equal to) 0.
2. leading $k \times k$ block $\left(a_{ij}\right)_{\substack{1 \leq i \leq k \\ 1 \leq j \leq k}}$ of A is also positive (semi-)definite for all k. We can see it by considering $\mathbf{x}^\top A \mathbf{x}$ after setting $\mathbf{x} = (x_1, \ldots, x_k, 0, \ldots, 0)^\top$.

Since $\mathbf{x}^\top A\mathbf{x}$ is a scalar,

$$\mathbf{x}^\top A\mathbf{x} = \frac{\mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top A^\top \mathbf{x}}{2} = \frac{1}{2}\mathbf{x}^\top (A + A^\top)\mathbf{x}$$

holds for any square matrix A with any vector $\mathbf{x}$. So, the positive (semi-)definiteness of A is equivalent to the positive (semi-)definiteness of the associated symmetric matrix $A + A^\top$. For this reason, we study the positive definiteness through only symmetric matrices.

7.2 Cholesky Factorization of Positive Definite Matrix

A positive definite matrix A can be expressed as $A = B^\top B$ with an invertible matrix B, according to Fact 7.2. We can further decompose B using QR decomposition as the product of an orthogonal matrix Q and an upper triangular matrix R with positive diagonals, i.e., $B = QR$. Since $Q^\top Q = I$,

$$A = B^\top B = (QR)^\top QR = R^\top Q^\top QR = R^\top R.$$

We call this Cholesky decomposition.

Fact 7.4 (Cholesky decomposition) *For a positive definite matrix A, there exists a unique upper triangular matrix R with positive diagonal entries such that $A = R^\top R$.*

Proof: Since we showed its existence already, we show the uniqueness here. Assume there exist two upper triangular matrices, $R = (r_{ij})$ and $S = (s_{ij})$, that satisfy $A = R^\top R = S^\top S$. Because S and R are both invertible,

$$(S^{-1})^\top R^\top = SR^{-1}.$$

The inverse of an upper triangular matrix is upper triangular by Theorem 2.1, and the product of upper(lower)-triangular matrices is an upper(lower)-triangular matrix by Fact 2.1. Thus, the left-hand side of this equation is a lower triangular matrix, and the right-hand side an upper triangular matrix. In other words, SR^{-1} is a diagonal matrix. For all i,

$$\frac{r_{ii}}{s_{ii}} = \frac{s_{ii}}{r_{ii}},$$

because $(S^{-1})^\top R^\top = \mathrm{diag}(r_{ii}/s_{ii})$ and $SR^{-1} = \mathrm{diag}(s_{ii}/r_{ii})$. Since the diagonal entries of both R and S are positive, $s_{ii} = r_{ii}$, which results in $SR^{-1} = I$. Therefore, $S = R$.

■

An Algorithm for Cholesky Decomposition

Here, we describe how to find $R = (r_{ij})$ such that $A = R^\top R$. $r_{11} = \sqrt{a_{11}}$ from $a_{11} = r_{11}^2 > 0$. We notice that the first row of $R^\top R$ consists of $r_{11}r_{1j}$'s. Combining these two, we get $a_{1j} = r_{11}r_{1j} = \sqrt{a_{11}}\, r_{1j}$ from which we can determine r_{1j}. Once we know r_{12}, we can also determine r_{22}, because $0 < a_{22} = r_{12}^2 + r_{22}^2$. The rest of the second row can be determined from $a_{2j} = r_{12}r_{1j} + r_{22}r_{2j}$, because we already know r_{12}, r_{1j} and r_{22}. We can repeat this procedure to determine all the entries of the rest of the rows, and based on the uniqueness of Cholesky decomposition, we know that the resulting matrix is the desired one.

7.3 Square Root of Positive Semi-definite Matrix

For any positive real number a, there exists a unique positive real b that satisfies $a = b^2$. An analogy holds for a positive definite matrix. For any positive definite matrix A, there exists a positive definite matrix B that satisfies $A = B^2$. Without the symmetry of B, this property is different from Fact 7.2, although two statements are quite similar.

Let A be a symmetric positive (semi-)definite matrix. We can use real spectral decomposition to express it as

$$A = V\Lambda V^\top,$$

where V is an orthogonal matrix and Λ is a non-negative diagonal matrix. If we let $D = \mathrm{diag}(\sqrt{\lambda_i})$, D is also positive-definite or positive semi-definite, identically to Λ, and $\Lambda = D^2$. If we let $B = VDV^\top$,

$$A = VD^2V^\top = VDV^\top VDV^\top = BB = B^2,$$

because $V^\top V = I$. A and B share the same positive definiteness.

Fact 7.5 *For a positive (semi-)definite matrix A, there exists a unique positive (semi-)definite matrix B such that $A = B^2 = B^\top B$. We denote this B as $A^{\frac{1}{2}}$.*

Proof: Since we already showed the existence above, we show the uniqueness here. Using the projection form of real spectral decomposition (5.14), we write A as $A = \sum_{i=1}^{r} \lambda_i \mathbf{P}_{A,\lambda_i}$ and a symmetric matrix B as $B = \sum_{j=1}^{s} \mu_j \mathbf{P}_{B,\mu_j}$. Because it is orthogonal projection, ${\mathbf{P}_{B,\mu_j}}^2 = \mathbf{P}_{B,\mu_j}$, ${\mathbf{P}_{B,\mu_j}}^\top = \mathbf{P}_{B,\mu_j}$, and $\mathbf{P}_{B,\mu_j}\mathbf{P}_{B,\mu_k} = \mathbf{0}$ for $j \neq k$. Thus, we get $B^2 = \left(\sum_{j=1}^{s} \mu_j \mathbf{P}_{B,\mu_j}\right)^2 = \sum_{j=1}^{s} \mu_j^2 \mathbf{P}_{B,\mu_j}$. For $A = B^2$ to hold, $r = s$ must hold, and for some i, $\mu_j = \sqrt{\lambda_i}$ and $\mathbf{P}_{B,\mu_j} = \mathbf{P}_{A,\lambda_i}$. Therefore, B must be expressed as

$$B = \sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{P}_{A,\lambda_i},$$

which is unique.

■

7.4 Variational Characterization of Symmetric Eigenvalues

An $n \times n$ real symmetric matrix A has n real eigenvalues. We sort these eigenvalues as follows:

$$\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) = \lambda_{\min}(A)\,.$$

We can relate these eigenvalues to the maximum and minimum values of the *Rayleigh quotient* defined as

$$\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}\,, \quad \mathbf{x} \neq \mathbf{0}.$$

For any $\mathbf{x} \neq \mathbf{0}$, we get

$$\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \mathbf{y}^\top A\mathbf{y}\,, \quad |\mathbf{y}| = 1$$

by setting $\mathbf{y} = \frac{\mathbf{x}}{|\mathbf{x}|}$, meaning that we may consider a quadratic form over unit vectors instead to investigate the Rayleigh quotient over all non-zero vectors.

The Rayleigh quotients over a subspace spanned by a subset of eigenvectors are bounded by the minimum and maximum corresponding eigenvalues. This result will be used occasionally throughout the book.

Lemma 7.1 *Consider an $n \times n$ symmetric matrix A and its real spectral decomposition $A = \sum_{i=1}^{n} \lambda_i(A)\mathbf{v}_i\mathbf{v}_i^\top$. Let $1 \leq p \leq q \leq n$. Then, for any non-zero vector $\mathbf{x} \in \mathrm{span}\{\mathbf{v}_p, \mathbf{v}_{p+1}, \ldots, \mathbf{v}_q\}$,*

$$\lambda_q(A) \;\leq\; \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \;\leq\; \lambda_p(A)\,.$$

The upper and lower bounds are achieved by $\mathbf{v}_p$ *and* $\mathbf{v}_q$*, respectively.*

Proof: With $\mathbf{x} = \sum_{i=p}^{q} x_i \mathbf{v}_i$ and the orthonormality of $\mathbf{v}_i$'s,

$$\mathbf{x}^\top A\mathbf{x} = \Big(\sum_{i=p}^{q} x_i \mathbf{v}_i^\top\Big)\Big(\sum_{j=1}^{n} \lambda_j(A)\mathbf{v}_j\mathbf{v}_j^\top\Big)\Big(\sum_{k=p}^{q} x_k \mathbf{v}_k\Big) = \sum_{i=p}^{q} \lambda_i(A) x_i^2.$$

Because $\lambda_p(A) \geq \lambda_i(A) \geq \lambda_q(A)$ for $p \leq i \leq q$, it is easy to see the following inequalities

$$\lambda_q(A)\sum_{i=p}^{q} x_i^2 \leq \sum_{i=p}^{q} \lambda_i(A) x_i^2 \leq \lambda_p(A)\sum_{i=p}^{q} x_i^2,$$

which can be re-written as, by using $\mathbf{x}^\top\mathbf{x} = \sum_{i=p}^{q} x_i^2$,

$$\lambda_q(A)\mathbf{x}^\top\mathbf{x} \leq \mathbf{x}^\top A\mathbf{x} \leq \lambda_p(A)\mathbf{x}^\top\mathbf{x}.$$

In addition, $\lambda_p(A) = \mathbf{v}_p^\top A\mathbf{v}_p$ and $\lambda_q(A) = \mathbf{v}_q^\top A\mathbf{v}_q$ hold as well as $|\mathbf{v}_p| = |\mathbf{v}_q| = 1$. ■

As a special case of this lemma, we get the following result.

Theorem 7.1 (Rayleigh quotients) *For any* $n \times n$ *symmetric matrix* A*,*

$$\lambda_{\min}(A) \leq \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_{\max}(A) \qquad \textit{for all } \mathbf{x} \neq \mathbf{0}\,.$$

Moreover,

$$\lambda_{\max}(A) = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \quad \textit{and} \quad \lambda_{\min}(A) = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}$$

and the maximum and minimum are attained for $\mathbf{x} = \mathbf{v}_1$ *and for* $\mathbf{x} = \mathbf{v}_n$*, respectively, where* $\mathbf{v}_1$ *(resp.* $\mathbf{v}_n$*) is the unit-norm eigenvector of* A *associated with its largest (resp. smallest) eigenvalue of* A*.*

Proof: By real spectral decomposition, we can express a symmetric matrix A using orthonormal vectors, $\mathbf{v}_1, \ldots, \mathbf{v}_n$, as

$$A = \sum_{i=1}^{n} \lambda_i(A)\mathbf{v}_i\mathbf{v}_i^\top.$$

From Lemma 7.1, we get

$$\lambda_1(A)\mathbf{x}^\top\mathbf{x} \geq \mathbf{x}^\top A\mathbf{x} \geq \lambda_n(A)\mathbf{x}^\top\mathbf{x}$$

when $p = 1$ and $q = n$. From this, we see that $\lambda_1(A)$ and $\lambda_n(A)$ are the maximum and minimum of the Rayleigh quotients, respectively.

■

We can generalize the results in Lemma 7.1 and Theorem 7.1 to derive the bounds for the maximum/minimum values of the Rayleigh quotient over a k-dimensional subspace.

Lemma 7.2 (Poincare inequality) *Consider an $n \times n$ symmetric matrix A and a k-dimensional subspace $\mathbb{W}$ of $\mathbb{R}^n$. Then,*

$$\min_{\substack{\mathbf{x}\in\mathbb{W}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_k(A)\,, \quad \max_{\substack{\mathbf{y}\in\mathbb{W}\\ \mathbf{y}\neq\mathbf{0}}} \frac{\mathbf{y}^\top A\mathbf{y}}{\mathbf{y}^\top\mathbf{y}} \geq \lambda_{n-k+1}(A)\,.$$

Proof: Let the real spectral decomposition of A be $A = \sum_{i=1}^n \lambda_i(A)\mathbf{v}_i\mathbf{v}_i^\top$. For $\mathbb{W}' = \text{span}\{\mathbf{v}_k, \ldots, \mathbf{v}_n\}$, Fact 3.6 implies that $\dim(\mathbb{W}\cap\mathbb{W}') \geq \dim\mathbb{W} + \dim\mathbb{W}' - \dim\mathbb{V} = k + (n-k+1) - n = 1$. Thus, there must be $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{W}\cap\mathbb{W}'$. According to Lemma 7.1, $\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_k(A)$, because $\mathbf{x} \in \mathbb{W}'$. Since $\mathbf{x}$ is also in $\mathbb{W}$, we get the first inequality. We can prove the second inequality following the same steps starting from $\mathbb{W}' = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-k+1}\}$.

■

Combining the results above, we derive the following useful result showing that every eigenvalue of a symmetric matrix can be expressed as minimax or maxmin of the Rayleigh quotient.

Theorem 7.2 (Minimax Principle) *Consider an $n \times n$ symmetric matrix A. Then, for $1 \leq k \leq n$,*

$$\lambda_k(A) = \max_{\mathbb{W}:\dim\mathbb{W}=k} \min_{\substack{\mathbf{x}\in\mathbb{W}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \tag{7.1}$$

$$= \min_{\mathbb{W}:\dim\mathbb{W}=n-k+1} \max_{\substack{\mathbf{x}\in\mathbb{W}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}\,. \tag{7.2}$$

Proof: Let the real spectral decomposition of A be $A = \sum_{i=1}^n \lambda_i(A)\mathbf{v}_i\mathbf{v}_i^\top$. By the first inequality in Lemma 7.2,

$$\min_{\substack{\mathbf{x}\in\mathbb{W}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_k(A),$$

for any subspace $\mathbb{W}$ with $\dim \mathbb{W} = k$. In order to prove (7.1), we then need to find the k-dimensional subspace where the minimum Rayleigh quotient is $\lambda_k(A)$. When $\mathbb{W}^* = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, $\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \geq \lambda_k(A)$ for $\mathbf{0} \neq \mathbf{x} \in \mathbb{W}^*$, according to Lemma 7.1. This implies $\min_{\substack{\mathbf{x}\in\mathbb{W}^* \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_k(A)$, as $\mathbf{v}_k^\top A\mathbf{v}_k = \lambda_k(A)$. In other words, because the minimum Rayleigh quotient in the k-dimensional subspace $\mathbb{W}^*$ is $\lambda_k(A)$, it holds that

$$\lambda_k(A) \;=\; \max_{\mathbb{W}:\dim\mathbb{W}=k} \; \min_{\substack{\mathbf{x}\in\mathbb{W} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}.$$

We can prove the second equality using the remaining inequalities from Lemma 7.2.

■

From Theorem 7.2, we now know that

$$\lambda_{\max}(A) = \lambda_1(A) = \max_{|\mathbf{x}|=1} \mathbf{x}^\top A\mathbf{x}\,, \quad \lambda_{\min}(A) = \lambda_n(A) = \min_{|\mathbf{x}|=1} \mathbf{x}^\top A\mathbf{x}. \tag{7.3}$$

By replacing A with $A^\top A$, we find the following relationship against the spectral norm of A:

$$\sqrt{\lambda_1(A^\top A)} = \max_{|\mathbf{x}|=1} |A\mathbf{x}| = \|A\|_2\,, \quad \sqrt{\lambda_n(A^\top A)} = \min_{|\mathbf{x}|=1} |A\mathbf{x}|.$$

7.4.1 Eigenvalues and Singular Values of Matrix Sum

We can derive the following result on the eigenvalues of a matrix sum, called the Weyl's inequality, from the minimax principle above.

Theorem 7.3 (Weyl's inequality of eigenvalues) *Let A and B be $n \times n$ symmetric matrices with eigenvalues $\lambda_i(A)$ and $\lambda_i(B)$, respectively. Then*

$$\lambda_{k+\ell+1}(A+B) \leq \lambda_{k+1}(A) + \lambda_{\ell+1}(B)$$

for $k, \ell = 0, 1, 2, \ldots$.

Proof: Let us bound $\lambda_{k+\ell+1}(A+B)$ in terms of the eigenvalues of A and B. By the minimax principle (Theorem 7.2), we have

$$\lambda_{k+\ell+1}(A+B) = \min_{\mathbb{W}:\dim\mathbb{W}=n-k-\ell} \; \max_{\substack{\mathbf{x}\in\mathbb{W} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top (A+B)\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}\,.$$

Again by the minimax principle, we can find subspaces $\mathbb{W}_A$ and $\mathbb{W}_B$ of $\mathbb{R}^n$ of dimensions $n-k$ and $n-\ell$, respectively, such that

$$\lambda_{k+1}(A) = \max_{\substack{\mathbf{x}\in\mathbb{W}_A \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \quad \text{and} \quad \lambda_{\ell+1}(B) = \max_{\substack{\mathbf{x}\in\mathbb{W}_B \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}\,.$$

If we let $\mathbb{W}_1 = \mathbb{W}_A \cap \mathbb{W}_B$ be their intersection, then it is clear that

$$\max_{\substack{\mathbf{x}\in\mathbb{W}_1 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \le \lambda_{k+1}(A) \quad \text{and} \quad \max_{\substack{\mathbf{x}\in\mathbb{W}_1 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \le \lambda_{\ell+1}(B)\,.$$

This intersection $\mathbb{W}_1$ has dimension at least $n-k-\ell$ due to Fact 3.6. Let $\mathbb{W}_2$ be any $(n-k-\ell)$-dimensional subspace of $\mathbb{W}_1$. We then have

$$\begin{aligned}
\lambda_{k+\ell+1}(A+B) &\le \max_{\substack{\mathbf{x}\in\mathbb{W}_2 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top(A+B)\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \\
&\le \max_{\substack{\mathbf{x}\in\mathbb{W}_1 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top(A+B)\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \qquad (\text{since } \mathbb{W}_2 \subset \mathbb{W}_1) \\
&\le \max_{\substack{\mathbf{x}\in\mathbb{W}_1 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \max_{\substack{\mathbf{x}\in\mathbb{W}_1 \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \\
&\le \lambda_{k+1}(A) + \lambda_{\ell+1}(B)\,.
\end{aligned}$$

■

By symmetrization (5.15), we can turn a result on eigenvalues of a symmetric matrix into that on singular values of an arbitrary matrix. Along this line, we can rewrite Theorem 7.3 to be about singular values of an arbitrary matrix. For notational convenience, $\sigma_k(A)$ denotes the k-th largest singular value of A.

Theorem 7.4 (Weyl's inequality of singular Values) *Let A and B be matrices of same size with singular values $\sigma_i(A)$ and $\sigma_i(B)$, respectively. Then, for $k, \ell = 0, 1, 2, \ldots$,*

$$\sigma_{k+\ell+1}(A+B) \le \sigma_{k+1}(A) + \sigma_{\ell+1}(B)\,.$$

Proof: According to Lemma 5.4, $\sigma_i(A) = \lambda_i(s(A))$, $\sigma_i(B) = \lambda_i(s(B))$ and $\sigma_i(A+B) = \lambda_i(s(A+B))$. Because a symmetrization is symmetric, we get

$$\begin{aligned}
\sigma_{k+\ell+1}(A+B) &= \lambda_{k+\ell+1}(s(A+B)) = \lambda_{k+\ell+1}(s(A)+s(B)) \\
&\le \lambda_{k+1}(s(A)) + \lambda_{\ell+1}(s(B)) = \sigma_{k+1}(A) + \sigma_{\ell+1}(B),
\end{aligned}$$

where the inequality comes from Theorem 7.3. ■

This result provides us with a bound on the rank of the sum of two matrices in terms of their ranks. If $\operatorname{rank} A = k$ and $\operatorname{rank} B = \ell$, then $\sigma_{k+1}(A) = 0$ and $\sigma_{\ell+1}(B) = 0$. According to Theorem 7.4, $\sigma_{k+\ell+1}(A+B) = 0$ and hence $\operatorname{rank}(A+B) \le k+\ell$. Even when $\sigma_{k+1}(A)$ and $\sigma_{\ell+1}(B)$ were not exactly zeros but very close to zeros, we often treat them as zeros and consider the rank of $A+B$ as at most $k+\ell$, in practice.

When $\ell = 1$ in Theorem 7.4, the inequality reduces to $\sigma_{k+1}(A+B) \le \sigma_{k+1}(A) + \sigma_1(B)$, from which we get the following observation by replacing A and B appropriately.

Fact 7.6 (Bound of additive perturbation) *Let A and B be $m \times n$ matrices. Then, for $1 \le k \le \min\{m,n\}$,*

$$|\sigma_k(A) - \sigma_k(B)| \le \sigma_1(A-B) = \|A-B\|_2 .$$

Proof: When $\ell = 1$ in Theorem 7.4,

$$\sigma_k(A) \le \sigma_k(B) + \sigma_1(A-B), \quad \sigma_k(B) \le \sigma_k(A) + \sigma_1(B-A),$$

with appropriate choices of A and B. Combining with $\sigma_1(\cdot) = \|\cdot\|_2$, we obtain the statement. ■

The following inequality involving an eigenvalue of a sum of two diagonal matrices comes handy later.

Fact 7.7 *Let A and B be $n \times n$ symmetric matrices. Then, for $1 \le k \le n$,*

$$\lambda_k(A) + \lambda_{\min}(B) \le \lambda_k(A+B) \le \lambda_k(A) + \lambda_{\max}(B) .$$

Proof: By (7.3), it holds that $\lambda_n(B) \le \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \le \lambda_1(B)$ for all $\mathbf{x} \ne \mathbf{0}$. Therefore, for all $\mathbf{x} \ne \mathbf{0}$, the following also holds:

$$\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \lambda_n(B) \le \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \frac{\mathbf{x}^\top (A+B)\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \le \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \lambda_1(B),$$

from which we get

$$\begin{aligned} \max_{\dim \mathbb{W}=k} \min_{\substack{\mathbf{x}\in\mathbb{W} \\ \mathbf{x}\ne\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \lambda_n(B) &\le \max_{\dim \mathbb{W}=k} \min_{\substack{\mathbf{x}\in\mathbb{W} \\ \mathbf{x}\ne\mathbf{0}}} \frac{\mathbf{x}^\top (A+B)\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \\ &\le \max_{\dim \mathbb{W}=k} \min_{\substack{\mathbf{x}\in\mathbb{W} \\ \mathbf{x}\ne\mathbf{0}}} \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} + \lambda_1(B). \end{aligned}$$

We derive the desired inequalities using Theorem 7.2. ■

Note that we proved both inequalities at the same time, rather than using Theorem 7.3 to prove the second inequality, which would have required a separate proof for the first inequality.

From this result, we observe that the eigenvalue increases when we add a positive (semi-)definite matrix to a symmetric matrix.

Fact 7.8 *Let A and B be $n \times n$ symmetric matrices. Let $1 \leq k \leq n$. If B is positive semi-definite, then*

$$\lambda_k(A) \leq \lambda_k(A+B)\,.$$

If B is positive definite, then

$$\lambda_k(A) < \lambda_k(A+B)\,.$$

If we add a symmetric positive semi-definite matrix of rank-one to a symmetric matrix, we can further obtain an upper bound on the eigenvalues as follows,

$$\lambda_k(A) \leq \lambda_k(A+\mathbf{q}\mathbf{q}^\top) \leq \lambda_k(A) + |\mathbf{q}|^2$$

for all $1 \leq k \leq n$.

Proof: Since B is positive semi-definite, $\lambda_{\min}(B) \geq 0$. Hence, Fact 7.7 implies the first result. If B is positive definite, $\lambda_{\min}(B) > 0$ implies the second result. A symmetric positive semi-definite matrix of rank-one is always in a form of $\mathbf{q}\mathbf{q}^\top$ for a non-zero vector $\mathbf{q}$. It is easy to see that $\lambda_{\min}(\mathbf{q}\mathbf{q}^\top) = 0$ and $\lambda_{\max}(\mathbf{q}\mathbf{q}^\top) = |\mathbf{q}|^2$. Fact 7.7 then implies the third result. ■

We arrive at the following result on eigenvalue interlacing by systematically analyzing how eigenvalues change when a rank-one positive semi-definite matrix is added to a symmetric matrix.

Theorem 7.5 (Eigenvalue Interlacing) *Let A be an $n \times n$ symmetric matrix and B an $n \times n$ symmetric positive semi-definite matrix of rank-one. Then,*

$$\lambda_{k+1}(A+B) \leq \lambda_k(A) \leq \lambda_k(A+B)\,, \quad \textit{for all } k = 1, \ldots, n-1$$

and

$$\lambda_{k+1}(A) \leq \lambda_k(A-B) \leq \lambda_k(A)\,, \quad \textit{for all } k = 1, \ldots, n-1.$$

Proof: Given appropriate orthonormal vector sets, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, we can write $A + B$ and A as

$$A + B = \sum_{i=1}^{n} \lambda_i(A + B)\mathbf{v}_i\mathbf{v}_i^\top \quad \text{and} \quad A = \sum_{i=1}^{n} \lambda_i(A)\mathbf{u}_i\mathbf{u}_i^\top,$$

using real spectral decomposition. Consider the following three subspaces of $\mathbb{R}^n$: $\mathbb{U} = \text{span}\{\mathbf{u}_k, \ldots, \mathbf{u}_n\}$, $\mathbb{V} = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$, and $\mathbb{W} = \text{null}\, B$. Because $\dim \mathbb{U} + \dim \mathbb{V} = n + 2$, $\dim(\mathbb{U} \cap \mathbb{V}) \geq 2$ due to Fact 3.6. Furthermore, because $\dim \mathbb{W} = n - \text{rank}(B) = n - 1$, $\dim(\mathbb{U} \cap \mathbb{V} \cap \mathbb{W}) \geq 1$ also due to Fact 3.6. Let $\hat{\mathbf{x}}$ be a unit vector in $\mathbb{U} \cap \mathbb{V} \cap \mathbb{W}$. Since $B\hat{\mathbf{x}} = \mathbf{0}$, $\hat{\mathbf{x}} \in \mathbb{V}$, and $\hat{\mathbf{x}} \in \mathbb{U}$,

$$\begin{aligned} \lambda_{k+1}(A + B) &\leq \hat{\mathbf{x}}^\top (A + B)\hat{\mathbf{x}} && \text{by Lemma 7.1 and } \hat{\mathbf{x}} \in \mathbb{V} \\ &= \hat{\mathbf{x}}^\top A\hat{\mathbf{x}} && \text{by } B\hat{\mathbf{x}} = \mathbf{0} \\ &\leq \lambda_k(A) && \text{by Lemma 7.1 and } \hat{\mathbf{x}} \in \mathbb{U}. \end{aligned}$$

Combining this with Fact 7.8, we obtain the two inequalities in the first line. We can prove the inequalities in the second line by replacing A with $A - B$ in the proof here.

■

7.5 Ellipsoidal Geometry of Positive Definite Matrices

A geometric interpretation of an $n \times n$ symmetric positive definite matrix Σ can be related to an ellipsoid in $\mathbb{R}^n$. Consider the following ellipsoidal set based on a quadratic inequality:

$$\mathcal{E} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \Sigma^{-1} \mathbf{x} \leq 1\}.$$

Since we can easily translate this ellipsoid to be centered on $\mathbf{a}$ by $\{\mathbf{x} \in \mathbb{R}^n : (\mathbf{x}-\mathbf{a})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{a}) \leq 1\}$, we will only consider $\mathcal{E}$ centered at the origin. According to Fact 7.3, Σ^{-1} is also positive definite, and we can write it using a set of orthonormal vectors, $\mathbf{u}_1, \ldots, \mathbf{u}_n$, as

$$\Sigma^{-1} = \sum_{i=1}^{n} \lambda_i(\Sigma)^{-1}\mathbf{u}_i\mathbf{u}_i^\top.$$

We can write an arbitrary vector $\mathbf{x} \in \mathbb{R}^n$ as $\mathbf{x} = \sum_{i=1}^n y_i\mathbf{u}_i$ with $y_i = \langle \mathbf{x}, \mathbf{u}_i \rangle$. From this we get

$$\mathbf{x}^\top \Sigma^{-1} \mathbf{x} = \sum_{i=1}^{n} \frac{y_i^2}{\lambda_i(\Sigma)}.$$

We can then represent an ellipsoid in terms of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ as

$$\mathcal{E} = \Big\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \sum_{i=1}^{n} y_i \mathbf{u}_i,\ \sum_{i=1}^{n} \frac{y_i^2}{\lambda_i(\Sigma)} \leq 1\Big\}. \tag{7.4}$$

Let us convert this expression into a form that is more familiar to us where each axis of the ellipsoid coincides with each standard basic vector $\mathbf{e}_i$. We start from $\mathcal{E}' = \{\mathbf{y} \in \mathbb{R}^n : \sum_{i=1}^{n} y_i^2/\lambda_i(\Sigma) \leq 1\}$ and linearly transform $\mathbf{e}_i$ to $\mathbf{u}_i$, in order to obtain $\mathcal{E}$. The i-th longest axis of this ellipsoid has the length of $2\sqrt{\lambda_i(\Sigma)}$ and the direction of $\mathbf{u}_i$. Since this transformation just shuffles the axes of the ellipsoid $\mathcal{E}$ keeping their orthogonality, it holds intuitively that two n-dimensional ellipsoids are of equal volumes,[3] that is, $\mathbf{vol}(\mathcal{E}) = \mathbf{vol}(\mathcal{E}')$. Let $B_n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{x} \leq 1\}$ be an n-dimensional unit sphere. We can transform this unit sphere B_n into the ellipsoid $\mathcal{E}'$ by linearly transforming $\mathbf{e}_i$ to $\sqrt{\lambda_i(\Sigma)}\mathbf{e}_i$. Since the length of each orthogonal direction $\mathbf{e}_i$ grows from 1 to $\sqrt{\lambda_i(\Sigma)}$, the volume of $\mathcal{E}'$ is $\prod_{i=1}^{n} \sqrt{\lambda_i(\Sigma)}$ times that of the unit sphere. That is,

$$\mathbf{vol}(\mathcal{E}) = \mathbf{vol}(\mathcal{E}') = \sqrt{\prod_{i=1}^{n} \lambda_i(\Sigma)}\ \mathbf{vol}(B_n). \tag{7.5}$$

Let us consider a related concept called Mahalanobis distance. Given a positive definite matrix Σ, both $\mathbf{x}^\top \Sigma \mathbf{y}$ and $\mathbf{x}^\top \Sigma^{-1} \mathbf{y}$ are inner products, according to Theorem 4.1. The inner product induces a norm $f(\mathbf{x}) = \sqrt{\mathbf{x}^\top \Sigma^{-1} \mathbf{x}}$, and further we can view $f(\mathbf{x} - \mathbf{y})$ as a distance between $\mathbf{x}$ and $\mathbf{y}$, as in

$$d(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} - \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \Sigma^{-1} (\mathbf{x} - \mathbf{y})}.$$

We call such a distance the Mahalanobis distance with respect to the positive definite matrix A. From the perspective of data science, this is a distance between two vectors taking into account the covariance of a data distribution. An ellipsoid from above can be thought of as a set of all vectors whose Mahalanobis distances to a center vector are less than or equal to 1.

Let us use SVD to interpret the geometry of a matrix that transforms a unit sphere. In the beginning of this section, we transformed linearly a unit sphere to an ellipsoid to compute the volume of the ellipsoid via matching $\mathbf{e}_i$ with $\mathbf{u}_i$. Similarly, let us regard an $n \times d$ matrix A as a linear transformation mapping $\mathbf{v}_i$ to $\sigma_i \mathbf{u}_i$, where $(\sigma_i, \mathbf{v}_i, \mathbf{u}_i), i = 1, \ldots, k$ are singular triplets. $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is

[3] We introduce the concept of volume in $\mathbb{R}^n$ more precisely in Chapter 8.

an orthonormal columns of V. Then, the image of a unit sphere $B_d = \{\mathbf{x} \in \mathbb{R}^d, |\mathbf{x}| \le 1\}$ through A is

$$\begin{aligned} AB_d &= \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, |\mathbf{x}| \le 1\} \\ &= \{A\mathbf{v} : \mathbf{v} = z_1\mathbf{v}_1 + \cdots + z_d\mathbf{v}_d,\ z_1^2 + \cdots + z_d^2 \le 1\} \\ &= \{z_1\sigma_1\mathbf{u}_1 + \cdots + z_k\sigma_k\mathbf{u}_k : z_1^2 + \cdots + z_k^2 \le 1\} \\ &= \Big\{y_1\mathbf{u}_1 + \cdots + y_k\mathbf{u}_k : \frac{y_1^2}{\sigma_1^2} + \cdots + \frac{y_k^2}{\sigma_k^2} \le 1\Big\}, \end{aligned}$$

where the last description coincide with (7.4) once we replace $\sqrt{\lambda_i(\Sigma)}$ by σ_i. The AB_d is an k-dimensional ellipsoid in $\mathbb{R}^n$ whose axes have lengths of σ_i's, which determine how far to stretch the sphere into an ellipsoid. This interpretation also highlights the importance of the rank of matrix A in determining the dimensionality of the resulting ellipsoid. Connecting to Section 5.4.1, the diagonal part of SVD determines the shape of ellipsoid since a sphere is invariant under rotation or reflection.

7.6 Application: Kernel Trick in Machine Learning

We base our discussion here on [5]. Consider a particular instance of classification, where our goal is to find a linear function[4] that classifies data points into two groups. There are n data points:

$$\big\{(\mathbf{z}_i, \ell_i) \in \mathbb{R}^{d+1} \times \{+1, -1\} : i = 1, \ldots, n\big\}.$$

We use $\mathcal{I}_+ = \{i : \ell_i = +1\}$ and $\mathcal{I}_-\{i : \ell_i = -1\}$ as index sets of the positive and negative examples, respectively. For each group, we use $n_+ = \big|\mathcal{I}_+\big|$ ($n_- = \big|\mathcal{I}_-\big|$) as its size and $\boldsymbol{\mu}_+ = \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\mathbf{z}_i$ ($\boldsymbol{\mu}_- = \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\mathbf{z}_i$) as its centroid.

A new observation $\mathbf{z}$ is classified based on whether the new vector is pointing toward the same direction from the mean of the group means, $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)$, as the direction from the negative group mean to the positive group mean. That is,

$$\ell = \operatorname{sign}\langle \mathbf{z} - \boldsymbol{\mu},\ \boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \rangle.$$

If we expand this classification rule further, we see that this is expressed as the

[4] In practice, we use an affine function by adding a constant term to a linear function.

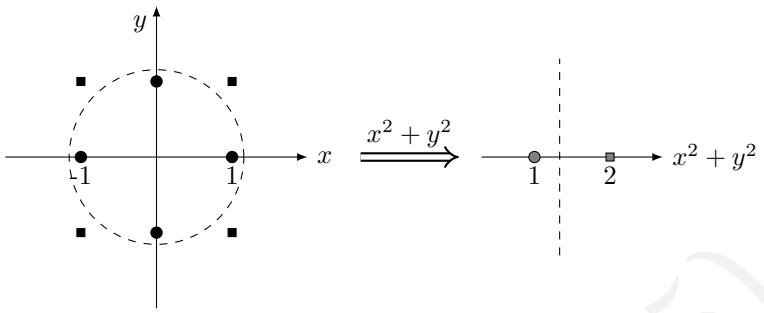


Figure 7.1: Example of Quadratic Embedding

sum of inner products against the data points, as in

$$\begin{aligned}\langle \mathbf{z}-\boldsymbol{\mu}, \boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \rangle &= \langle \mathbf{z}, \boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \rangle - \frac{1}{2}\langle \boldsymbol{\mu}_+ + \boldsymbol{\mu}_-, \boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \rangle \\ &= \langle \mathbf{z}, \boldsymbol{\mu}_+ \rangle - \langle \mathbf{z}, \boldsymbol{\mu}_- \rangle - \frac{1}{2}\big(\langle \boldsymbol{\mu}_+, \boldsymbol{\mu}_+ \rangle - \langle \boldsymbol{\mu}_-, \boldsymbol{\mu}_- \rangle\big) \\ &= \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\langle \mathbf{z}, \mathbf{z}_i\rangle - \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\langle \mathbf{z}, \mathbf{z}_i\rangle - b,\end{aligned}$$

where $b = \frac{1}{2}\big(\langle \boldsymbol{\mu}_+, \boldsymbol{\mu}_+ \rangle - \langle \boldsymbol{\mu}_-, \boldsymbol{\mu}_- \rangle\big) = \frac{1}{2}\big(|\boldsymbol{\mu}_+|^2 - |\boldsymbol{\mu}_-|^2\big)$ is a constant.

We use a kernel trick to handle data points that are not linearly separable. This is especially useful when the data points in the original data $\big\{(\mathbf{x}_i, \ell_i) \in \mathbb{R}^{d+1} \times \{+1, -1\} : i = 1, \ldots, n\big\}$ are not linearly separated into positive and negative classes, but they are linearly separable after they are embedded into a higher-dimensional space, called a feature space, using a nonlinear transformation $\psi : \mathbb{R}^d \to \mathbb{R}^N$. N is often greater than d and can even be infinitely large. Then, the classification rule in this feature space becomes

$$\ell = \operatorname{sign}\Big(\frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\langle \psi(\mathbf{x}), \psi(\mathbf{x}_i)\rangle - \frac{1}{n_-}\sum_{i\in\mathcal{I}_-}\langle \psi(\mathbf{x}), \psi(\mathbf{x}_i)\rangle - b\Big),$$

and b is also similarly defined using $\langle \psi(\mathbf{x}_i), \psi(\mathbf{x}_j)\rangle$.

Example 7.2 Consider two sets $S_- = \{(1,0), (-1,0), (0,1), (0,-1)\}$ and $S_+ = \{(1,1), (-1,1), (1,-1), (-1,-1)\}$ in $\mathbb{R}^2$. As they are, these two sets cannot be separated by a linear function. We can easily check this by visual inspection. See Figure 7.1, where S_- and S_+ are marked with circles and squares, respectively.

It is however possible to separate them once we embed the points in $\mathbb{R}^6$ by transforming each data point (x, y) into $(1, x, y, x^2, xy, y^2)$. Once embedded in

$\mathbb{R}^6$, $\boldsymbol{\mu}_+ = (1,0,0,1,0,1)^\top$ and $\boldsymbol{\mu}_- = (1,0,0,1/2,0,1/2)^\top$, and the two sets are separated with $\boldsymbol{\mu} = (1,0,0,3/4,0,3/4)^\top$ and $\boldsymbol{\mu}_+ - \boldsymbol{\mu}_- = (0,0,0,1/2,0,1/2)$. ∎

We often call the inner product between the embedded vectors a kernel and use $K(\cdot,\cdot)$ to refer to it:

$$K(\mathbf{x},\mathbf{y}) = \big\langle \psi(\mathbf{x}), \psi(\mathbf{y}) \big\rangle. \tag{7.6}$$

This helps us simplify the classification rule into

$$\ell = \operatorname{sign}\Big(\frac{1}{n_+}\sum_{i\in\mathcal{I}_+} K(\mathbf{x},\mathbf{x}_i) - \frac{1}{n_-}\sum_{i\in\mathcal{I}_-} K(\mathbf{x},\mathbf{x}_i) - b\Big).$$

This kernel-based perspective is useful when it is computationally more favorable to compute $K(\cdot,\cdot)$ directly than to compute the inner product of two vectors after embedding them using $\psi(\mathbf{x})$. We consider two examples below.

A kernel in (7.6) is symmetric, i.e., $K(\mathbf{x},\mathbf{y}) = K(\mathbf{y},\mathbf{x})$, because it is an inner product, and a matrix constructed from $K(\mathbf{x}_i,\mathbf{x}_j)$ is positive definite, according to Theorem 4.1. We thus refer to such a kernel as a symmetric positive definite kernel.

Let us introduce a few notations to facilitate analyzing kernels. Similar to the standard inner product in a finite-dimensional Euclidean space, we can define an inner product between two vectors in an infinite-dimensional space as

$$\big\langle (a_k), (b_k) \big\rangle = \sum_{k=1}^{\infty} a_k b_k,$$

where $(a_k) = (a_1, a_2, \ldots) \in \mathbb{R}^\infty$ and $(b_k) = (b_1, b_2, \ldots) \in \mathbb{R}^\infty$. We use ; to indicate that we are vertically stacking two vectors. For instance, $(\mathbf{u};\mathbf{v}) = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}$. We also define $\mathbf{j} = (j_1, j_2, \ldots, j_k)^\top \in \{1,\ldots,n\}^k$ as a vector of positive integers.

A Polynomial Kernel

For two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ and a positive integer k, we can expand the k-th power of an inner product as

$$\big(\mathbf{x}^\top\mathbf{y}\big)^k \;=\; \underbrace{\Big(\sum_{j=1}^n x_j y_j\Big)\cdots\Big(\sum_{j=1}^n x_j y_j\Big)}_{k \text{ times}}$$

$$= \sum_{\mathbf{j}\in\{1,\cdots,n\}^k} \prod_{i=1}^{k} x_{j_i} y_{j_i}$$
$$= \sum_{\mathbf{j}\in\{1,\cdots,n\}^k} \prod_{i=1}^{k} x_{j_i} \prod_{i=1}^{k} y_{j_i} .$$

If we denote all n^k vectors in $\{1,\ldots,n\}^k$ by $\mathbf{j}^{(1)},\ldots,\mathbf{j}^{(n^k)}$, $\{\mathbf{j}^{(1)},\ldots,\mathbf{j}^{(n^k)}\} = \{1,\ldots,n\}^k$. We use $(\mathbf{j}^{(p)})_i$ or ${j^{(p)}}_i$ to refer to the i-th entry of $\mathbf{j}^{(p)}$. Using these notations and re-arranging terms, we get

$$\left(\mathbf{x}^\top \mathbf{y}\right)^k = \sum_{p=1}^{n^k} \prod_{i=1}^{k} x_{(\mathbf{j}^{(p)})_i} \prod_{i=1}^{k} y_{(\mathbf{j}^{(p)})_i} = \sum_{p=1}^{n^k} \prod_{i=1}^{k} x_{{j^{(p)}}_i} \prod_{i=1}^{k} y_{{j^{(p)}}_i} .$$

We introduce the following nonlinear embedding transformation $\psi_k : \mathbb{R}^n \to \mathbb{R}^{n^k}$ to interpret the last summation with n^k summands as a standard inner product:

$$\psi_k(\mathbf{x}) = \Big(\prod_{i=1}^{k} x_{{j^{(1)}}_i}, \prod_{i=1}^{k} x_{{j^{(2)}}_i}, \ldots, \prod_{i=1}^{k} x_{{j^{(n^k)}}_i} \Big)^\top \in \mathbb{R}^{n^k} . \tag{7.7}$$

This embedding results in a simplified representation:

$$\left(\mathbf{x}^\top \mathbf{y}\right)^k = \langle \psi_k(\mathbf{x}), \psi_k(\mathbf{y}) \rangle .$$

Because ψ_k depends on the specification of each $\mathbf{j}^{(p)}$ which is not uniquely determined, ψ_k is not uniquely determined either.

In other words, the k-th order polynomial kernel in $\mathbb{R}^n$

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y})^k$$

can be represented by a standard inner product in $\mathbb{R}^{n^k}$

$$\langle \psi_k(\mathbf{x}), \psi_k(\mathbf{y}) \rangle .$$

Let us augment $\mathbf{x}$ as $\mathbf{x}' = (1; \mathbf{x})$ to handle constant terms of polynomials. By setting $\mathbf{x}' = (1; \mathbf{x})$ and $\mathbf{y}' = (1; \mathbf{y})$, we can obtain

$$(1 + \mathbf{x}^\top \mathbf{y})^k = \left(\mathbf{x}'^\top \mathbf{y}'\right)^k = \langle \psi_{k+1}(\mathbf{x}'), \psi_{k+1}(\mathbf{y}') \rangle$$

just as same as the above case with embedding ψ_{k+1}.

Gaussian Kernel

One of the most popularly used kernels is the Gaussian kernel. It is defined as $K(x,y) = e^{-\frac{1}{2}(x-y)^2}$ in $\mathbb{R}^1$. If we expand this kernel, we get

$$\begin{aligned} e^{-\frac{1}{2}(x-y)^2} &= e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}y^2} e^{xy} \\ &= e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}y^2} \sum_{k=0}^{\infty} \frac{1}{k!}(xy)^k \\ &= \sum_{k=0}^{\infty} e^{-\frac{1}{2}x^2} \frac{x^k}{\sqrt{k!}} \times e^{-\frac{1}{2}y^2} \frac{y^k}{\sqrt{k!}} \,. \end{aligned}$$

We interpret the last infinite sum as an inner product in $\mathbb{R}^\infty$ by introducing the following embedding $\psi_{G1} : \mathbb{R} \to \mathbb{R}^\infty$:

$$\psi_{G1}(x) = e^{-\frac{1}{2}x^2}(1; x; \frac{x^2}{\sqrt{2!}}; \ldots; \frac{x^k}{\sqrt{k!}}; \ldots) \in \mathbb{R}^\infty \,. \tag{7.8}$$

This leads to the following simplified representation:

$$K(x,y) = \langle \psi_{G1}(x), \psi_{G1}(y) \rangle \,.$$

Let us generalize the inner product representation to the Gaussian kernel in $\mathbb{R}^n$. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we derive this kernel as $K(\mathbf{x},\mathbf{y}) = e^{-\frac{1}{2}|\mathbf{x}-\mathbf{y}|^2}$ using the embedding function (7.7), following

$$\begin{aligned} e^{-\frac{1}{2}|\mathbf{x}-\mathbf{y}|^2} &= e^{-\frac{1}{2}|\mathbf{x}|^2} e^{-\frac{1}{2}|\mathbf{y}|^2} e^{\mathbf{x}^\top \mathbf{y}} \\ &= e^{-\frac{1}{2}|\mathbf{x}|^2} e^{-\frac{1}{2}|\mathbf{y}|^2} \sum_{k=0} \frac{1}{k!} (\mathbf{x}^\top \mathbf{y})^k \\ &= e^{-\frac{1}{2}|\mathbf{x}|^2} e^{-\frac{1}{2}|\mathbf{y}|^2} \sum_{k=0} \frac{1}{k!} \langle \psi_k(\mathbf{x}), \psi_k(\mathbf{y}) \rangle \\ &= \sum_{k=0} \Big\langle e^{-\frac{1}{2}|\mathbf{x}|^2} \frac{1}{\sqrt{k!}} \psi_k(\mathbf{x}),\, e^{-\frac{1}{2}|\mathbf{y}|^2} \frac{1}{\sqrt{k!}} \psi_k(\mathbf{y}) \Big\rangle \,. \end{aligned}$$

We rewrite the sum of the inner products as an inner product in $\mathbb{R}^\infty$ by defining the generalized embedding $\psi_{Gk} : \mathbb{R}^n \to \mathbb{R}^\infty$ as

$$\psi_{Gk}(\mathbf{x}) = e^{-\frac{1}{2}|\mathbf{x}|^2} \Big(1;\ \psi_1(\mathbf{x});\ \frac{1}{\sqrt{2!}}\psi_2(\mathbf{x});\ \frac{1}{\sqrt{3!}}\psi_3(\mathbf{x});\ \ldots \Big) \in \mathbb{R}^\infty \,.$$

With this embedding[5] we see that we can write the Gaussian kernel as an inner product in the embedding space:

$$K(\mathbf{x},\mathbf{y}) = \langle \psi_{Gk}(x), \psi_{Gk}(y) \rangle \,.$$

[5] Already with only the leading four terms of ψ_{Gk}, we notice the explosion of the embedding dimension: $\Big(1; \psi_1(\mathbf{x}); \frac{1}{\sqrt{2!}}\psi_2(\mathbf{x}); \frac{1}{\sqrt{3!}}\psi_3(\mathbf{x})\Big) \in \mathbb{R}^{1+n^1+n^2+n^3}$.

Kernel tricks have been extensively studied both theoretically and practically and are widely used in practice. We suggest you refer to other materials for further discussion.

Chapter 8

Determinants

Given n vectors in $\mathbb{R}^n$, a volume of parallelopiped with the n vectors as its edges is an important quantity for many scientific works. Even though a volume seems intrinsic to the Euclidean space, it needs an agreement on what a volume in $\mathbb{R}^n$ for $n \geq 4$ means. A starting point is a volume of the unit cube $\{\mathbf{x} \in \mathbb{R}^n : 0 \leq x_i \leq 1 \text{ for all } i = 1, \ldots, n\}$ in $\mathbb{R}^n$. We set the volume of the unit cube in $\mathbb{R}^n$ as 1 regardless of its dimension. It would also make sense that lengthening an edge of a parallelopiped twice inflates the volume twice as well as that adding volumes of two parallelopipeds sharing $n-1$ edges equals the volume of a single parallelopiped built by adding two unmatched edge vectors and keeping the other $n-1$ edges. In addition, swapping of the coordinates doesn't change the volumes of objects in $\mathbb{R}^n$. These properties encapsulate the volume in high-dimensional Euclidean spaces.

Mathematicians invented a function called determinant, with its symbol det, that obeys these rules on square matrices. If we build an $n \times n$ matrix A by stacking n vectors in $\mathbb{R}^n$ row-wise, $\det(A)$ is a signed volume of the parallelopiped with the n rows of A as its edges. $\det(I_n) = 1$ since the identity matrix corresponds to the unit cube. $\det(A)$ is linear in each row of A, in other words, multilinear in the rows of A. In addition, $\det(A)$ changes its sign, not its absolute value if we swap two rows. In fact, there is only one function satisfying the above three properties, which we will show soon. Therefore, we may set $|\det(A)|$ as a volume of parallelopiped built by n rows as its edges.

The det function further satisfies the following properties:

- $\det(A) = 0$ if and only if A is not invertible;

- $\det(A^\top) = \det(A)$;
- $\det(A^{-1}) = \det(A)^{-1}$ for invertible A.

It can be written down a function of all entries of the matrix, which is referred to as Laplace expansion just like $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$ for 2×2 matrices.

Finally, it is useful to relate determinants of various matrices in practice. For instance, it is useful in optimization to know the relationship between the determinants of an inverse matrix before and after adding a rank-one matrix to the original matrix, which is given by the Sherman-Morrison formula. We can also use the Cramer's rule to express a solution to a linear system using the determinant of a coefficient matrix, although it is not computationally efficient in practice.

8.1 Definition and Properties

We start by defining the determinant more formally.

Definition 8.1 *We call a function* det *on $n \times n$ square matrices a determinant when it satisfies the following properties:*

1. $\det(I_n) = 1$*;*
2. *The sign of* det *flips if a pair of rows in a matrix are swapped. That is,* $\det(A) = -\det(\hat{A})$ *where $\hat{A}$ results from swapping two rows of A.*
3. *It is linear with respect to each rows of A. When a row of A can be expressed as $\mathbf{b} + \alpha\mathbf{c}$, let us construct two matrices, B and C, from A by replacing this particular row with $\mathbf{b}$ and $\mathbf{c}$, respectively. Then,* $\det(A) = \det(B) + \alpha \det(C)$.[a]

[a] We can get this result by combining the second property and that the determinant is linear with the first row of the matrix. If we want a minimal definition, we can thus change this third property to be about the first row of a matrix.

Starting from these three properties, we can derive many other properties of the determinant.

Fact 8.1 *If two rows of a square matrix A are equal, then* $\det(A) = 0$.

Proof: Because swapping these two rows does not change the matrix itself, $\det(A) = -\det(A)$, which implies that $\det(A) = 0$. ■

Fact 8.2 *Adding a multiple of one row to another row leaves the same determinant.*

Proof: Let $\mathbf{b}$ and $\mathbf{c}$ be the i-th and j-th rows of a square matrix B, respectively, with $i \neq j$. We construct two matrices from B, C by replacing the i-th row of B with $\mathbf{c}$, and A by replacing the i-th row of B with $\mathbf{b} + \alpha\mathbf{c}$. Because the i-th and j-th rows of C are same, $\det(C) = 0$, according to Fact 8.1, and according to the third property in Definition 8.1, $\det(A) = \det(B) + \alpha \det(C) = \det(B)$. ■

Fact 8.3 *If A has a row of zeros, then* $\det(A) = 0$.

Proof: We set $\alpha = -1$ and $\mathbf{b} = \mathbf{c}$ in the third property from Definition 8.1. Then, because $B = C$, the determinant of A is 0. ■

Fact 8.4 *If A is triangular, then* $\det(A)$ *is the product of the diagonal entries.*

Proof: Assume A is an upper triangular matrix. If $a_{nn} = 0$, the n-th row of A is zero, and thus $\det(A) = 0$, proving the statement. On the other hand, consider the case where $a_{nn} \neq 0$. According to Fact 8.2, the determinant does not change even if we add a scalar multiple of the last row to another row in the matrix. Because only a_{nn} is non-zero in the last row of an upper triangular matrix, we can replace all the other entries in the last column of the matrix with 0, without altering its determinant. Let such a matrix be A_{n-1}. If we repeat this procedure with $a_{(n-1)(n-1)}$, we end up with a matrix whose last two rows and columns constitute a diagonal matrix, again while keeping the determinant the same. We can repeat this procedure until we end up with a diagonal matrix D such that $\det(A) = \det(D)$. Together with the first and third properties from Definition 8.1, $\det(D)$ is the product of all diagonal entries, which proves this fact. ■

If a diagonal matrix has zero on its diagonal, its determinant is zero, according to Fact 8.4. More generally, the determinant of a non-invertible matrix is zero.

Fact 8.5 *A is invertible if and only if* $\det(A) \neq 0$.

Proof: Consider LU-decomposition of a matrix A. For an appropriate choice of a permutation matrix P, there exist a lower triangular matrix $\tilde{L}$ and an upper triangular matrix U that allow us to obtain a row-echelon form of A, and this can be expressed as $\tilde{L}PA = U$. Because $\tilde{L}$ represents adding scalar multiples of rows to other rows, $\det(\tilde{L}PA) = \det(PA)$ according to Fact 8.2. Similarly, because P represents swapping rows of A, $\det(PA) = \pm \det A$ according to the second property from Definition 8.1. Combining these two together, we get $\det A = \pm \det U$, as $\det U = \det(\tilde{L}PA) = \pm \det A$. A necessary and sufficient condition for an invertible square matrix A from the previous chapter is that all diagonal entries of the row echelon form U must be non-zero pivot elements. This condition is equivalent to $\det(U) \neq 0$, according to Fact 8.4. ■

We now investigate whether a function satisfying three conditions in Definition 8.1 exists and if so whether it is unique. Let f be a function that satisfies the second and third conditions in Definition 8.1. Since none of Fact 8.1, 8.2, and 8.3 used the first property in their derivations, we can assume that they are applicable to f. Let us construct an $(n-1) \times n$ matrix A' by taking all the rows except for the first row from A. We also construct the following $n \times n$ matrix A_j by keeping only the j-th entry of the first row while setting all the other entries to 0, that is, $A_j = \begin{bmatrix} a_{1j}\mathbf{e}_j^\top \\ A' \end{bmatrix}$. Since the first row of A is $a_{11}\mathbf{e}_1^\top + a_{12}\mathbf{e}_2^\top + \cdots + a_{1n}\mathbf{e}_n^\top$, $f(A) = f(A_1) + \cdots + f(A_n)$ according to the third condition in the definition of the determinant.

We can repeat the same procedure to the second row by creating an $(n-2) \times n$ matrix A'' by removing the first two rows from A. Then, we can construct $A_{jk} = \begin{bmatrix} a_{1j}\mathbf{e}_j^\top \\ a_{2k}\mathbf{e}_k^\top \\ A'' \end{bmatrix}$, and see that $f(A_j) = f(A_{j1}) + \cdots + f(A_{jn})$. When the sub-indices coincide, i.e. A_{jj}, we notice that $f(A_{jj}) = 0$, because $f(A_{jj}) = a_{1j}a_{2j} f \begin{bmatrix} \mathbf{e}_j^\top \\ \mathbf{e}_j^\top \\ A'' \end{bmatrix}$.

After repeating this procedure to all n rows, $A_{j_1 j_2 \ldots j_n}$ is a matrix with its k-th row being $a_{kj_k}\mathbf{e}_{j_k}^\top$, and as we have seen above, if $j_k = j_\ell$ for two rows k and ℓ, $f(A_{j_1 j_2 \ldots j_n}) = 0$, and it does not contribute to computing f. We thus need to consider $A_{j_1 j_2 \ldots j_n}$'s only for which j_k's are all different. In those cases, we get

$j_1, j_2, \ldots, j_n$ by re-ordering $1, 2, \ldots, n$, that is, as a permutation of $1, 2, \ldots, n$. If we use $\sigma(i)$ to denote the positive integer corresponding to the i-th position in a permutation, $A_{j_1 j_2 \ldots j_n} = A_{\sigma(1)\sigma(2)\ldots\sigma(n)}$. Then,

$$f(A) = \sum_{\substack{\sigma:\text{ permutation} \\ \text{of } \{1,\ldots,n\}}} f(A_{\sigma(1)\sigma(2)\ldots\sigma(n)}).$$

Let $P_\sigma = \begin{bmatrix} \mathbf{e}_{\sigma(1)}^\top \\ \vdots \\ \mathbf{e}_{\sigma(n)}^\top \end{bmatrix}$ be the permutation matrix corresponding to the permutation σ. Although we do not prove it in this book, there are many ways to swap rows to turn P_σ into the identity matrix I, but the corresponding numbers of row swaps share the parity. That is, they are either all odd or all even. With this fact, we define a sign function on permutation to return 1 for even permutation and -1 for odd permutation. Then,

$$f(A_{\sigma(1)\sigma(2)\ldots\sigma(n)}) = f\begin{bmatrix} a_{1\sigma(1)}\mathbf{e}_{\sigma(1)}^\top \\ a_{2\sigma(2)}\mathbf{e}_{\sigma(2)}^\top \\ \vdots \\ a_{n\sigma(n)}\mathbf{e}_{\sigma(n)}^\top \end{bmatrix} = a_{1\sigma(1)}a_{2\sigma(2)}\cdots a_{n\sigma(n)} f\begin{bmatrix} \mathbf{e}_{\sigma(1)}^\top \\ \mathbf{e}_{\sigma(2)}^\top \\ \vdots \\ \mathbf{e}_{\sigma(n)}^\top \end{bmatrix}$$

$$= \quad a_{1\sigma(1)}a_{2\sigma(2)}\cdots a_{n\sigma(n)} f(P_\sigma) \;=\; \operatorname{sign}(\sigma)\, a_{1\sigma(1)}a_{2\sigma(2)}\cdots a_{n\sigma(n)} f(I).$$

This allows us to write $f(A)$ as

$$f(A) = \sum_{\substack{\sigma:\text{ permutation} \\ \text{of } \{1,\ldots,n\}}} \operatorname{sign}(\sigma)\, a_{1\sigma(1)}a_{2\sigma(2)}\cdots a_{n\sigma(n)} f(I). \tag{8.1}$$

We arrived at this expression by using the second and third properties from Definition 8.1 only, and the determinant satisfies this equation as well. Conversely, any f defined as in (8.1) satisfies the second and third properties of Definition 8.1.

Theorem 8.1 *The function satisfying Definition 8.1 is unique.*

Proof: Because f in (8.1) already satisfies the second and third conditions in Definition 8.1, f is a determinant as long as $f(I) = 1$. That is, by defining $f(I) = 1$, there exists at least one function that satisfies all three conditions in Definition 8.1, to which we refer by det.

Let g be a function that satisfies both the second and third conditions in Definition 8.1. We define h for any $n \times n$ matrix A by

$$h(A) = g(A) - \det(A)g(I).$$

Then, $h(I) = g(I) - \det(I)g(I) = 0$. Because both g and det satisfy the second and third conditions, so does h. We can thus expand h in the form of (8.1) by

$$h(A) = \sum_{\substack{\sigma: \text{permutation} \\ \text{of } \{1,\dots,n\}}} \operatorname{sign}(\sigma)\, a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} h(I).$$

In this form, $h(A) = 0$ for any A, because $h(I) = 0$. That is, $g(A) = \det(A)g(I)$ holds for all A. If $g(I) = 1$, g and det coincide, implying that there is only one function that satisfies all three properties in Definition 8.1.

■

In addition to the existence and uniqueness of the determinant, we also obtained the following expanded form of (8.1) in this proof:

$$\det A = \sum_{\substack{\sigma: \text{permutation} \\ \text{of } \{1,\dots,n\}}} \operatorname{sign}(\sigma)\, a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}. \tag{8.2}$$

As described in Appendix B, σ^{-1} is also a permutation if σ is a permutation. Furthermore, $\operatorname{sign}(\sigma) = \operatorname{sign}(\sigma^{-1})$ since $\sigma^{-1} \circ \sigma$ is the identity permutation whose sign is clearly even. Applying these to the summand in (8.2), we get

$$\operatorname{sign}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)} = \operatorname{sign}(\sigma^{-1}) \prod_{j=1}^{n} a_{\sigma^{-1}(j)j}.$$

Combining this with a fact $\{\sigma : \sigma \text{ is a permutation}\} = \{\sigma^{-1} : \sigma \text{ is a permutation}\}$, we get the following alternative expression:

$$\det A = \sum_{\substack{\sigma: \text{permutation} \\ \text{of } \{1,\dots,n\}}} \operatorname{sign}(\sigma)\, a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}\,. \tag{8.3}$$

These expansions are useful later when for instance we compute the determinant of a small matrix or derive some theoretical results on matrices. By observing $a_{\sigma(i)i} = (A^\top)_{i\sigma(i)}$ in the expansion (8.3), we also get the following result on the determinant of transposed matrices.

Fact 8.6 *For a square matrix A, $\det(A^\top) = \det(A)$. That is, the determinant and transpose commute.*

Proof: Expanding $\det A^\top$ along (8.2) results in (8.3). ■

Thanks to Fact 8.6, all conditions and properties of the determinant in terms of rows can be rephrased in terms of columns. For example, a column swapping changes the sign of the determinant, and determinant function is linear with respect to each column.

Theorem 8.1 allows us to observe an important rule to compute the determinant of matrix products.

Fact 8.7 *For two square matrix A and B of same size,* $\det(AB) = \det(A)\det(B)$.

Proof: Let us fix two matrices A and B. If B is non-invertible, the equality holds as both sides are trivially 0. Otherwise, $\det B \neq 0$. Consider a real-valued function ρ for any $n \times n$ square matrix C, defined as follows:

$$\rho(C) = \frac{\det(CB)}{\det(B)}.$$

1. If $C = I$, $\rho(I) = \frac{\det(IB)}{\det(B)} = 1$;
2. Since each row of CB is the product of the row of C and the matrix B, by swapping the i-th and j-th rows of C, the corresponding rows in CB are also swapped. That is, the sign of $\det(CB)$ flips. Since the sign of $\det(B)$ is maintained, the sign of $\rho(C)$ flips;
3. When a row of C is $\mathbf{b}^\top + \alpha\mathbf{c}^\top$, the corresponding row of CB is $\mathbf{b}^\top B + \alpha\mathbf{c}^\top B$. Since the denominator does not change, ρ is linear with respect to each row.

This new function ρ satisfies all three conditions in Definition 8.1, implying that ρ is the matrix determinant. In other words, $\rho(C) = \det(C)$ for all C. Thus, when $C = A$,

$$\rho(A) = \frac{\det(AB)}{\det(B)} = \det(A),$$

from which we see that the determinant of the product of two matrices is equal to the product of the determinants of two matrices.

■

The transpose and determinant commute, and so do the inverse and determinant. If A were invertible, $\det(A) \neq 0$, according to Fact 8.5. Using $AA^{-1} = I$ with Fact 8.7, we get

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

Based on these results in this section, we now know a few ways to compute the determinant of a matrix. First, we can use LU decomposition of a matrix. Let $\Pi A = LU$. Because L has a unit diagonal, $\det L = 1$. Because Π is a permutation matrix, $\det \Pi$ is either 1 or -1. Because U is upper-triangular, $\det U$ is simply the product of all diagonal entries. Then, $\det A = \underbrace{\det \Pi}_{-1 \text{ or } 1} \det U$. Another approach is based on QR decomposition and is particular useful for computing the absolute value of $\det A$. As the determinant of an orthogonal matrix is either 1 or -1. We simply need to compute $\det R$ as the product of all diagonal entries, and it corresponds to the absolute determinant of the original matrix A. These approaches are often used in practice to compute the determinant of a matrix numerically.

8.2 Formulas for Determinant

Let us use (8.2) to compute the determinant of a small matrix. Consider first a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Because $(a, b) = (a, 0) + (0, b)$ and $(c, d) = (c, 0) + (0, d)$, we can compute the determinant of this matrix, following

$$\begin{aligned} \det(A) &= \det \begin{bmatrix} a & 0 \\ c & d \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ c & d \end{bmatrix} \\ &= \det \begin{bmatrix} a & 0 \\ c & 0 \end{bmatrix} + \det \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ 0 & d \end{bmatrix} \\ &= \quad 0 \quad + \det \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} - \det \begin{bmatrix} c & 0 \\ 0 & b \end{bmatrix} + \quad 0 \\ &= ad - bc. \end{aligned}$$

Along this line, let us try to compute the determinant of the following 3×3 matrix:

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

This can be done as follows:

$$\det(A) = \det \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det \begin{bmatrix} 0 & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det \begin{bmatrix} 0 & 0 & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$
\begin{aligned}
= \quad & \det\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & 0 & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ 0 & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & 0 & a_{13} \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & 0 & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\
= \quad & \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & 0 & a_{23} \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & 0 & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & 0 \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ a_{21} & 0 & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & 0 & a_{13} \\ a_{21} & 0 & 0 \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ a_{21} & 0 & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \\
& + \det\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & a_{22} & 0 \\ a_{31} & 0 & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & a_{22} & 0 \\ 0 & a_{32} & 0 \end{bmatrix} + \det\begin{bmatrix} 0 & 0 & a_{13} \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \\
= \quad & a_{11}a_{22}a_{33} \times \det\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + a_{11}a_{23}a_{32} \times \det\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \\
& + a_{12}a_{21}a_{33} \times \det\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + a_{12}a_{23}a_{31} \times \det\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
& +a_{13}a_{21}a_{32} \times \det \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + a_{13}a_{22}a_{31} \times \det \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \\
= & \sum_{\substack{\sigma: \text{ permutation} \\ \text{of } \{1,2,3\}}} a_{1\sigma(1)}a_{2\sigma(2)}a_{3\sigma(3)} \det(P_\sigma) \\
= & \ a_{11}a_{22}a_{33} \times 1 + a_{11}a_{23}a_{32} \times (-1) + a_{12}a_{21}a_{33} \times (-1) \\
& +a_{12}a_{23}a_{31} \times (-1)^2 + a_{13}a_{21}a_{32} \times (-1)^2 + a_{13}a_{22}a_{31} \times (-1) \\
= & \ a_{11}\big(a_{22}a_{33} - a_{23}a_{32}\big) - a_{12}\big(a_{21}a_{33} - a_{23}a_{31}\big) + a_{13}\big(a_{21}a_{32} - a_{22}a_{31}\big) \\
= & \ a_{11} \times \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12} \times \det \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + a_{13} \times \det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \\
= & \ \sum_{j=1}^{3} (-1)^{1+j} a_{1j} \det(A_{1j}) \,.
\end{aligned}
$$

In general, A_{1j} above is generally an $(n-1) \times (n-1)$ submatrix of an $n \times n$ matrix A after removing the first row and the j-th column of A. Observe that the red-colored terms above cancel out or vanish to zero, and only the blue-colored terms may remain non-zero.

The last line above generalizes this computation to an $n \times n$ matrix from the 3×3 matrix as

$$\sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}) \,,$$

though we will not discuss any explicit proof here. While taking this as correct, let us try to prove the following result showing that this expansion could be done with any arbitrary row rather than the first one:

$$\sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}) \,.$$

Similarly to A_{1j}, A_{ij} is an $(n-1) \times (n-1)$ submatrix of A after removing the i-th row and j-th column. If we use $\hat{A}$ to denote a matrix resulting from swapping the first and i-th row of A, $\det(A) = -\det(\hat{A})$. We now let $\hat{A}_{1j}$ be the $(n-1) \times (n-1)$ submatrix of $\hat{A}$ after removing the first row and the j-th column, in fact, which corresponds to rearranging all rows but the i-th one in A in the order of $2, \ldots, (i-1), 1, (i+1), \ldots, n$ and removing the j-th column. The first row of A is the $(i-1)$-th row in $\hat{A}_{1j}$. If we swap this row $(i-2)$ times to move it to the top of $\hat{A}_{1j}$, we end up with A_{ij}. As the sign of the determinant

flips every time a pair of rows swap, $\det(\hat{A}_{1j}) = (-1)^{i-2}\det(A_{ij})$. This results in the desired expansion:

$$\begin{aligned}\det(\hat{A}) &= \sum_{j=1}^{n}(-1)^{1+j}a_{ij}\det(\hat{A}_{1j}) = \sum_{j=1}^{n}(-1)^{1+j}a_{ij}(-1)^{i-2}\det(A_{ij}) \\ &= \sum_{j=1}^{n}(-1)^{i+j-1}a_{ij}\det(A_{ij}).\end{aligned}$$

Using a shorthand notation $C_{ij} = (-1)^{i+j}\det(A_{ij})$, to which we refer as a cofactor, we get the following cofactor expansion of the determinant given any i:

$$\det(A) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(A_{ij}) = \sum_{j=1}^{n}a_{ij}C_{ij}\,.$$

Fact 8.8 *Let A be an $n \times n$ square matrix. For any i,*

$$\det(A) = \sum_{j=1}^{n}a_{ij}C_{ij} = \sum_{j=1}^{n}a_{ji}C_{ji}\,.$$

Proof: We already showed that $\det(A) = \sum_{j=1}^{n}a_{ij}C_{ij}$. If we apply this to $A^\top$, $(-1)^{i+j}\det(A^\top{}_{ij}) = (-1)^{i+j}\det(A_{ji}{}^\top) = (-1)^{i+j}\det(A_{ji}) = C_{ji}$. Therefore,

$$\det(A^\top) = \sum_{j=1}^{n}(A^\top)_{ij}(-1)^{i+j}\det(A^\top{}_{ij}) = \sum_{j=1}^{n}a_{ji}C_{ji}\,.$$

■

This tells us that cofactor expansion does not only work for an arbitrary row but also for an arbitrary column. If there is a mismatch between the element a_{kj} and the cofactor C_{ij}, that is, $k \neq i$, this expansion vanishes to 0.

Fact 8.9 $\sum_{j=1}^{n}a_{kj}C_{ij} = 0$ *for $i \neq k$.*

Proof: We construct B from A by replacing the i-th row with the k-th row. In other words, the i-th and k-th rows of B are the same, meaning that $\det(B) = 0$. All the entries of A and B are the same except for the i-th row, and thus the cofactors of the i-th row also coincide with each other. That is, the B's cofactor is also C_{ij}. Then, the cofactor expansion of B with respect to the i-th row is

$$\det(B) = \sum_{j=1}^{n}b_{ij}C_{ij} = \sum_{j=1}^{n}a_{kj}C_{ij} = 0,$$

which proves the statement. ∎

We use these facts, 8.8 and 8.9, later to derive the inverse of a matrix in terms of its determinant and cofactors, in (8.8).

Example 8.1 Let us compute the determinant of (the second difference matrix)

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

For a product L of appropriate elementary matrices,

$$LA = U = \begin{bmatrix} 2 & -1 & & & \\ & \frac{3}{2} & -1 & & \\ & & \ddots & \ddots & \\ & & & \frac{n}{n-1} & -1 \\ & & & & \frac{n+1}{n} \end{bmatrix},$$

and therefore, $\det A = 2 \times \frac{3}{2} \times \cdots \times \frac{n}{n-1} \times \frac{n+1}{n} = n+1$ since $\det L = 1$. ∎

8.2.1 Determinant of Block Matrix

Consider the determinant of the following $(n_1 + n_2) \times (n_1 + n_2)$ square block matrix consisting of A_{11} and A_{22} of sizes $n_1 \times n_1$ and $n_2 \times n_2$, respectively:

$$A = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix}.$$

If both of these blocks were of size 1×1, as in $\begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}$, the determinant would be $a_{11}a_{22} = \det([a_{11}]) \det([a_{22}])$. From this observation, it is natural to deduce

$$\det(A) = \det(A_{11}) \det(A_{22}).$$

Let σ be a permutation of $\{1, \ldots, n_1 + n_2\}$, and σ_1 be a permutation where only the first n_1 indices are permuted and the rest $\{n_1 + 1, \ldots, n_1 + n_2\}$ are

maintained (as an identity). On the other hand, σ_2 is a permutation that only permutes the latter n_2 indices $\{n_1+1,\ldots,n_1+n_2\}$, while keeping the rest $\{1,\ldots,n_1\}$ as they are. When P_σ is the permutation matrix of σ, $\mathrm{sign}(\sigma) = \det P_\sigma$. We make one important observation here. $\sigma_1 \circ \sigma_2$ is a permutation over $\{1,\ldots,n_1+n_2\}$, although the first n_1 and the latter n_2 are permuted within each other only, implying that $\sigma_1 \circ \sigma_2 = \sigma_2 \circ \sigma_1$. This also implies that such a permutation must be expressible in the form of $\sigma_1 \circ \sigma_2$. Furthermore, any permutation σ, that cannot be expressed as $\sigma_1 \circ \sigma_2$, must map either $i \le n_1$ to $\sigma(i) > n_1$ or $i > n_1$ to $\sigma(i) \le n_1$. Since such a pair $(i, \sigma(i))$ corresponds to an entry in A outside A_{11} and A_{22}, the block diagonal structure imposes $a_{i\sigma(i)} = 0$. We can hence limit ourselves to only permutations in the form of $\sigma_1 \circ \sigma_2$ to compute the determinant of A.

Let $\mathcal{S}$ be a set of permutations of $\{1,\ldots,n_1+n_2\}$, $\mathcal{S}_1$ a set of permutations of $\{1,\ldots,n_1\}$, and $\mathcal{S}_2$ a set of permutations of $\{n_1+1,\ldots,n_1+n_2\}$. Then, we get the following expression for the determinant of A:

$$\begin{aligned}
\det(A) &= \sum_{\sigma\in\mathcal{S}} a_{1\sigma(1)}\cdots a_{(n_1+n_2)\sigma(n_1+n_2)}\,\mathrm{sign}(\sigma) \\
&= \sum_{\sigma_1\in\mathcal{S}_1,\sigma_2\in\mathcal{S}_2} a_{1\sigma_1\circ\sigma_2(1)}\cdots a_{(n_1+n_2)\sigma_1\circ\sigma_2(n_1+n_2)}\,\mathrm{sign}(\sigma_1\circ\sigma_2) \\
&= \sum_{\sigma_1\in\mathcal{S}_1}\sum_{\sigma_2\in\mathcal{S}_2} a_{1\sigma_1(1)}\cdots a_{n_1\sigma_1(n_1)} a_{(n_1+1)\sigma_2(n_1+1)}\cdots a_{(n_1+n_2)\sigma_2(n_1+n_2)} \\
&\qquad \times \mathrm{sign}(\sigma_1)\,\mathrm{sign}(\sigma_2) \\
&= \sum_{\sigma_1\in\mathcal{S}_1} a_{1\sigma_1(1)}\cdots a_{n_1\sigma_1(n_1)}\,\mathrm{sign}(\sigma_1) \\
&\qquad \times \sum_{\sigma_2\in\mathcal{S}_2} a_{(n_1+1)\sigma_2(n_1+1)}\cdots a_{(n_1+n_2)\sigma_2(n_1+n_2)}\,\mathrm{sign}(\sigma_2) \\
&= \det(A_{11})\det(A_{22}).
\end{aligned}$$

Next, consider the following block lower triangular matrix:

$$A = \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{bmatrix}.$$

Taking into account that the determinant of a regular matrix is expressed as the product of diagonal entries, we can make an educated guess that that of the block matrix must be

$$\det(A) = \det(A_{11})\det(A_{22}).$$

If $\det(A_{11}) = 0$, $\text{rank}(A_{11}) < n_1$, implying that $\text{rank}(A) < n_1 + n_2$ and $\det(A) = 0$. In this case, the expression above holds. When A_{11} is invertible, we can perform Gaussian elimination on A, as follows:

$$\begin{bmatrix} I_{11} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix}.$$

Because the first matrix on the left hand side is a lower triangular matrix with unit diagonals, the determinant is 1. We already showed earlier that the determinant of the right hand side is $\det(A_{11})\det(A_{22})$. Combining these two, we see that the equation above holds for the block lower triangular matrix.

Finally, consider a general block matrix, where A_{11} is invertible. We can perform Gaussian elimination to remove A_{21} as follows:

$$\begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I_{11} & A_{11}^{-1}A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}.$$

If we consider the determinants of both sides of the equation, we get

$$\det(A_{11}^{-1})\det(A) = \det(A_{22} - A_{21}A_{11}^{-1}A_{12}).$$

Since the determinant of the inverse is the inverse of the determinant of the original matrix, we arrive at

$$\det(A) = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12}), \tag{8.4}$$

where $A_{22} - A_{21}A_{11}^{-1}A_{12}$ is the Schur complement.

8.2.2 Matrix Determinant Lemma

We derive the matrix determinant lemma from the results on the determinant of a block matrix above. From there on, we continue to deriving Sherman-Morrison, or Woodbury formula that explains how the inverse changes when we modify an invertible matrix by adding a product of lower rank matrices, in Section 8.5.

Although there is no intuitive way to explain why we choose these three matrices, let us consider the product of the following three matrices.

$$\begin{bmatrix} I_n & \mathbf{0} \\ V^\top & I_k \end{bmatrix} \begin{bmatrix} I_n + UV^\top & U \\ \mathbf{0} & I_k \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -V^\top & I_k \end{bmatrix},$$

where U and V are both $n \times k$ matrices. Because the product of the first two matrices is

$$\begin{bmatrix} I_n + UV^\top & U \\ V^\top + V^\top UV^\top & I_k + V^\top U \end{bmatrix},$$

the product of all three matrices is

$$\begin{bmatrix} I_n + UV^\top & U \\ V^\top + V^\top UV^\top & I_k + V^\top U \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -V^\top & I_k \end{bmatrix} = \begin{bmatrix} I_n & U \\ \mathbf{0} & I_k + V^\top U \end{bmatrix}.$$

As the determinants of the original expression and the final expression must coincide, we get

$$\det(I_n + UV^\top) = \det(I_k + V^\top U), \tag{8.5}$$

from which we derive the following general result.

Theorem 8.2 (Matrix Determinant Lemma) *Let A be an $n \times n$ invertible matrix, and U and V be $n \times k$ matrices. Then,*

$$\det(A + UV^\top) = \det(A)\det(I_k + V^\top A^{-1}U)\,. \tag{8.6}$$

Proof: Since $A + UV^\top = A(I_n + A^{-1}UV^\top)$, we get

$$\begin{aligned} \det(A + UV^\top) &= \det(A(I_n + A^{-1}UV^\top)) \\ &= \det(A)\det(I_n + A^{-1}UV^\top) \\ &= \det(A)\det(I_k + V^\top A^{-1}U) \quad \text{by plugging } A^{-1}U \text{ into } U \text{ in (8.5)}\,. \end{aligned}$$

■

One potential advantage of using the matrix determinant lemma is the reduction of computational effort when $k < n$, and the determinant and inverse of A are known in advance. This is because the size of $I_k + V^\top A^{-1}U$ is $k \times k$ and is smaller than $A + UV^\top$ which is $n \times n$.

In addition to using this result for computing the determinant, it is also used often to show that the inverse of $A + UV^\top$ exists only when $I_k + V^\top A^{-1}U$ is invertible. When $k = 1$, U and V are respectively $\mathbb{R}^n$ vectors, $\mathbf{u}$ and $\mathbf{v}$, and the matrix determinant formula simplifies to

$$\det(A + \mathbf{u}\mathbf{v}^\top) = \det(A)\det(1 + \mathbf{v}^\top A^{-1}\mathbf{u}). \tag{8.7}$$

Therefore, for an invertible A, $A + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $\mathbf{v}^\top A^{-1}\mathbf{u} \neq -1$.

8.3 Volume of parallelopiped in Euclidean Space

Consider n vectors, $\mathbf{a}_1, \ldots, \mathbf{a}_n$, in $\mathbb{R}^n$, and let us compute the volume V of the n-dimensional parallelopiped defined by the origin and these vectors. We assume these vectors are linearly independent, as otherwise the volume vanishes.

Let us start with an assumption that $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are orthogonal, and let $A = [\mathbf{a}_1 \mid \ldots \mid \mathbf{a}_n]$. Since these vectors are orthogonal, the volume V corresponds to the product of their magnitudes/norms. That is, $V = \prod_{i=1}^n |\mathbf{a}_i|$. By Fact 8.6, we know that

$$\det(A)^2 = \det(A^\top A) = \prod_{i=1}^n |\mathbf{a}_i|^2 = V^2 \iff V = \big|\det(A)\big|,$$

because

$$A^\top A = \begin{bmatrix} |\mathbf{a}_1|^2 & 0 & 0 & \cdots & 0 \\ 0 & |\mathbf{a}_2|^2 & 0 & \cdots & 0 \\ 0 & 0 & |\mathbf{a}_3|^2 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & |\mathbf{a}_n|^2 \end{bmatrix}.$$

We now relax the orthogonality constraint on $\mathbf{a}_1, \ldots, \mathbf{a}_n$ so that they are linearly independent but not necessarily orthogonal. We use QR decomposition to re-write A as QR, where Q is an orthogonal matrix and R is an upper triangular matrix. Once we computed the volume of the $(i-1)$-dimensional parallelopiped defined by $\{\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}\}$ together with the origin in the $(i-1)$-dimensional subspace $\text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{i-1}\}$, the contribution to the volume of the i-dimensional parallelopiped by $\mathbf{a}_i$ is by its orthogonal component to $\text{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{i-1}\}$. Since the i-th column of R is the coordinate of $\mathbf{a}_i$ under the basis $\{\mathbf{q}_1, \ldots, \mathbf{q}_i\}$, the absolute value of $(R)_{ii} = \langle \mathbf{a}_i, \mathbf{q}_i \rangle$ is precisely the contribution by $\mathbf{a}_i$ to the parallelopiped's volume in the i-dimensional subspace. Thus, according to Fact 8.4, we know that

$$V = \Big| \prod_{i=1}^n (R)_{ii} \Big| = |\det(R)|.$$

Since $A^\top A = R^\top Q^\top Q R = R^\top R$, $\det(A)^2 = \det(R)^2$. Therefore, $V = |\det(A)|$. In other words, the volume of n-dimensional parallelopiped is just the determinant of a matrix whose columns are the edges of the parallelopiped, up to its sign.

8.4 Closed-Form Expressions using Determinant

In this section, we consider two applications of the determinant; matrix inverse and linear systems. These applications reveal interesting insights into the relationship between the determinant and other practically useful quantities in linear algebra. They are however not used in practice due to their inefficiency.

8.4.1 Closed-Form Expression for Matrix Inverse

One observation we draw by considering a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is that

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{\det(A)} \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix},$$

where C_{ij} is a cofactor. If we multiply A to its left,

$$\frac{1}{\det(A)} A \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix} = \frac{1}{\det(A)} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{bmatrix} = I.$$

Let $C = (C_{ij})$ be a cofactor matrix. Then, we see that

$$A^{-1} = \frac{1}{\det(A)} C^\top,$$

for an 2×2 matrix A. We generalize this result to an arbitrary invertible square matrix by using the cofactor expansion with respect to an arbitrary row i,

$$\det(A) = \sum_{j=1}^{n} a_{ij} C_{ij} = [a_{i1}, \ldots, a_{in}][C_{i1}, \ldots, C_{in}]^\top,$$

and Fact 8.9, which results in

$$\begin{aligned} AC^\top &= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} C_{11} & C_{21} & \ldots & C_{n1} \\ C_{12} & C_{22} & \ldots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \ldots & C_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \det(A) & 0 & \ldots & 0 \\ 0 & \det(A) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \det(A) \end{bmatrix} \end{aligned}$$

$$= \det(A)I.$$

In short,

$$A^{-1} = \frac{1}{\det(A)} C^\top \tag{8.8}$$

where C is a cofactor matrix.

8.4.2 Cramer's Rule: Closed-Form Solution of Linear System

When A is an invertible matrix, we can use (8.8) to derive the solution to $A\mathbf{x} = \mathbf{b}$ as follows:

$$\begin{aligned} \mathbf{x} &= A^{-1}\mathbf{b} \\ &= \frac{1}{\det(A)} C^\top \mathbf{b} \\ &= \frac{1}{\det(A)} \Big(\sum_{i=1}^{n} C_{ij} b_i \Big). \end{aligned}$$

If we replace the j-th column of A with $\mathbf{b}$, the determinant is $\sum_{i=1}^{n} C_{ij} b_i$. This allows us to derive the following Cramer's rule:

$$x_j = \frac{1}{\det(A)} \det \big[\mathbf{a}_1 \,|\, \dots \,|\, \mathbf{a}_{j-1} \,|\, \mathbf{b} \,|\, \mathbf{a}_{j+1} \,|\, \dots \,|\, \mathbf{a}_n \big].$$

The Cramer's rule is concise and is useful for mathematical reasoning, but is not computationally efficient enough for solving real-world linear systems.

8.5 Sherman-Morrison and Woodbury Formulas

The matrix determinant lemma allows us to progressively compute the matrix determinant by multiplying the matrix determinants of smaller matrices. In this section, we derive a similar approach to progressively computing the inverse of a matrix by inverting smaller matrices, called the Sherman-Morrison and Woodbury formulas.

Consider the inverse of $A + UV^\top$, where A is an $n \times n$ invertible matrix, and U and V are $n \times k$ matrices. We want to express this inverse using the inverse of A, which is assumed to be known. According to the matrix determinant lemma (8.6), the inverse of $A + UV^\top$ exists only when $I_k + V^\top A^{-1} U$ is invertible.

Let us start with the inverse of $I_n + UV^\top$. Note that $I_n + UV^\top$ is invertible if and only if $I_k + VU^\top$ is invertible. Try $I_n + UBV^\top$ as the inverse with an

appropriate choice of a $k \times k$ matrix B (though, it is definitely not intuitive how we decide to start from here.) Then,

$$\begin{aligned}(I_n + UV^\top)(I_n + UBV^\top) &= I_n + UV^\top + UBV^\top + UV^\top UBV^\top \\ &= I_n + U\big(I_k + B + V^\top UB\big)V^\top,\end{aligned}$$

which tells us that a sufficient condition for the invertibility of $I_n + UV^\top$ is $I_k + B + V^\top UB = \mathbf{0}$. Then, from

$$-I_k = B + V^\top UB = (I_k + V^\top U)B\,,$$

we get

$$B = -(I_k + V^\top U)^{-1}.$$

According to the matrix determinant lemma, the invertibility of $I_k + V^\top U$ is the necessary and sufficient condition for the invertibility of $I_n + UV^\top$. That is, $I_k + V^\top U$ is invertible, and by plugging it into the original form $I_n + UBV^\top$, we get

$$\big(I_n + UV^\top\big)^{-1} = I_n - U\big(I_k + V^\top U\big)^{-1}V^\top. \tag{8.9}$$

Now, let us derive the inverse of $A + UV^\top$:

$$\begin{aligned}\big(A + UV^\top\big)^{-1} &= \big(A(I_n + A^{-1}UV^\top)\big)^{-1} \\ &= (I_n + A^{-1}UV^\top)^{-1}A^{-1} \\ &= \big(I_n - A^{-1}U(I_k + V^\top A^{-1}U)^{-1}V^\top\big)A^{-1} \\ &= A^{-1} - A^{-1}U\big(I_k + V^\top A^{-1}U\big)^{-1}V^\top A^{-1}.\end{aligned}$$

According to this rule, we can compute the inverse of an $n \times n$ matrix by only computing the inverse of a smaller $k \times k$ matrix, if we know A^{-1} already.

Theorem 8.3 (Woodbury Formula) *Let A be an $n \times n$ invertible matrix, and U and V be $n \times k$ matrices. Then, $I_k + V^\top A^{-1}U$ is invertible if and only if $A + UV^\top$ is invertible. Then,*

$$\big(A + UV^\top\big)^{-1} = A^{-1} - A^{-1}U\big(I_k + V^\top A^{-1}U\big)^{-1}V^\top A^{-1}\,. \tag{8.10}$$

Proof: It is clear from the matrix determinant lemma (8.6) that these two conditions are necessary and sufficient conditions of each other. We already derived the inverse above.

■

According to this theorem, the invertibility of $I_k + V^\top A^{-1}U$ is a necessary and sufficient condition for the invertibility of $A+UV^\top$, once we know that A is invertible. It is important to consider the case of $k = 1$. In that case, U and V are $\mathbb{R}^n$ vectors, $\mathbf{u}$ and $\mathbf{v}$, respectively, and the necessary and sufficient condition for the invertibility of $A + \mathbf{u}\mathbf{v}^\top$ is $\mathbf{v}^\top A^{-1}\mathbf{u} \neq -1$.

Corollary 8.1 (Sherman-Morrison Formula) *Let A be an $n \times n$ invertible matrix, and $\mathbf{u}$ and $\mathbf{v}$ be $\mathbb{R}^n$-vectors. Then, $1 + \mathbf{v}^\top A^{-1}\mathbf{u} \neq 0$ if and only if $A + \mathbf{u}\mathbf{v}^\top$ is invertible. Then,*

$$\left(A + \mathbf{u}\mathbf{v}^\top\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}}\,. \tag{8.11}$$

The significance of the Sherman-Morrison formula is that we can compute the inverse of a matrix obtained by adding a rank-one matrix to the original matrix by a combination of matrix-vector multiplications, when we knew the invers of the original matrix. This progressive approach is computationally advantageous, as we will see in the next section.

8.6 Application: Rank-one Update of Inverse Hessian

In machine learning, the process of learning is often implemented as optimization. That is, it is the process of iteratively finding a set of parameters that minimizes or maximizes the learning objective function. Let us use $\mathbf{w} \in \mathbb{R}^n$ to denote the parameter vector. A representative example is the parameters of a neural network from Section 3.10. We use $f(\mathbf{w})$ to denote the learning objective function, which is often referred to as a loss function. We prefer it to be smaller, which means that learning corresponds to solving $\mathbf{w}^* = \text{argmin}_{\mathbf{w}}\, f(\mathbf{w})$. Among numerous algorithms that have been proposed to solve this problem, we consider a symmetric rank-one update based algorithm.

Most nonlinear optimization algorithms aim to find $\mathbf{w}^*$ that makes the gradient vanishes to $\mathbf{0}$, i.e., $\nabla f(\mathbf{w}) = \mathbf{0}$. In doing so, consider the first-order derivative (the gradient vector ∇f) and the second-order derivative (the Hessian matrix

$\nabla^2 f$):

$$\nabla f = \begin{bmatrix} \frac{\partial}{\partial w_1} f \\ \vdots \\ \frac{\partial}{\partial w_n} f \end{bmatrix}, \quad H = \nabla^2 f = \begin{bmatrix} \frac{\partial}{\partial w_1}\frac{\partial}{\partial w_1} f & \cdots & \frac{\partial}{\partial w_n}\frac{\partial}{\partial w_1} f \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_1}\frac{\partial}{\partial w_n} f & \cdots & \frac{\partial}{\partial w_n}\frac{\partial}{\partial w_n} f \end{bmatrix}.$$

The second-order Taylor approximation to f near $\mathbf{w}_k$ is given as

$$f(\mathbf{w}) \approx f(\mathbf{w}_k) + \nabla f(\mathbf{w}_k)^\top (\mathbf{w} - \mathbf{w}_k) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_k)^\top H(\mathbf{w}_k)(\mathbf{w} - \mathbf{w}_k)\,.$$

Because $\frac{\partial}{\partial w_i}\frac{\partial}{\partial w_j} f(\mathbf{w}) = \frac{\partial}{\partial w_j}\frac{\partial}{\partial w_i} f(\mathbf{w})$ for most of functions f, $H(\mathbf{w})$ is symmetric, and we can approximate the gradient by

$$\nabla f(\mathbf{w}) \approx \nabla f(\mathbf{w}_k) + H(\mathbf{w}_k)(\mathbf{w} - \mathbf{w}_k),$$

according to Fact 4.19. In order to find $\mathbf{w}$ that satisfies $\nabla f(\mathbf{w}) = \mathbf{0}$, we solve the following secant condition

$$\nabla f(\mathbf{w}_k) + H(\mathbf{w}_k)(\mathbf{w} - \mathbf{w}_k) = \mathbf{0}$$

and get $\mathbf{w} = \mathbf{w}_k - H(\mathbf{w}_k)^{-1}\nabla f(\mathbf{w}_k)$.

Since we started from approximation, $\nabla f(\mathbf{w}) \neq \mathbf{0}$ in general, and we thus iteratively compute next-step vectors $\mathbf{w}_{k+1}$ until the iteration converges to the desired solution $\mathbf{w}^*$, as follows:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - H(\mathbf{w}_k)^{-1}\nabla f(\mathbf{w}_k)\,. \tag{8.12}$$

We call this procedure the Newton-Raphson method. Under suitable conditions, we can show that $\mathbf{w}_k$ converges to $\mathbf{w}^*$, and the convergence rate is fast if we can compute all quantities exactly. It however becomes computationally infeasible to compute the inverse of the Hessian matrix every time as the number of parameters n grows. Here, we thus seek a similar iterative algorithm that does not assume access to the Hessian matrix $H(\mathbf{w}_k)$ but only to the gradient.

First, we replace $H(\mathbf{w}_k)$ with a similar, symmetric matrix B_k in (8.12), resulting in

$$\mathbf{w}_{k+1} = \mathbf{w}_k - B_k^{-1}\nabla f(\mathbf{w}_k). \tag{8.13}$$

After we determine $\mathbf{w}_{k+1}$ with this rule, we find B_{k+1} that closely approximates the second-order derivative $H(\mathbf{w}_{k+1})$ using the gradients at $\mathbf{w}_k$ and $\mathbf{w}_{k+1}$, by solving

$$\nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k) = B_{k+1}(\mathbf{w}_{k+1} - \mathbf{w}_k). \tag{8.14}$$

Because this system has n equations while the number of entries of $n \times n$ symmetric matrix to be determined is order of n^2, there are many solutions of B_{k+1}. We call optimization algorithms that use any of these solutions collectively as quasi-Newton methods.

Among these quasi-Newton methods is the symmetric rank-one method, where we simply add a rank-one matrix to B_k to obtain B_{k+1}. The update rule for B_{k+1} in this method is

$$B_{k+1} = B_k + \frac{1}{(\mathbf{d}_k - B_k \Delta \mathbf{w}_k)^\top \Delta \mathbf{w}_k} (\mathbf{d}_k - B_k \Delta \mathbf{w}_k)(\mathbf{d}_k - B_k \Delta \mathbf{w}_k)^\top, \quad (8.15)$$

where $\Delta \mathbf{w}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$ and $\mathbf{d}_k = \nabla f(\mathbf{w}_{k+1}) - \nabla f(\mathbf{w}_k)$. Show yourself that B_{k+1} satisfies (8.14).

Because we use (8.13) to update the parameters at each iteration, we need to know B_{k+1}^{-1}. We can use the Sherman-Morrison formula (8.11) and get the following update rule to compute B_{k+1}^{-1} directly from B_k^{-1} without B_{k+1}:

$$B_{k+1}^{-1} = B_k^{-1} - \frac{1}{(\Delta \mathbf{w}_k - B_k^{-1} \mathbf{d}_k)^\top \mathbf{d}_k} (\Delta \mathbf{w}_k - B_k^{-1} \mathbf{d}_k)(\Delta \mathbf{w}_k - B_k^{-1} \mathbf{d}_k)^\top. \quad (8.16)$$

Once we know B_k^{-1}, we only need to add a rank-one matrix to efficiently compute B_{k+1}^{-1}. In practice, it is usual to maintain and update B_k^{-1} directly rather than B_k. It is however more convenient to use (8.15) for analyzing mathematically the update rule.

Chapter 9

Further Results on Eigenvalues and Eigenvectors

Although we introduced eigenvalues and eigenvectors in Definition 5.1, we waited until this chapter to delve deeper into these two concepts, as it requires the notion of determinants.

What is a simple interpretable transformation? As we have seen in Section 3.8.2, scaling transformations would be the simplest one. Then, for the linear transformation corresponding to a matrix A, is there any way to interpret the matrix through scaling transformations? It is difficult to get such an interpretation on the whole space at once, but there may be hope if we search for a one-dimensional subspace on which the transformation can be understood as a scaling. This amounts to finding a scalar scaling factor λ and a vector $\mathbf{v}$ such that $A = \lambda I$ on a subspace spanned by a single non-zero vector $\mathbf{v}$, that is, for vectors $x\mathbf{v}$ in the one-dimensional subspace, $A(x\mathbf{v}) = (\lambda I)(x\mathbf{v}) = x\lambda\mathbf{v}$ holds. This relation is simply $A\mathbf{v} = \lambda\mathbf{v}$ and we say λ is an eigenvalue and $\mathbf{v}$ an eigenvector as well as $(\lambda, \mathbf{v})$ an eigenpair. There can be more than one eigenvectors per eigenvalue.[1] For each eigenvalue, we can span a subspace by all associated eigenvectors, on which the transformation works as a scaling by the eigenvalue. This story still holds for complex matrices with complex eigenvalues

[1] We refer to the maximum number of linearly independent eigenvectors for an eigenvalue as the geometric multiplicity.

and eigenvectors.

It is difficult to find both eigenvalue and eigenvector at once since the term $\lambda\mathbf{v}$ is not linear in unknowns. So, we first find eigenvalues and then search for eigenvectors. When $A-\lambda I$ is not invertible, that is $\det(A-\lambda I)=0$, $A\mathbf{v}=\lambda\mathbf{v}$ admits non-zero solution vector $\mathbf{v}$. This allows us to find an eigenvalue by treating λ as a variable and solving $\det(A-\lambda I)=0$. We call this equation a characteristic equation, which is useful for theoretical derivation and abstract reasoning. This is however not helpful in computing eigenvalues of a general matrix in practice. With an eigenvalue λ in hand, we find eigenvectors as a basis of $\mathrm{Null}\,(A-\lambda I)$.

An n-th order complex polynomial equation admits n roots, including simple and multiple roots. Based on this (though we do not prove it here), every $n\times n$ matrix has n eigenvalues, again including multiple roots. In the appendix, we summarize a minimal set of results on complex numbers needed for studying this chapter.

In the rest of the chapter, we present the spectral decomposition of a real symmetric matrix in Section 5.5 once more, as this is the most popular eigen-decomposition result in applications. In Chapter 3, we discussed that a linear transformation has a corresponding matrix representation given a basis of a vector space. When we can build a basis consisting only of the eigenvectors of such a matrix, the same linear transformation under this basis corresponds to a diagonal matrix. We refer to the process of finding such a diagonal matrix by diagonalization. Although diagonalization does not work for all matrices, we will study in Chapter 11 that each linear transformation admits a corresponding Jordan-form matrix under the choice of an appropriate basis, where a Jordan form refers to a diagonal matrix or a matrix similar to a diagonal matrix.

9.1 Examples of Eigendecomposition

We can study various cases and properties of eigenpairs using a 2×2 matrix.

1. A real matrix with two real eigenvalues: When $A=\begin{bmatrix}3 & 2\\ -2 & -2\end{bmatrix}$, $\det(A-\lambda I)=(3-\lambda)(-2-\lambda)+4=\lambda^2-\lambda-2=(\lambda-2)(\lambda+1)$. The eigenvalues are thus $\lambda=2$ and -1, and their corresponding eigenvectors are $(2,-1)^\top$ and $(1,-2)^\top$, respectively, because $A-\lambda I=\begin{bmatrix}1 & 2\\ -2 & -4\end{bmatrix}$ and $\begin{bmatrix}4 & 2\\ -2 & -1\end{bmatrix}$.

2. A real diagonal matrix: When $A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$, the eigenvalues are $\lambda = a$ and b, because $\det(A - \lambda I) = (a - \lambda)(b - \lambda)$.

   - $a \neq b$: the eigenvectors are $(1, 0)^\top$ and $(0, 1)^\top$, because $A - \lambda I = \begin{bmatrix} 0 & 0 \\ 0 & b - a \end{bmatrix}$ and $\begin{bmatrix} a - b & 0 \\ 0 & 0 \end{bmatrix}$.
   - $a = b$: Any arbitrary pair of linearly independent vectors can be eigenvectors, since $A - \lambda I = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

3. A real upper triangular matrix: When $A = \begin{bmatrix} a & c \\ 0 & b \end{bmatrix}$, $c \neq 0$, the eigenvalues are $\lambda = a$ and b, because $\det(A - \lambda I) = (a - \lambda)(b - \lambda)$.

   - $a \neq b$: The eigenvectors are $(1, 0)^\top$ and $(1, \frac{b-a}{c})^\top$, because $A - \lambda I = \begin{bmatrix} 0 & c \\ 0 & b - a \end{bmatrix}$ and $\begin{bmatrix} a - b & c \\ 0 & 0 \end{bmatrix}$.
   - $a = b$: The eigenvector can be any scalar multiple of $(1, 0)^\top$, because $A - \lambda I = \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix}$. That is, this matrix has only one eigenvector. This is an example of a real asymmetric matrix that does not possess as many eigenvectors as the number of roots (algebraic multiplicity).

4. A real matrix with two complex eigenvalues: When $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, the eigenvalues are $\lambda = \mathtt{i}$ and $-\mathtt{i}$, because $\det(A - \lambda I) = \lambda^2 + 1$. The corresponding eigenvectors are $(-\mathtt{i}, 1)^\top$ and $(\mathtt{i}, 1)^\top$, respectively, since $A - \lambda I = \begin{bmatrix} \mathtt{i} & -1 \\ 1 & \mathtt{i} \end{bmatrix}$ and $\begin{bmatrix} -\mathtt{i} & -1 \\ 1 & -\mathtt{i} \end{bmatrix}$.

5. Eigendecomposition of a non-trivial projection matrix P satisfying $P^2 = P$ (neither I nor $\mathbf{0}$): Assume an eigenpair $(\lambda, \mathbf{v})$ satisfying $P\mathbf{v} = \lambda\mathbf{v}$. By multiplying both sides with P, we get $P^2\mathbf{v} = \lambda P\mathbf{v}$ which simplifies to $P\mathbf{v} = \lambda^2\mathbf{v}$, since $P^2 = P$ and $P\mathbf{v} = \lambda\mathbf{v}$. Thus, $\lambda = 0$ or 1, as $(\lambda^2 - \lambda)\mathbf{v} = \mathbf{0}$. Since a non-trivial projection matrix satisfies $0 < \operatorname{rank}(P) < n$, any vector in the null space of P is the eigenvector corresponding to the eigenvalue 0. Since a vector $\mathbf{v}$ already projected onto a subspace satisfies $P\mathbf{v} = \mathbf{v}$, such a vector is the eigenvector corresponding to the eigenvalue of 1.

From the examples above, we can deduce that eigenpairs can come in many different ways.

9.2 Properties of Eigenpair

We start with the linear independence of eigenvectors associated with distinct eigenvalues.

Fact 9.1 *If eigenvectors* $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *correspond to different eigenvalues* $\lambda_1, \ldots, \lambda_k$, *then the eigenvectors are linearly independent.*

Proof: Suppose that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is linearly dependent. Then, there exist $c_1, \ldots, c_k \in \mathbb{R}$ with at least one of them being a non-zero scalar, that satisfies $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$. As we can permute the order within $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, we assume $c_1 \neq 0$ without loss of generality. By multiplying both side of the equation by A, we get $A(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) = c_1\lambda_1\mathbf{v}_1 + \cdots + c_k\lambda_k\mathbf{v}_k = \mathbf{0}$. We then multiply the original equation with λ_k and subtract it from this new equation to get $(c_1\lambda_1\mathbf{v}_1 + \cdots + c_{k-1}\lambda_{k-1}\mathbf{v}_{k-1} + c_k\lambda_k\mathbf{v}_k) - (c_1\lambda_k\mathbf{v}_1 + \cdots + c_{k-1}\lambda_k\mathbf{v}_{k-1} + c_k\lambda_k\mathbf{v}_k) = c_1(\lambda_1 - \lambda_k)\mathbf{v}_1 + \cdots + c_{k-1}(\lambda_{k-1} - \lambda_k)\mathbf{v}_{k-1} = \mathbf{0}$. Since $c_1(\lambda_1 - \lambda_k) \neq 0$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$ is also linearly dependent. If we repeat this argument $k-2$ times more, we end up with $c_1'\mathbf{v}_1 = \mathbf{0}$ where $c_1' = c_1(\lambda_1 - \lambda_k)\cdots(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_2)$. Then, $c_1' = 0$ and $c_1 = 0$, which is contradictory. Therefore, $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is linearly independent. ■

It is also interesting that the determinant and the trace of a matrix can be expressed only in terms of its eigenvalues.

Fact 9.2 *Let* $A = (a_{ij})$ *be an* $n \times n$ *matrix and* $\lambda_1, \ldots, \lambda_n$ *be its eigenvalues. Then,*

- $\det(A) = \lambda_1\lambda_2\cdots\lambda_n = \prod_{i=1}^n \lambda_i$;
- $\mathrm{trace}(A) = a_{11} + a_{22} + \cdots + a_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^n \lambda_i$.

Proof:

According to (8.2), $\det(A - \lambda I)$ is an n-th order polynomial of λ, and the coefficient of its n-th term is $(-1)^n$. Since the roots of this polynomial are eigenvalues,

$$\det(A - \lambda I) = (-1)^n(\lambda - \lambda_1)\cdots(\lambda - \lambda_n) = (\lambda_1 - \lambda)\cdots(\lambda_n - \lambda). \qquad (9.1)$$

- If we set $\lambda = 0$ in (9.1), $\det(A) = \lambda_1 \cdots \lambda_n$.
- The entries of $A - \lambda I$ that contain λ are only on the n diagonal entries, $a_{ii} - \lambda$. Let us do the cofactor expansion of $B = A - \lambda I$ along the first row. Consider the summand $b_{1i} \det(B_{1i})$ for $i \geq 2$ where B_{1i} contains $n-2$ diagonal entries[2] including λ except for $a_{11} - \lambda$ and $a_{ii} - \lambda$. Therefore, the degree of λ in $\det(B_{1i})$ is at most $n-2$ since B_{1i} contains at most $n-2$ entries including λ. Therefore, the λ^{n-1} is contained only in $b_{11} \det(B_{11}) = (a_{11} - \lambda) \det(B_{11})$. If we repeat the same argument to identify λ^{n-2} in $\det(B_{11})$ and so forth, we conclude that λ^{n-1} appears only in $(a_{11} - \lambda) \cdots (a_{nn} - \lambda)$ in the expansion (8.2) of $\det(A - \lambda I)$. Here, the coefficient of λ^{n-1} is $(-1)^{n-1}(a_{11} + \cdots + a_{nn})$, and hence $\text{trace}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$ since the coefficient of λ^{n-1} in $\det(A - \lambda I)$ is also $(-1)^{n-1}(\lambda_1 + \cdots + \lambda_n)$ by expanding (9.1).

■

Because we define the eigenvalue through determinants, some properties of determinant have counterpart observations for the eigenvalue. For example, since the determinants of A and $A^\top$ are the same, $\det(A - \lambda I) = \det(A^\top - \lambda I)$, and therefore, the eigenvalues of A and $A^\top$ are also shared.

Fact 9.3 *The eigenvalues of A and $A^\top$ coincide.*

By algebraically manipulating $A\mathbf{v} = \lambda\mathbf{v}$ in various ways, we can derive properties of eigenpairs. Let $(\lambda, \mathbf{v})$ be an eigenpair of A. Then, $(\lambda^2, \mathbf{v})$ is an eigenpair of A^2, because $A^2\mathbf{v} = A(A\mathbf{v}) = A(\lambda\mathbf{v}) = \lambda A\mathbf{v} = \lambda^2\mathbf{v}$.

Fact 9.4 *If $(\lambda, \mathbf{v})$ is an eigenpair of A, then $(\lambda^2, \mathbf{v})$ is an eigenpair of A^2.*

When A is invertible and has an eigenpair $(\lambda, \mathbf{v})$, $\lambda \neq 0$ from the invertibility of A and $A^{-1}\mathbf{v} = \lambda^{-1}\mathbf{v}$ by multiplying $\lambda^{-1}A^{-1}$ to both sides of $A\mathbf{v} = \lambda\mathbf{v}$. Therefore, $(\lambda^{-1}, \mathbf{v})$ is an eigenpair of A^{-1}.

Fact 9.5 *If A is invertible and $(\lambda, \mathbf{v})$ is an eigenpair of A, then $\lambda \neq 0$ and $(\lambda^{-1}, \mathbf{v})$ is an eigenpair of A^{-1}.*

Interestingly, if a complex matrix is Hermitian,[3] which includes the case of a real symmetric matrix, all eigenvalues are real.

[2] Recall that B_{ij} is a submatrix built by removing i-th row and j-th column.

[3] We define the Hermitian matrix and also discuss various results on complex numbers in Appendix E for your review.

Lemma 9.1 *Let A be a Hermitian matrix. Then, all eigenvalues of A are real. Furthermore, if A is a real symmetric matrix, not only real eigenvalues but also real eigenvectors exist.*

Proof: Let A be Hermitian and $(\lambda, \mathbf{v})$ an eigenpair of A. Recall that $|\mathbf{v}| \neq 0$. From $A\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v}^{\mathsf{H}}A\mathbf{v} = \lambda\mathbf{v}^{\mathsf{H}}\mathbf{v} = \lambda|\mathbf{v}|^2$. So,

$$(\lambda|\mathbf{v}|^2)^{\mathsf{H}} = \lambda^{\mathsf{H}}|\mathbf{v}|^2 = (\mathbf{v}^{\mathsf{H}}A\mathbf{v})^{\mathsf{H}} = \mathbf{v}^{\mathsf{H}}A^{\mathsf{H}}(\mathbf{v}^{\mathsf{H}})^{\mathsf{H}} = \mathbf{v}^{\mathsf{H}}A\mathbf{v} = \lambda|\mathbf{v}|^2, \text{ that is, } \lambda^{\mathsf{H}} = \lambda$$

which implies that λ is real by Fact E.1 in Appendix E. In addition, let us assume that A is real and $(\lambda, \mathbf{v} + \mathtt{i}\mathbf{w})$ be an eigenpair where λ is real, and $\mathbf{v}$ and $\mathbf{w}$ are real n-vectors. Then, the eigenpair satisfies

$$A\mathbf{v} + \mathtt{i}A\mathbf{w} = A(\mathbf{v} + \mathtt{i}\mathbf{w}) = \lambda(\mathbf{v} + \mathtt{i}\mathbf{w}) = \lambda\mathbf{v} + \mathtt{i}\lambda\mathbf{w}$$

which implies $(\lambda, \mathbf{v})$ and $(\lambda, \mathbf{w})$ are also eigenpairs. ■

Similarly to the linear independence of eigenvectors for distinct eigenvalues in Fact 9.1, if a matrix is Hermitian, eigenvectors of distinct eigenvalues are not only linearly independent but also orthogonal. Recall Fact 4.4 saying that the orthogonality of vectors implies their linear independence.

Fact 9.6 *Let A be Hermitian. If eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ correspond to different eigenvalues λ_1 and λ_2, those eigenvectors are orthogonal.*

Proof: By Lemma 9.1, the eigenvalues are real. From $A\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $A\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, $\mathbf{v}_2^{\mathsf{H}}A\mathbf{v}_1 = \lambda_1\mathbf{v}_2^{\mathsf{H}}\mathbf{v}_1$ and $\mathbf{v}_1^{\mathsf{H}}A\mathbf{v}_2 = \lambda_2\mathbf{v}_1^{\mathsf{H}}\mathbf{v}_2$. Since $\lambda_1 \neq \lambda_2$ and

$$0 = (\mathbf{v}_1^{\mathsf{H}}A\mathbf{v}_2)^{\mathsf{H}} - \mathbf{v}_2^{\mathsf{H}}A\mathbf{v}_1 = (\lambda_2\mathbf{v}_1^{\mathsf{H}}\mathbf{v}_2)^{\mathsf{H}} - \lambda_1\mathbf{v}_2^{\mathsf{H}}\mathbf{v}_1 = (\lambda_2 - \lambda_1)\mathbf{v}_2^{\mathsf{H}}\mathbf{v}_1\,,$$

we conclude $\mathbf{v}_2^{\mathsf{H}}\mathbf{v}_1 = 0$. ■

If a matrix is orthogonal, we can observe that the absolute values of the eigenvalues are all 1, and the eigenvectors associated with distinct eigenvalues are orthogonal as for Hermitian matrices.

Fact 9.7 *Let Q be a real orthogonal matrix and $(\lambda, \mathbf{v})$ an eigenpair of Q. λ and $\mathbf{v}$ may be complex-valued. Then*

1. *$|\lambda| = 1$.*
2. *$(\lambda^{-1}, \mathbf{v})$ is an eigenpair of $Q^{\top}$.*
3. *If eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ correspond to different eigenvalues λ_1 and λ_2, then those eigenvectors are orthogonal.*

Proof: Recall that $Q^\top Q = I$ and $Q\mathbf{v} = \lambda\mathbf{v}$.

1. $|\mathbf{v}|^2 = \mathbf{v}^\mathsf{H}\mathbf{v} = \mathbf{v}^\mathsf{H}(Q^\top Q)\mathbf{v} = (Q\mathbf{v})^\mathsf{H}(Q\mathbf{v}) = (\lambda\mathbf{v})^\mathsf{H}(\lambda\mathbf{v}) = |\lambda|^2|\mathbf{v}|^2$ and $|\lambda|^2 = 1$.
2. $\mathbf{v} = (Q^\top Q)\mathbf{v} = Q^\top(\lambda\mathbf{v}) = \lambda Q^\top\mathbf{v}$. So, $Q^\top\mathbf{v} = \lambda^{-1}\mathbf{v}$.
3. Let $Q\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $Q\mathbf{v}_2 = \lambda_2\mathbf{v}_2$. Then,

$$\mathbf{v}_1^\mathsf{H}\mathbf{v}_2 = \mathbf{v}_1^\mathsf{H}(Q^\top Q)\mathbf{v}_2 = (Q\mathbf{v}_1)^\mathsf{H}(Q\mathbf{v}_2) = \lambda_1^\mathsf{H}\lambda_2\mathbf{v}_1^\mathsf{H}\mathbf{v}_2$$

and we must have $\lambda_1^\mathsf{H}\lambda_2 = \overline{\lambda_1}\lambda_2 = 1$ or $\mathbf{v}_1^\mathsf{H}\mathbf{v}_2 = 0$. However, $1 = |\lambda_1|^2 = \lambda_1^\mathsf{H}\lambda_1 = \overline{\lambda_1}\lambda_1$ implies $\overline{\lambda_1}\lambda_2 = \frac{\lambda_2}{\lambda_1} \neq 1$ since $\lambda_1 \neq \lambda_2$. Therefore, $\mathbf{v}_1^\mathsf{H}\mathbf{v}_2 = 0$.

■

Fact 9.8 *If A is a real matrix and $(\lambda, \mathbf{v})$ is an eigenpair of A, then $(\overline{\lambda}, \overline{\mathbf{v}})$ is also an eigenpair of A, where $\overline{\lambda}$ and $\overline{\mathbf{v}}$ are the complex conjugates of λ and $\mathbf{v}$, respectively.*

Proof: By taking the conjugation on both sides of $A\mathbf{v} = \lambda\mathbf{v}$, we get $\overline{A\mathbf{v}} = \overline{\lambda\mathbf{v}}$. By complex arithmetic, we see that $\overline{A}\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Since A is a real matrix, $\overline{A} = A$, and thus $A\overline{\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$. Therefore, $(\overline{\lambda}, \overline{\mathbf{v}})$ is also an eigenpair.

■

Using the results from Section 7.5, we can express the volume of an arbitrary n-dimensional ellipsoid as the multiple of the volume of a unit sphere. We further characterize this in terms of the determinant of the positive definite matrix inducing the ellipsoid.

Fact 9.9 *For a positive definite matrix A, an n-dimensional ellipsoid is given as $\mathcal{E} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top A^{-1}\mathbf{x} \leq 1\}$. Then,*

$$\mathbf{vol}(\mathcal{E}) = \sqrt{\det(A)}\,\mathbf{vol}(B_n)\,. \tag{9.2}$$

Proof: By Fact 9.2, $\prod_{i=1}^n \sqrt{\lambda_i(A)} = \sqrt{\det(A)}$, and therefore, $\mathbf{vol}(\mathcal{E}) = \sqrt{\det(A)} \times \mathbf{vol}(B_n)$.

■

9.3 Similarity and Change of Basis

We say that two matrices are similar when they represent the same linear transformation under appropriate choices of bases.

9.3.1 Change of Basis

Given a basis $\mathcal{B}_1$ of a vector space $\mathbb{V}$, we want to consider another basis $\mathcal{B}_2$. In order to translate the properties expressed by the basic vectors in $\mathcal{B}_1$ into the descriptions in terms of basic vectors in $\mathcal{B}_2$, we must first establish the relationship between $\mathcal{B}_1$ and $\mathcal{B}_2$, using the ideas developed in Section 3.8.1.

Let $\mathcal{B}_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\mathcal{B}_2 = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. Since $\mathcal{B}_2$ is a basis itself, we can express each $\mathbf{v}_j$ as the linear combination of the basic vectors of $\mathcal{B}_2$, as follows

$$\mathbf{v}_j = \sum_{i=1}^{n} b_{ij}\mathbf{w}_i. \tag{9.3}$$

With this, let us try to compute the coordinate vector $\mathbf{y} \in \mathbb{R}^n$ of the vector $\mathbf{v}$ with respect to $\mathcal{B}_2$, when its coordinate vector under $\mathcal{B}_1$ is $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, that is,

$$\mathbf{v} = \sum_{j=1}^{n} x_j\mathbf{v}_j.$$

Plugging in (9.3), we get

$$\mathbf{v} = \sum_{j=1}^{n} x_j\Big(\sum_{i=1}^{n} b_{ij}\mathbf{w}_i\Big) = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} b_{ij}x_j\Big)\mathbf{w}_i. \tag{9.4}$$

With a square matrix $B = (b_{ij})$, we can write the relationship between $\mathbf{x}$ and $\mathbf{y}$ as

$$y_i = \sum_{j=1}^{n} b_{ij}x_j \quad \text{or} \quad \mathbf{y} = B\mathbf{x}. \tag{9.5}$$

Since b_{ij} relates the i-th and j-th basic vectors from two bases, (9.5) holds for all vectors with the $B = (b_{ij})$ as long as we fix the two bases.

We can swap the roles of $\mathcal{B}_1$ and $\mathcal{B}_2$ and repeat the argument above to obtain another square matrix $\hat{B}$ that maps $\mathbf{y}$ to $\mathbf{x}$, that is $\mathbf{x} = \hat{B}\mathbf{y}$. Together with (9.5),

$$\mathbf{x} = \hat{B}B\mathbf{x} \;\text{ for all } \mathbf{x} \in \mathbb{R}^n \;\text{ and }\; \mathbf{y} = B\hat{B}\mathbf{y} \;\text{ for all } \mathbf{y} \in \mathbb{R}^n,$$

which implies that $\hat{B}B = B\hat{B} = I_n$. In summary, the matrix that represents the change of basis is invertible, and two matrices that represent the mapping between two bases are inverses of each other.

Conversely, we can build a new basis $\mathcal{B}' = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ satisfying (9.3) once a basis $\mathcal{B} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ and an arbitrary invertible matrix $B = (b_{ij})$ are given. It is enough to check whether $\mathcal{B}'$ is linearly independent to confirm that it is

really a basis. Assume $\sum_{j=1}^{n} \lambda_j \mathbf{v}_j = \mathbf{0}$ with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top \in \mathbb{R}^n$. Then,

$$\sum_{j=1}^{n} \lambda_j \mathbf{v}_j = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{n} b_{ij} \mathbf{w}_i = \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} b_{ij} \lambda_j \Big) \mathbf{w}_i = \mathbf{0}$$

holds, and hence $\sum_{j=1}^{n} b_{ij} \lambda_j = 0$ for all i, that is, $B\boldsymbol{\lambda} = \mathbf{0}$ since $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are linearly independent. Therefore, $\boldsymbol{\lambda} = B^{-1}\mathbf{0} = \mathbf{0}$ and $\mathcal{B}'$ is linearly independent and a basis.

Example 9.1 [Change of Orthonormal Basis] Consider the case where both bases $\mathcal{B}_1$ and $\mathcal{B}_2$ are orthonormal. Since $\mathbf{w}_i^\top \mathbf{w}_\ell$ is 1 when $i = \ell$ and otherwise 0, we get

$$\begin{aligned} \mathbf{v}_j^\top \mathbf{v}_k &= \Big(\sum_{i=1}^{n} b_{ij} \mathbf{w}_i \Big)^\top \Big(\sum_{\ell=1}^{n} b_{\ell k} \mathbf{w}_\ell \Big) = \Big(\sum_{i=1}^{n} b_{ij} \mathbf{w}_i^\top \Big) \Big(\sum_{\ell=1}^{n} b_{\ell k} \mathbf{w}_\ell \Big) \\ &= \sum_{i=1}^{n} \sum_{\ell=1}^{n} b_{ij} b_{\ell k} \mathbf{w}_i^\top \mathbf{w}_\ell \\ &= \sum_{i=1}^{n} b_{ij} b_{ik}. \end{aligned}$$

$\mathbf{v}_j^\top \mathbf{v}_k$ is also 1 only when $j = k$ and otherwise 0. Because $\sum_{i=1}^{n} b_{ij} b_{ik}$ is the (j,k)-th entry of $B^\top B$,

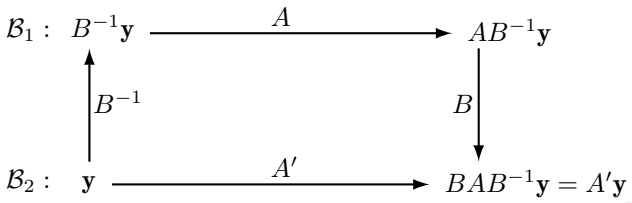
$$B^\top B = I_n.$$

In other words, B is an orthogonal matrix, and the change of basis matrix between two orthonormal bases is orthogonal. ∎

9.3.2 Similarity

Consider a linear transformation $T : \mathbb{V} \to \mathbb{V}$ defined in a vector space $\mathbb{V}$ and its two bases $\mathcal{B}_1$ and $\mathcal{B}_2$. Let A be the transformation matrix of T with respect to $\mathcal{B}_1$ such that the $\mathcal{B}_1$-coordinate of $T(\mathbf{v})$ is $A\mathbf{x}$ when $\mathbf{v}$'s $\mathcal{B}_1$-coordinate is $\mathbf{x}$. In addition, B is the change of basis matrix between $\mathcal{B}_1$ and $\mathcal{B}_2$. We now derive the matrix representation of T with respect to $\mathcal{B}_2$.

According to (9.5), the $\mathcal{B}_1$-coordinate of a vector whose $\mathcal{B}_2$-coordinate is $\mathbf{y}$ is $B^{-1}\mathbf{y}$. And, the $\mathcal{B}_1$-coordinate of vector $\mathbf{y}$ after transformation by T is then $AB^{-1}\mathbf{y}$. The $\mathcal{B}_2$-coordinate of a vector whose $\mathcal{B}_1$-coordinate is $AB^{-1}\mathbf{y}$ is $BAB^{-1}\mathbf{y}$.



If we let A' be the transformation matrix of T under $\mathcal{B}_2$, we arrive at $A'\mathbf{y} = BAB^{-1}\mathbf{y}$ applies to all $\mathbf{y} \in \mathbb{R}^n$, which implies

$$A' = BAB^{-1} \quad \text{and} \quad A = B^{-1}A'B\,.$$

As A and BAB^{-1} represent the same transformation T under two different bases, respectively, we can view them as equivalent in terms of how a vector transforms in $\mathbb{V}$. We thus say they are similar.

Definition 9.1 *Square matrices A and A' are similar if there is an invertible matrix B such that $A = B^{-1}A'B$.*

Fact 9.10 *If A and A' are similar, they share the same eigenvalues.*

Proof: Suppose that $A = B^{-1}A'B$ and $(\lambda, \mathbf{v})$ is an eigenpair of A satisfying $A\mathbf{v} = \lambda\mathbf{v}$. Then,

$$A'(B\mathbf{v}) = (A'B)\mathbf{v} = (BA)\mathbf{v} = B(A\mathbf{v}) = B(\lambda\mathbf{v}) = \lambda(B\mathbf{v})$$

holds, and implies that $(\lambda, B\mathbf{v})$ is an eigenpair of A'. ■

9.4 Diagonalization

A diagonal matrix is considered the simplest form of a matrix. If we can find a basis with respect to which a transformation matrix is diagonal by a change of basis, this can help us understand and analyze the corresponding linear transformation. When this is possible, we call such a transformation matrix diagonalizable.

Definition 9.2 *A is diagonalizable if A is similar to a diagonal matrix.*

We first investigate the most basic necessary and sufficient condition for a diagonalizable matrix.

Fact 9.11 *An $n \times n$ matrix A is diagonalizable if and only if A has n linearly independent eigenvectors.*

Proof: Define a diagonal matrix Λ as

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$

- **if**: Let $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_n, \mathbf{v}_n)$ be n eigenpairs with linearly independent $\mathbf{v}_i$'s, although some eigenvalues may coincide. With invertible $B = \begin{bmatrix} \mathbf{v}_1 \mid \cdots \mid \mathbf{v}_n \end{bmatrix}$ and $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$, we know $AB = B\Lambda$ and that $B^{-1}AB = \Lambda$. Thus, A is diagonalizable.
- **only if**: Because A is diagonalizable, there exist an invertible B and a diagonal matrix Λ such that $A = B^{-1}\Lambda B$. Because then $AB^{-1} = B^{-1}\Lambda$, each column of B^{-1} is an eigenvector. Since B^{-1} is invertible, its columns are linearly independent.

■

We can ask when the eigenvectors of two matrices coincide perfectly. There is no known result for two $n \times n$ matrices in general. We however know that the necessary and sufficient condition for two diagonalizable matrices, A and B, to share their eigenvectors is for them to commute, that is, $AB = BA$.

Theorem 9.1 *Let A and B be $n \times n$ diagonalizable matrices. Then, A and B share the same eigenvectors if and only if $AB = BA$.*

Proof: Let A and B share the same eigenvectors and denote the square matrix of those shared eigenvectors as S. Then $A = S^{-1}\Lambda_1 S$ and $B = S^{-1}\Lambda_2 S$. Since diagonal matrices commute,

$$AB = S^{-1}\Lambda_1 S S^{-1}\Lambda_2 S = S^{-1}\Lambda_1\Lambda_2 S = S^{-1}\Lambda_2\Lambda_1 S = S^{-1}\Lambda_2 S S^{-1}\Lambda_1 S = BA\,.$$

Conversely, assume that diagonalizable matrices A and B commute. Let S be an invertible matrix of eigenvectors of A such that $A = S^{-1}\Lambda S$. Set $B = S^{-1}\hat{B}S$. The commutativity of A and B implies the commutativity of Λ and $\hat{B}$. Assume $\Lambda = \operatorname{diag}(\lambda_1 I_{n_1}, \ldots, \lambda_k I_{n_k})$ where each I_{n_i} is an identity matrix of size n_i such that $n = n_1 + \cdots + n_k$. Of course, λ_i's are assumed to be different

from each other. Denote the matrix $\hat{B}$ by a partitioned matrix $(\hat{B}_{ij})$ where $\hat{B}_{ij}$ is an $n_i \times n_j$ matrix. Then, since

$$\Lambda\hat{B} = \begin{bmatrix} \lambda_1\hat{B}_{11} & \lambda_1\hat{B}_{12} & \cdots & \lambda_1\hat{B}_{1k} \\ \lambda_2\hat{B}_{21} & \lambda_2\hat{B}_{22} & \cdots & \lambda_2\hat{B}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_k\hat{B}_{k1} & \lambda_k\hat{B}_{k2} & \cdots & \lambda_k\hat{B}_{kk} \end{bmatrix} \text{ and } \hat{B}\Lambda = \begin{bmatrix} \lambda_1\hat{B}_{11} & \lambda_2\hat{B}_{12} & \cdots & \lambda_k\hat{B}_{1k} \\ \lambda_1\hat{B}_{21} & \lambda_2\hat{B}_{22} & \cdots & \lambda_k\hat{B}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1\hat{B}_{k1} & \lambda_2\hat{B}_{k2} & \cdots & \lambda_k\hat{B}_{kk} \end{bmatrix},$$

$\Lambda\hat{B} = \hat{B}\Lambda$ holds if and only if $\hat{B}_{ij} = \mathbf{0}$ for $i \neq j$. That is, $\hat{B} = \text{diag}(\hat{B}_{11}, \ldots, \hat{B}_{kk})$. From the diagonalizability of B, $\hat{B}$ is also diagonalizable, and each $\hat{B}_{ii}$ is diagonalizable in turn. Let some $\hat{S}_i$ satisfy $\hat{B}_{ii} = \hat{S}_i^{-1}\hat{\Lambda}_i\hat{S}_i$, where $\hat{\Lambda}_i$ is a diagonal matrix. Denote $\hat{S} = \text{diag}(\hat{S}_1, \ldots, \hat{S}_k)$ and $\hat{\Lambda} = \text{diag}(\hat{\Lambda}_1, \ldots, \hat{\Lambda}_k)$ such that $\hat{B} = \hat{S}^{-1}\hat{\Lambda}\hat{S}$. By recalling the common block structure of $\hat{S}$ and Λ, we also observe $\Lambda = \hat{S}^{-1}\Lambda\hat{S}$. If we combine these, we obtain

$$A = S^{-1}\Lambda S = S^{-1}\hat{S}^{-1}\Lambda\hat{S}S \quad \text{and} \quad B = S^{-1}\hat{B}S = S^{-1}\hat{S}^{-1}\hat{\Lambda}\hat{S}S.$$

Therefore, A and B share the common matrix $(\hat{S}S)^{-1}$ of eigenvectors. ■

According to the real spectral theorem (Theorem 5.2), a symmetric matrix is diagonalizable. Furthermore, since symmetric matrices satisfy $(AB)^\top = B^\top A^\top = BA$, $AB = BA$ is equivalent to the symmetry of AB when A and B are both symmetric. Combining these two, we get the following result.

Corollary 9.1 *Let A and B be $n \times n$ symmetric matrices. Then, A and B share the same eigenvector matrix if and only if AB is symmetric.*

When the eigenvectors of a diagonalizable matrix are orthogonal, we can find orthonormal eigenvectors by diagonalization. When this is possible, we call such a matrix orthogonally diagonalizable.

Definition 9.3 *A square matrix A is orthogonally diagonalizable if there exists an orthogonal matrix Q such that $Q^{-1}AQ = Q^\top AQ$ is diagonal.*

When an $n \times n$ matrix A is orthogonally diagonalizable, there exists an orthogonal matrix Q, i.e. $Q^{-1} = Q^\top$ that diagonalizes A by $Q^{-1}AQ = \Lambda$, where $Q = [\,\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\,]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Not all λ_i's may be different. Recall (3.8) in Corollary 3.1, and we see that we can rewrite an orthogonally diagonalizable matrix A as a sum of rank-one matrices induced from orthonormal

vectors, as follows:

$$A = Q\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top. \tag{9.6}$$

We can see that such an orthogonally diagonalizable matrix is symmetric. Conversely, spectral decomposition tells us that a symmetric matrix is orthogonally diagonalizable as well. Combining these two observations, we learn that this property is unique for symmetric matrices.

Theorem 9.2 (Fundamental Theorem of Symmetric Matrices) *A real matrix A is orthogonally diagonalizable if and only if A is symmetric.*

Invertibility and Diagonalizability in the Lens of Eigenpairs

Every $n \times n$ matrix has n (including multiple roots) complex eigenvalues. From these eigenvalues, we can make the following observations.

- If all eigenvalues are non-zero, the matrix is invertible;
- If there are n linearly independent eigenvectors, the matrix is diagonalizable. When an eigenvalue λ is a multiple root of the characteristic equation $\det(A - xI)$, this equation contains $(x - \lambda)^k$ with $k > 1$, and we say the algebraic multiplicity is k. Although we are not proving it here, the number of linearly independent eigenvectors corresponding to each λ, to which we refer as geometric multiplicity, is at most k. The sum of the algebraic multiplicities of all eigenvalues coincides with the number of either rows or columns of the matrix. A matrix is diagonalizable if the sum of the geometric multiplicities matches the number of columns.

An Example of a Non-diagonalizable Matrix

Let us consider $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. As its characteristic polynomial is $p_A(x) = (x-1)^2$, A has one eigenvalue $\lambda = 1$, and its algebraic multiplicity is 2. The null space of A has dimension of 1, since $A - \lambda I = A - I = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $2 - \operatorname{rank} A = 1$. In other words, A has one eigenvector, and the geometric multiplicity of the eigenvalue 1 is 1. Since the geometric multiplicity is smaller than the algebraic multiplicity, this matrix is not diagonalizable.

Let us show that this matrix is not diagonalizable without relying on the relationship between two types of multiplicity. Assume the existence of an invertible matrix $B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ that diagonalizes A, that is, $B^{-1}AB$ is a diagonal matrix. Because $B^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, $B^{-1}AB = \frac{1}{ad-bc}\begin{bmatrix} ad+cd-bc & d^2 \\ -c^2 & ad-cd-bc \end{bmatrix}$. For the latter to be diagonal, both c and d must be 0, in which case B is not invertible. This contradicts the invertibility assumption, and therefore A is not diagonalizable. Although it is not diagonalizable, we will see later in Section 11.4 that we can express such a non-diagonalizable matrix as an upper triangular matrix that closely resembles a diagonal matrix. We call such an upper triangular matrix the Jordan form.

9.5 Spectral Decomposition Theorem

A real $n \times n$ matrix A may have complex eigenvalues, even if it contains purely real entries. A Hermitian matrix on the other hand only has real eigenvalues, according to Lemma 9.1. Furthermore, a real symmetric matrix, which is a special case of a Hermitian matrix, does not only have real eigenvalues but also n real eigenvectors that are orthonormal. These results and observations were already implied in Theorem 5.2.

(The Real Spectral Decomposition Theorem revisited) *Let A be a real symmetric matrix. Then, A is orthogonally diagonalizable. That is,*

$$A = V\Lambda V^\top = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top ,$$

where V is an orthogonal matrix with orthonormal columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, $|\mathbf{v}_i| = 1$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

See Appendix F for the proof that does not rely on SVD.

9.6 How to Compute Eigenvalues and Eigenvectors

We can find an eigenpair $(\lambda, \mathbf{v})$ of an $n \times n$ square matrix A by solving

$$A\mathbf{v} = \lambda\mathbf{v}$$

for $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{C}^n$. It does not matter whether A consists solely of real values, since its eigenvalues and eigenvectors may very well contain complex numbers. If we are forced to find eigenvalues and eigenvectors without using a computer, we can try the following procedure:

- Carefully inspect the matrix A to find any clue that allows us to readily compute eigenvalues and their corresponding eigenvectors.
- If there is no such clue, we find eigenvalues by solving the characteristic equation $\det(A-\lambda I) = 0$. Once eigenvalues λ are found, we can use Gaussian elimination to find $(n - \text{rank}(A - \lambda I))$ linearly independent vectors that span Null $(A - \lambda I)$.

If A were symmetric, we can narrow down the search spaces for eigenvalues and eigenvectors to be real only. More specifically, we can use real spectral decomposition (Theorem 5.2) to express the matrix A as the sum of rank-one matrices, as in (5.12), and exploit this structure to identify eigenvalues and eigenvectors. Regardless, there is no one standard approach to solving this eigenvalue problem.

If we are allowed to use computers, we can use any of many numerical methods, such as the power method and QR method based on QR decomposition. We refer readers to any text book on numerical analysis.

9.7 Application: Power Iteration

Let $\mathbf{x}_t = (x_t^1, x_t^2, \ldots, x_t^n)^\top \in \mathbb{R}^n$ describe the state of a system at time t. This system's state at time t can be expressed as either

$$\mathbf{x}_t = A^t\mathbf{x}_0\,,\ t = 1, 2, \ldots$$

in the case of discrete-time system, or

$$\mathbf{x}_t = e^{tA}\mathbf{x}_0 = \Big(I + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \cdots\Big)\mathbf{x}_0\,,\ t \geq 0$$

in the case of continuous-time system,[4] when the system's dynamics is given correspondingly as

$$\mathbf{x}_{t+1} = A\mathbf{x}_t\,,\ t = 0, 1, 2, \ldots,$$

or

$$\frac{d}{dt}\mathbf{x}_t = A\mathbf{x}_t\,,\ t \geq 0.$$

Either way, we need to compute A^k for $k = 1, 2, \ldots$. If A were diagonalizable, i.e. $A = V\Lambda V^{-1}$, it is conceptually easy to compute this, since

$$A^k = V\Lambda^k V^{-1},$$

and matrix exponential is similarly simplified to

$$\begin{aligned} e^{tA} &= V\Big(I + \frac{t}{1!}\Lambda + \frac{t^2}{2!}\Lambda^2 + \frac{t^3}{3!}\Lambda^3 + \cdots\Big)V^{-1} \\ &= V\,\mathrm{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t})V^{-1}. \end{aligned}$$

Even when A is diagonalizable, it may not be feasible to compute eigenpairs if the matrix is large. Here, we consider approximately computing the eigenpair with the largest absolute eigenvalue.

Let $A = V\Lambda V^{-1}$, $V = [\,\mathbf{v}_1|\mathbf{v}_2|\cdots|\,\mathbf{v}_n\,]$ be a diagonalizable matrix, and further assume that $|\mathbf{v}_i| = 1$ for all i and $\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$. First, sample a vector $\mathbf{x}$ from a Gaussian distribution. With probability 1, $w_1 \neq 0$ where $\mathbf{w} = V^{-1}\mathbf{x}$, because $\mathbf{x}$ follows a Gaussian distribution.[5] Let $w_1 > 0$ (if necessary, we can multiply $\mathbf{w}$ with -1.) As we have seen already, $A^kV = V\Lambda^k$, which allows us to get

$$A^k\mathbf{x} = A^kV\mathbf{w} = V\Lambda^k\mathbf{w} = \sum_{i=1}^{n} w_i\lambda_i^k\mathbf{v}_i.$$

If we continue to rearrange terms to be explicit about λ_1 and $\mathbf{v}_1$,

$$\begin{aligned} A^k\mathbf{x} &= w_1\lambda_1^k\mathbf{v}_1 + \sum_{i=2}^{n} w_i\lambda_i^k\mathbf{v}_i \\ &= w_1\lambda_1^k\Big(\mathbf{v}_1 + \sum_{i=2}^{n}\frac{w_i}{w_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k\mathbf{v}_i\Big). \end{aligned}$$

[4] We refer to e^B as the matrix exponential of a square matrix B and define it as

$$e^B = I + \frac{1}{1!}B + \frac{1}{2!}B^2 + \frac{1}{3!}B^3 + \cdots = \sum_{k=0}^{\infty}\frac{1}{k!}B^k.$$

[5] In fact, it is fine to use any continuous distribution to sample $\mathbf{x}$.

If we use $\mathbf{z}_k$ to denote the summation term inside the parentheses on the right hand side,

$$\begin{aligned} |\mathbf{z}_k| &\leq \sum_{i=2}^{n} \left|\frac{w_i}{w_1}\right| \left|\frac{\lambda_i}{\lambda_1}\right|^k |\mathbf{v}_i| \\ &= \sum_{i=2}^{n} \left|\frac{w_i}{w_1}\right| \left|\frac{\lambda_i}{\lambda_1}\right|^k \\ &\leq \left|\frac{\lambda_2}{\lambda_1}\right|^k \frac{\sum_{i=2}^{n} |w_i|}{|w_1|}. \end{aligned} \tag{9.7}$$

As $\frac{1}{|w_1|}\sum_{i=2}^{n}|w_i| > 0$ and $\rho = |\frac{\lambda_2}{\lambda_1}| < 1$, it is guaranteed that $\lim_{k\to\infty}\mathbf{z}_k = \mathbf{0}$, although the convergence can slow down with a large ρ.[6]

Now we design a practical algorithm to estimate the eigenvector of the largest eigenvalue. Because

$$\frac{A^k\mathbf{x}}{|A^k\mathbf{x}|} = \frac{w_1\lambda_1^k(\mathbf{v}_1+\mathbf{z}_k)}{w_1|\lambda_1|^k|\mathbf{v}_1+\mathbf{z}_k|} = \frac{\mathbf{v}_1+\mathbf{z}_k}{|\mathbf{v}_1+\mathbf{z}_k|},$$

in the limit of $k\to\infty$, this quantity converges to the correct vector, as follows:

$$\lim_{k\to\infty}\frac{A^k\mathbf{x}}{|A^k\mathbf{x}|} = \lim_{k\to\infty}\frac{\mathbf{v}_1+\mathbf{z}_k}{|\mathbf{v}_1+\mathbf{z}_k|} = \mathbf{v}_1$$

The algorithm then can be described by the following three steps:

1. $\mathbf{x}_0 = \mathbf{x} \sim N(\mathbf{0}, I)$;
2. $\mathbf{y}_{k+1} = A\mathbf{x}_k$, $\mathbf{x}_{k+1} = \frac{1}{|\mathbf{y}_{k+1}|}\mathbf{y}_{k+1}$;
3. $\alpha_{k+1} = \mathbf{x}_{k+1}^\top A\mathbf{x}_{k+1}$.

By repeating the second step, we get $\mathbf{x}_k \to \mathbf{v}_1$. In the case of the third step, $\lim_{k\to\infty}\alpha_k = \lambda_1$, because

$$\begin{aligned} \alpha_k &= \frac{(\mathbf{v}_1+\mathbf{z}_k)^\top A(\mathbf{v}_1+\mathbf{z}_k)}{|\mathbf{v}_1+\mathbf{z}_k|^2} = \frac{(\mathbf{v}_1^\top+\mathbf{z}_k^\top)(\lambda_1\mathbf{v}_1+A\mathbf{z}_k)}{|\mathbf{v}_1+\mathbf{z}_k|^2} \\ &= \frac{\lambda_1+\mathbf{v}_1^\top A\mathbf{z}_k+\lambda_1\mathbf{z}_k^\top\mathbf{v}_1+\mathbf{z}_k^\top A\mathbf{z}_k}{|\mathbf{v}_1+\mathbf{z}_k|^2}. \end{aligned}$$

By repeating these two steps, the algorithm converges to the eigenpair with the largest eigenvalue. This algorithm is computationally efficient, as each iteration requires matrix-vector multiplication only rather than matrix-matrix multiplication.

[6] Later in Example 10.5, we explain how we can add a rank-one matrix to make $\rho \leq 0.85$ to avoid slow convergence.

Chapter 10

Advanced Results in Linear Algebra

We discuss some remaining results on linear algebra and matrix theory. An exploration of a dual space of a vector space provides many insights into the original vector space. The dual space of a finite-dimensional vector space is characterized very well, but the characterization of the dual space for an infinite-dimensional vector space is tricky and covered in a functional analysis course. The transpose is a popular operation on matrices, but it is difficult to relate it to some aspect of the linear transformation corresponding to the matrix. It can be done implicitly by the adjoint of the linear transformation in a vector space equipped with an inner product. This implicit description of adjoints explains why we could not find an analogy for a transpose in a direct set-up. Through this adjoint, we can explain why a projection matrix should be symmetric. As an important result for the probability theory, Perron-Frobenius theorem says that a matrix with only positive entries must have a positive eigenvalue and an eigenvector with only positive elements. We also discuss the Schur triangularization saying we can find a unitary basis through which an arbitrary matrix is similar to an upper triangular matrix whose diagonals are eigenvalues.

10.1 Dual Space

We call a linear map from a vector space $\mathbb{V}$ to a real $\mathbb{R}$ or complex $\mathbb{C}$ space a linear functional. That is, a linear functional is a scalar-valued linear function.

We then refer to a set of all linear functionals on $\mathbb{V}$, that is,

$$\mathbb{V}^* = \{f : f \text{ is a linear functional on } \mathbb{V}\}$$

as a dual space. This dual space $\mathbb{V}^*$ is also a vector space, as the sum or the scalar multiple of a linear functional on $\mathbb{V}$ is also a linear functional. Then, we want to know the dimension of this vector space $\mathbb{V}^*$ when $\dim \mathbb{V} = n$.

We can think of $\mathbb{R}$ as a one-dimensional vector space and use 1 as its basic vector. Then, a linear functional $f : \mathbb{V} \to \mathbb{R}$ is a linear map from an n-dimensional vector space to a one-dimensional vector space. When $\mathcal{B}_\mathbb{V}$ is a basis of $\mathbb{V}$, there exists a $1 \times n$ matrix A such that $f(\mathbf{v}) = A\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$ is a coordinate of $\mathbf{v} \in \mathbb{V}$ with respect to $\mathcal{B}_\mathbb{V}$. When $\mathbf{a} \in \mathbb{R}^n$ is the only row vector of A, i.e. $A = [\mathbf{a}^\top]$, $f(\mathbf{v}) = \mathbf{a}^\top\mathbf{x}$. In other words, we can view a linear functional in an n-dimensional vector space as an n-dimensional Euclidean vector.

Already in Section 3.8.1, we showed that $a_i = f(\mathbf{v}_i)$ where $\mathcal{B}_\mathbb{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Using this fact, we now show that the correspondence between $\mathbb{V}^*$ and $\mathbb{R}^n$ is injective. If two linear functionals, f and g, are different, there has to be at least one basic vector $\mathbf{v}_j$ for which $f(\mathbf{v}_j) \neq g(\mathbf{v}_j)$. Then, $a_j \neq b_j$ if we use $\mathbf{a}$ and $\mathbf{b}$ to denote the vectors corresponding to functionals f and g, respectively. That is, two vectors in $\mathbb{R}^n$ corresponding to two different linear functionals in $\mathbb{V}^*$ differ. Therefore, this correspondence is injective. Furthermore, if we define $h(\mathbf{v}) = \mathbf{a}^\top\mathbf{x}$ given an arbitrary $\mathbf{a} \in \mathbb{R}^n$, i.e.,

$$h\Big(\sum_{i=1}^n x_i\mathbf{v}_i\Big) = \sum_{i=1}^n a_i x_i,$$

it is clear that h is a linear map from $\mathbb{V}^*$ to $\mathbb{R}$, which implies that this relationship is also surjective. The correspondence between a linear functional in $\mathbb{V}^*$ and an $\mathbb{R}^n$-vector is bijective.

Let us use $f_j \in \mathbb{V}^*$ to denote a linear functional corresponding to $\mathbf{a} = \mathbf{e}_j \in \mathbb{R}^n$. That is, $f_j\big(\sum_{i=1}^n x_i\mathbf{v}_i\big) = \mathbf{e}_j^\top\mathbf{x} = x_j$ for $\mathbf{v} = \sum_{i=1}^n x_i\mathbf{v}_i$. To check the linear independence of f_j's, assume that $\sum_{j=1}^n \alpha_j f_j = 0$, a functional equality, holds for some α_j's. Then, it holds that $\big(\sum_{j=1}^n \alpha_j f_j\big)(\mathbf{v}) = \sum_{j=1}^n \alpha_j f_j(\mathbf{v}) = 0$ for any $\mathbf{v} \in \mathbb{V}$. This is equivalent to $\sum_{j=1}^n \alpha_j f_j\big(\sum_{i=1}^n x_i\mathbf{v}_i\big) = \sum_{j=1}^n \alpha_j x_j = 0$ for any $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. Therefore, it must be that $\alpha_1 = \cdots = \alpha_n = 0$. In other words, $\{f_1, \ldots, f_n\} \subset \mathbb{V}^*$ is linearly independent. $\mathbb{V}^* = \text{span}\{f_1, \ldots, f_n\}$ is easy to see from the surjectivity of the correspondence. So $\{f_1, \ldots, f_n\}$ is a basis of the dual space $\mathbb{V}^*$, to which we refer as a dual basis, and therefore $\dim \mathbb{V}^* = n$. This allows us to treat $\mathbb{V}^*$ and $\mathbb{R}^n$ as if they were the same. We can furthermore

consider the dual space of $\mathbb{V}^*$, that is, $(\mathbb{V}^*)^* = \mathbb{V}^{**}$, since $\mathbb{V}^*$ is itself a vector space.

A linearly constrained optimization problem corresponds to finding a vector that maximizes or minimizes an objective function while satisfying a set of equalities and inequalities defined using linear functionals. We introduce a new dual variable for each linear functional in the original (primal) problem and redefine the optimization problem in terms of dual variables, which we call a dual problem. This serves an important role in optimization.

10.2 Transpose of Matrices and Adjoint of Linear Transformations

Take two vector spaces, $\mathbb{V}$ and $\mathbb{W}$, with inner products defined on them. We use $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ and $\langle \cdot, \cdot \rangle_{\mathbb{W}}$ to refer to them, respectively. We then define the adjoint of a linear transformation, which can be thought of as a functional version of matrix transpose.

Definition 10.1 *A function $f : \mathbb{W} \to \mathbb{V}$ is an adjoint of a linear transformation $T : \mathbb{V} \to \mathbb{W}$ if*

$$\big\langle T(\mathbf{v}), \mathbf{w} \big\rangle_{\mathbb{W}} = \big\langle \mathbf{v}, f(\mathbf{w}) \big\rangle_{\mathbb{V}} \tag{10.1}$$

for all $\mathbf{v} \in \mathbb{V}$ and $\mathbf{w} \in \mathbb{W}$.

We can show the uniqueness of adjoint using the following result derived from the inner product's property.

Fact 10.1 *Let f and g be functions from $\mathbb{W}$ to $\mathbb{V}$. If $\langle \mathbf{v}, f(\mathbf{w}) \rangle_{\mathbb{V}} = \langle \mathbf{v}, g(\mathbf{w}) \rangle_{\mathbb{V}}$ for all $\mathbf{v} \in \mathbb{V}$ and $\mathbf{w} \in \mathbb{W}$, then $f = g$.*

Proof: For $\mathbf{v}^\star = f(\mathbf{w}) - g(\mathbf{w}) \in \mathbb{V}$,

$$0 = \langle \mathbf{v}^\star, f(\mathbf{w}) \rangle_{\mathbb{V}} - \langle \mathbf{v}^\star, g(\mathbf{w}) \rangle_{\mathbb{V}} = \langle \mathbf{v}^\star, f(\mathbf{w}) - g(\mathbf{w}) \rangle_{\mathbb{V}} = \langle f(\mathbf{w}) - g(\mathbf{w}), f(\mathbf{w}) - g(\mathbf{w}) \rangle_{\mathbb{V}}$$

implies $f(\mathbf{w}) - g(\mathbf{w}) = 0$. ■

Using this result, we can show the uniqueness of adjoint, as shown below in Fact 10.2. It is however important to remember that the adjoint of a linear transformation may not exist in an infinite-dimensional vector space, as shown in Example 10.2.

Fact 10.2 *If a linear transformation has an adjoint, then it is unique.*

Proof: If f and g are adjoints of a linear transformation T, $\langle \mathbf{v}, f(\mathbf{w})\rangle_{\mathbb{V}} = \langle \mathbf{v}, g(\mathbf{w})\rangle_{\mathbb{V}}$ for all $\mathbf{v} \in \mathbb{V}$ and $\mathbf{w} \in \mathbb{W}$ according to (10.1). Then, $f = g$ by the result in Fact 10.1.

■

When the adjoint of a linear transformation T exists, we use T^* to refer to it. When $T^* = T$, we say T is self-adjoint. T^* is also a linear transformation, as shown below.

Fact 10.3 *If a linear transformation has an adjoint, then it is also linear.*

Proof: Assume T^* exists. Then, for any $\mathbf{v} \in \mathbb{V}$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{W}$,

$$\begin{aligned}
\langle \mathbf{v}, T^*(\alpha\mathbf{w}_1 + \mathbf{w}_2)\rangle_{\mathbb{V}} &= \langle T(\mathbf{v}), \alpha\mathbf{w}_1 + \mathbf{w}_2\rangle_{\mathbb{W}} \\
&= \langle T(\mathbf{v}), \alpha\mathbf{w}_1\rangle_{\mathbb{W}} + \langle T(\mathbf{v}), \mathbf{w}_2\rangle_{\mathbb{W}} \\
&= \alpha\langle T(\mathbf{v}), \mathbf{w}_1\rangle_{\mathbb{W}} + \langle T(\mathbf{v}), \mathbf{w}_2\rangle_{\mathbb{W}} \\
&= \alpha\langle \mathbf{v}, T^*(\mathbf{w}_1)\rangle_{\mathbb{V}} + \langle \mathbf{v}, T^*(\mathbf{w}_2)\rangle_{\mathbb{V}} \\
&= \langle \mathbf{v}, \alpha T^*(\mathbf{w}_1)\rangle_{\mathbb{V}} + \langle \mathbf{v}, T^*(\mathbf{w}_2)\rangle_{\mathbb{V}} \\
&= \langle \mathbf{v}, \alpha T^*(\mathbf{w}_1) + T^*(\mathbf{w}_2)\rangle_{\mathbb{V}}.
\end{aligned}$$

Since this holds for any $\mathbf{v} \in \mathbb{V}$, $T^*(\alpha\mathbf{w}_1 + \mathbf{w}_2) = \alpha T^*(\mathbf{w}_1) + T^*(\mathbf{w}_2)$ due to the result from Fact 10.1. Therefore, T^* is a linear transformation.

■

Example 10.1 Consider a vector space consisting of polynomials, that is, $\mathbb{P} = \{a_0 + a_1 x + \cdots + a_n x^n : a_i \in \mathbb{R} \text{ for } 0 \le i \le n,\ n = 0, 1, 2, \ldots\}$. We define an inner product as

$$\langle f, g\rangle_{\mathbb{P}} = \int_0^1 f(x)g(x)dx.$$

Let us find the adjoint of a linear transformation $T_p(f) = pf$ given a fixed $p \in \mathbb{P}$. Since

$$\begin{aligned}
\langle T_p(f), g\rangle_{\mathbb{P}} &= \langle pf, g\rangle_{\mathbb{P}} \\
&= \int_0^1 \big(p(x)f(x)\big)g(x)dx \\
&= \int_0^1 f(x)\big(p(x)g(x)\big)dx \\
&= \langle f, pg\rangle_{\mathbb{P}}
\end{aligned}$$

$$= \quad \langle f, T_p(g) \rangle_{\mathbb{P}},$$

$T_p^* = T_p$, which implies that T_p is self-adjoint.

∎

Example 10.2 [Non-existence of Adjoint] Consider the vector space $\mathbb{P}$ from Example 10.1 and a linear transformation $T(f) = f'$, where f' is the derivative of f. Assume the adjoint T^* exists. Let $h = T^*(g) \in \mathbb{P}$, where $g(x) \equiv 1$ is a constant function in $\mathbb{P}$. From the basic result in calculus, we get

$$\langle T(f), g \rangle_{\mathbb{P}} = \int_0^1 f'(x)dx = f(1) - f(0),$$

and for any $f \in \mathbb{P}$, it must hold that

$$\int_0^1 f'(x)dx = f(1) - f(0) = \int_0^1 f(x)h(x)dx.$$

If $f(x) = x^2(x-1)^2h(x)$ such that $f(1) = f(0) = 0$, then

$$0 = \int_0^1 f(x)h(x)dx = \int_0^1 x^2(x-1)^2h(x)^2dx.$$

In this case, for $q(x) = x(x-1)h(x) \in \mathbb{P}$, $q \equiv 0$ since $\langle q, q \rangle_{\mathbb{P}} = 0$, and hence $T^*(g) = h \equiv 0$.[1] Hence, for any $f \in \mathbb{P}$,

$$f(1) - f(0) = \langle T(f), g \rangle_{\mathbb{P}} = \langle f, T^*(g) \rangle_{\mathbb{P}} = \langle f, 0 \rangle_{\mathbb{P}} = 0\,,$$

which is equivalently to $f(1) = f(0)$. We can however find a contradictory example of f, such as $f(x) = x \in \mathbb{P}$. Therefore, the linear transformation that corresponds to differentiation does not have an adjoint.

∎

We now consider the relationship between the adjoint and matrix transpose.

Fact 10.4 *Consider vector spaces $\mathbb{V} = \mathbb{R}^n$ and $\mathbb{W} = \mathbb{R}^m$ with the standard inner products. Let A be an $m \times n$ matrix. Then, the adjoint of $T(\mathbf{v}) = A\mathbf{v}$ is $T^*(\mathbf{w}) = A^\top \mathbf{w}$.*

[1] If $h(x)$ takes a non-zero value for any interval longer than 0 within $[0, 1]$, the integral above cannot be 0, and since $h(x)$ is a polynomial, it must be that $h(x) \equiv 0$.

Proof: Because

$$\langle T(\mathbf{v}), \mathbf{w} \rangle_{\mathbb{W}} = \big(A\mathbf{v}\big)^{\top} \mathbf{w} = \mathbf{v}^{\top} A^{\top} \mathbf{w} = \langle \mathbf{v}, A^{\top} \mathbf{w} \rangle_{\mathbb{V}},$$

$T^*(\mathbf{w})$ is an adjoint of $T(\mathbf{v})$. By Fact 10.2, this is the adjoint of T. ■

Because of this fact, we often use transpose and adjoint interchangeably. Furthermore, in a finite-dimensional space, the matrix representing a self-adjoint operator is symmetric, and thereby we use self-adjoint and symmetric interchangeably as well.

Finally, in a finite-dimensional vector space, any linear transformation can be represented as a finite-size matrix. The adjoint of such a linear transformation is then represented by the transpose of this matrix, according to Fact 10.4. That is, there exists always the adjoint of a linear transformation between finite-dimensional spaces.

10.2.1 Adjoint and Projection

We learned in Fact 4.16 that an orthogonal projection matrix is symmetric and idempotent. It is natural to see that the projection matrix is idempotent, since a vector should not change once it has been projected onto the target subspace. Then, how does the symmetry relate to the fact that it is a projection?

Although we imply orthogonality when we say projection, a general projection does not have to be orthogonal. We can very well project a vector in a 2-dimensional space to a one-dimensional subspace along the 45°-tilted line. We can then ask what is the right way to express orthogonal projection more explicitly. Let $\mathbf{P}(\cdot)$ be a linear transformation that corresponds to the orthogonal projection onto the subspace $\mathbb{W}$ from the original vector space $\mathbb{V}$ in which an inner product $\langle \cdot, \cdot \rangle$ is defined. For any $\mathbf{v}, \mathbf{w} \in \mathbb{V}$, we get

$$0 = \big\langle \mathbf{v} - \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w}) \big\rangle = \big\langle \mathbf{v}, \mathbf{P}(\mathbf{w}) \big\rangle - \big\langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w}) \big\rangle,$$

because $\mathbf{v} - \mathbf{P}(\mathbf{v}) \perp \mathbb{W}$ and $\mathbf{P}(\mathbf{w}) \in \mathbb{W}$. If we swap the roles of $\mathbf{v}$ and $\mathbf{w}$, we also get $\big\langle \mathbf{P}(\mathbf{v}), \mathbf{w} \big\rangle = \big\langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w}) \big\rangle$. Combining these two together, we find

$$\big\langle \mathbf{P}(\mathbf{v}), \mathbf{w} \big\rangle = \big\langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w}) \big\rangle = \big\langle \mathbf{v}, \mathbf{P}(\mathbf{w}) \big\rangle,$$

which implies that orthogonal projection as a linear transformation is self-adjoint. The orthogonal projection matrix is symmetric since the matrix associated with a self-adjoint transformation in finite-dimensional space is symmetric due to Fact 10.4.

Conversely, let an idempotent $\mathbf{P}$ is self-adjoint. In other words, for any $\mathbf{v}, \mathbf{w} \in \mathbb{V}$, it holds that $\langle \mathbf{P}(\mathbf{v}), \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{P}(\mathbf{w}) \rangle$. If we replace $\mathbf{v}$ with $\mathbf{P}(\mathbf{v})$ and $\mathbf{w}$ with $\mathbf{v}$, we end up with $\langle \mathbf{P}(\mathbf{P}(\mathbf{v})), \mathbf{v} \rangle = \langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{v}) \rangle$. Since $\mathbf{P}$ is idempotent, this simplifies to $\langle \mathbf{P}(\mathbf{v}), \mathbf{v} \rangle = \langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{v}) \rangle$. By rearranging the terms, we arrive at

$$\langle \mathbf{P}(\mathbf{v}), \mathbf{v} - \mathbf{P}(\mathbf{v}) \rangle = 0,$$

and $\mathbf{P}$ is an orthogonal projection.

We summarize these observations into the following lemma.

Lemma 10.1 *Let an idempotent linear transformation represent a projection. Then, the projection is orthogonal if and only if the transformation is self-adjoint.*

10.3 Further Results on Positive Definite Matrices

Despite some of the restrictions, such as the lack of commutativity in multiplication, square matrices are similar to real numbers, as addition and multiplication are both well defined for them. Among square matrices, positive definite matrices correspond to positive real numbers and share with them various properties. When we model a complex system with many variables, some of the variables are often expressed as square matrices. If they were positive definite matrices, we can use numerous results on them to conveniently analyze such a system.

Before we continue, we list up symbols that are handy when dealing with positive definite matrices:

- $\mathbb{R}^{m,n}$: a set of $m \times n$ matrices with real entries;
- $\mathbb{S}^n$: a set of $n \times n$ symmetric matrices;
- $\mathbb{S}^n_+$: a set of $n \times n$ positive semi-definite matrices. $A \succeq \mathbf{0}$ if $A \in \mathbb{S}^n_+$;
- $\mathbb{S}^n_{++}$: a set of $n \times n$ positive definite matrices. $A \succ \mathbf{0}$ if $A \in \mathbb{S}^n_{++}$;
- $A \succ B$ (resp., $A \succeq B$) if and only if $A - B \succ \mathbf{0}$ (resp., $A - B \succeq \mathbf{0}$).

10.3.1 Congruence Transformations

The notion of similarity in Definition 9.1 was about the same linear transformation expressed in two different bases; if transformations are the same, the corresponding matrices are similar. On the other hand, here we define the notion of congruence as the preservation of positive (semi-)definiteness of a coefficient matrix of a quadratic form after linearly transforming the variables.

Theorem 10.1 *Let $A \in \mathbb{S}^n$ and $B \in \mathbb{R}^{n,m}$, and consider the product*

$$C = B^\top A B \in \mathbb{S}^m \,. \tag{10.2}$$

1. *If $A \succeq \mathbf{0}$, then $C \succeq \mathbf{0}$;*
2. *If $A \succ \mathbf{0}$, then $C \succ \mathbf{0}$ if and only if $\operatorname{rank} B = m$;*
3. *If B is square and invertible, then $A \succ \mathbf{0}$ (resp. $A \succeq \mathbf{0}$) if and only if $C \succ \mathbf{0}$ (resp. $C \succeq \mathbf{0}$).*

Proof: For $\mathbf{x} \in \mathbb{R}^m$, set $\mathbf{y} = B\mathbf{x} \in \mathbb{R}^n$.

1. $\mathbf{x}^\top C \mathbf{x} = \mathbf{y}^\top A \mathbf{y} \geq 0$ since $A \succeq \mathbf{0}$;
2. Assume $A \succ \mathbf{0}$. If $\operatorname{rank} B = m$, then $\mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0}$ since $\dim \operatorname{Null} B = m - \operatorname{rank} B = 0$. Therefore, $\mathbf{x}^\top C \mathbf{x} = \mathbf{y}^\top A \mathbf{y} > 0$ if $\mathbf{x} \neq \mathbf{0}$. That is, $C \succ \mathbf{0}$. Conversely, if $C \succ \mathbf{0}$, then $\mathbf{x} = \mathbf{0} \Leftrightarrow 0 = \mathbf{x}^\top C \mathbf{x} = \mathbf{y}^\top A \mathbf{y} \Leftrightarrow \mathbf{y} = \mathbf{0}$, that is, $\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{y} = \mathbf{0}$. This implies B has a full-column rank, i.e., $\operatorname{rank} B = m$;
3. If B is square and invertible, then $\operatorname{rank} B = m$ and 2 becomes "if $A \succ \mathbf{0}$, then $C \succ \mathbf{0}$." By applying 1 and 2 to both (10.2) and $A = B^{-\top} C B^{-1}$, we obtain the results.

■

We call (10.2) a congruence transformation. Compare this congruence transformation with the similarity transformation in Definition 9.1.

Corollary 10.1 *For any matrix $B \in \mathbb{R}^{m,n}$, it holds that:*

1. *$B^\top B \succeq \mathbf{0}$ and $B B^\top \succeq \mathbf{0}$;*
2. *$B^\top B \succ \mathbf{0}$ if and only if B has full-column rank, i.e., $\operatorname{rank} B = n$;*
3. *$B B^\top \succ \mathbf{0}$ if and only if B has full-row rank, i.e., $\operatorname{rank} B = m$.*

Proof: By setting $A = I_n$ in Theorem 10.1, we obtain all results. ■

The complexity of a mathematical model involving a matrix can often be lowered if the matrix is diagonalizable. Below, we present important results on two simultaneously diagonalizable matrices and on the diagonalizable product of two matrices.

Theorem 10.2 (Joint diagonalization by similarity transformation) *Let $A_1, A_2 \in \mathbb{S}^n$ and*

$$A = \alpha_1 A_1 + \alpha_2 A_2 \succ \mathbf{0}$$

for some scalars α_1 and α_2. Then, there exists an invertible matrix $B \in \mathbb{R}^{n,n}$ such that $B^\top A_1 B$ and $B^\top A_2 B$ are diagonal.

Proof: If both α_1 and α_2 vanish, $A \not\succ \mathbf{0}$. Hence, at least one of them should be non-zero, and we assume that $\alpha_2 \neq 0$. Since $A \succ \mathbf{0}$, $A = C^\top C$ for some invertible matrix C. If we plug it into the original equation, we get $C^\top C = \alpha_1 A_1 + \alpha_2 A_2$. We re-write it as

$$I_n = \alpha_1 C^{-\top} A_1 C^{-1} + \alpha_2 C^{-\top} A_2 C^{-1}\,.$$

Since $C^{-\top} A_i C^{-1}$ is still symmetric, symmetric spectral decomposition guarantees $C^{-\top} A_1 C^{-1} = Q \Lambda Q^\top$ where Q is orthogonal and Λ is a diagonal matrix with eigenvalues as its diagonal entries. We multiply $Q^\top$ and Q on both sides of the equation and obtain

$$I_n = Q^\top I_n Q = \alpha_1 Q^\top C^{-\top} A_1 C^{-1} Q + \alpha_2 Q^\top C^{-\top} A_2 C^{-1} Q = \alpha_1 \Lambda + \alpha_2 Q^\top C^{-\top} A_2 C^{-1} Q\,.$$

Since $\alpha_2 \neq 0$, we can modify the above equation as

$$\frac{1}{\alpha_2}\left(I_n - \alpha_1 \Lambda\right) = Q^\top C^{-\top} A_2 C^{-1} Q,$$

where the right side is diagonal because the left side is diagonal. Hence, if we set $B = C^{-1} Q$, then B is invertible and diagonalizes both A_1 and A_2. ■

Corollary 10.2 *Let $A \succ \mathbf{0}$ and $C \in \mathbb{S}^n$. Then, there exists an invertible matrix B such that $B^\top C B$ is diagonal and $B^\top A B = I_n$.*

Proof: If we apply Theorem 10.2 to $A = 1A + 0C \succ \mathbf{0}$, then there exists an invertible $\hat{B}$ such that both $\hat{B}^\top A \hat{B}$ and $\hat{B}^\top C \hat{B}$ are diagonal. By the third

property from Theorem 10.1, $\hat{B}^\top A\hat{B} \succ \mathbf{0}$. Denote $\hat{B}^\top A\hat{B} = \text{diag}(\lambda_i)$. Then, the positive definiteness implies $\lambda_i > 0$ for all i. Let $D = \text{diag}(\sqrt{\lambda_i})$, and set $B = \hat{B}D^{-1}$. Then, $B^\top AB = D^{-1}\hat{B}^\top A\hat{B}D^{-1} = D^{-1}D^2D^{-1} = I_n$, and $B^\top CB = D^{-1}\hat{B}^\top C\hat{B}D^{-1}$ is diagonal, since both $\hat{B}^\top C\hat{B}$ and D^{-1} are diagonal. ■

Corollary 10.3 *Let $A, B \in \mathbb{S}^n$ with $A \succ \mathbf{0}$. Then, the matrix AB is diagonalizable and has real eigenvalues only.*

Proof: Let $A^{1/2}$ be the square root of positive definite A. Then,

$$A^{-1/2}ABA^{1/2} = A^{1/2}BA^{1/2}.$$

The matrix on the right side is symmetric. By the spectral decomposition theorem, it is diagonalizable and has real eigenvalues. Since AB and $A^{1/2}BA^{1/2}$ are similar (as in Definition 9.1), both share the same eigenvalues as well as their diagonalizability. ■

10.3.2 Positive Semi-definite Cone and Partial Order

The convexity is an important structure we use to investigate and describe mathematical objects.[2] It often makes analysis and optimization easier and more intuitive to approach when we understand the convexity of target objects. In this section, we are particularly interested in the convexity of a cone consisting of positive (semi-)definite matrices, where a cone is defined as a set of half-infinite rays of elements, such as the set of all positive real numbers and the first quadrant of a real plane.

Definition 10.2 *Let $\mathbb{V}$ be a vector space. A subset K of $\mathbb{V}$ is a cone if $\lambda\mathbf{v} \in K$ for all $\mathbf{v} \in K$ and all $\lambda \geq 0$. A subset of $\mathbb{V}$ is a convex cone if it is convex and a cone.*

In the context of matrices, positive (semi-)definite matrices form a convex cone. It is intuitive to draw this conclusion by noticing the similarity between the positive definiteness of matrices and positivity of real values.

Fact 10.5 *$\mathbb{S}^n_+$ and $\mathbb{S}^n_{++}$ are convex cones.*

[2] If you are not familiar with convexity, refer Appendix A.

Proof: It is clear that $\mathbb{S}_+^n$ is a cone. For $A, B \in \mathbb{S}_+^n$, $\mathbf{x}^\top\big(\lambda A + (1-\lambda)B\big)\mathbf{x} = \lambda\mathbf{x}^\top A\mathbf{x} + (1-\lambda)\mathbf{x}^\top B\mathbf{x} \geq 0$, since $\mathbf{x}^\top A\mathbf{x} \geq 0$ and $\mathbf{x}^\top B\mathbf{x} \geq 0$. Hence, $\lambda A + (1-\lambda)B \in \mathbb{S}_+^n$. It is parallel to show that $\mathbb{S}_{++}^n$ is a convex cone. ■

A major difference between either $\mathbb{S}_+^n$ or $\mathbb{S}_{++}^n$ and positive real values is whether we can tell one is greater than the other given a pair of elements. In the former case of matrices, we often cannot tell this.

Example 10.3 Can we impose an order on positive semi-definite matrices in $\mathbb{S}_+^n$? Let us try to borrow the positive semi-definiteness introduced earlier, $A \succeq B$ if and only if $A - B \succeq \mathbf{0}$. Indeed, the transitivity holds, as $A \succeq C$ if $A \succeq B$ and $B \succeq C$. Consider however the following matrices: $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. It holds that $A \succeq C$ and $B \succeq C$, but neither $A \succeq B$ nor $B \succeq A$ hold between A and B. In other words, there may not be an order defined using $\succeq$ between two positive semi-definite matrices. Therefore, $\succeq$ defines only a partial order on $\mathbb{S}_+^n$.[3]

▮

When $a \geq b$ for two positive real numbers, a and b, $ba^{-1} \leq 1$. We derive a similar property for two positive semi-definite matrices below.

Theorem 10.3 *Let $A \succ \mathbf{0}$ and $B \succeq \mathbf{0}$, and denote by $\rho(\cdot)$ the spectral radius of a matrix (that is, the maximum absolute value of the eigenvalues of a matrix). Then,*

$$A \succeq B \quad \Leftrightarrow \quad \rho(BA^{-1}) \leq 1, \tag{10.3}$$

$$A \succ B \quad \Leftrightarrow \quad \rho(BA^{-1}) < 1. \tag{10.4}$$

Proof: Assume $A - B \succeq \mathbf{0}$. By Corollary 10.2, there exists an invertible matrix M such that $M^\top AM = I_n$ and $M^\top BM = D = \mathrm{diag}(d_i)$ is a diagonal matrix. Since $B \in \mathbb{S}_+^n$, D is also in $\mathbb{S}_+^n$ and $d_i \geq 0$. From $M^\top AM - M^\top BM = I_n - D$, we get $\mathbf{0} \preceq A - B = M^{-\top}(I_n - D)M^{-1}$, to which we apply Theorem 10.1. Then, we obtain that $I_n - D \succeq \mathbf{0}$. Therefore, $d_i \leq 1$ for all i and $\rho(D) \leq 1$.

[3] $\succeq$ is a partial order on A if $a \succeq a$ for all $a \in A$ (reflexivity), $a \succeq b$ and $b \succeq a$ together imply $a = b$ for $a, b \in A$ (antisymmetry), and $a \succeq b$ and $b \succeq c$ together imply $a \succeq c$ for $a, b, c \in A$ (transitivity).

Since $B = M^{-\top}DM^{-1}$ and $A^{-1} = MM^\top$, $BA^{-1} = M^{-\top}DM^{-1}MM^\top = M^{-\top}DM^\top$, that is, D and BA^{-1} are similar to each other. By Fact 9.10, D and BA^{-1} share the same eigenvalues, and $\rho(BA^{-1}) \leq 1$.

(10.4) can be proved parallel to the proof of (10.3). ■

For a pair of square matrices, A and B, AB and BA share the same set of eigenvalues,[4] and thus, $\rho(AB) = \rho(BA)$.Therefore, assuming $A \succ \mathbf{0}$ and $B \succ \mathbf{0}$,

$$A \succeq B \;\Leftrightarrow\; \rho(BA^{-1}) = \rho(A^{-1}B) \leq 1 \;\Leftrightarrow\; B^{-1} \succeq A^{-1}.$$

due to Theorem 10.3. This is similar to the relationship between a positive real number and its inverse.

There are other inequalities induced from the partial order relations of positive semi-definite matrices. Let us see an exemplary case. Consider two positive semi-definite matrices A and B. By Fact 7.8, the following holds for all i if $A \succeq B$:

$$\lambda_i(A) = \lambda_i(B + (A - B)) \geq \lambda_i(B).$$

From this relationship, we derive the following useful inequalities:

$$\begin{aligned} \det A &= \prod_{i=1}^{n} \lambda_i(A) \geq \prod_{i=1}^{n} \lambda_i(B) = \det B, \\ \operatorname{trace} A &= \sum_{i=1}^{n} \lambda_i(A) \geq \sum_{i=1}^{n} \lambda_i(B) = \operatorname{trace} B\,. \end{aligned}$$

In other words, matrix determinant and trace are monotonic functions on $\mathbb{S}^n_+$ with respect to $\succeq$.

As another example, we obtain the following result on the symmetric sum, which appears often in fields such as in control theory, useful as well.

Theorem 10.4 (Symmetric Sum) *Let $A \succ \mathbf{0}$ and $B \in \mathbb{S}^n$, and consider their symmetric sum*

$$S = AB + BA\,.$$

Then, $S \succeq \mathbf{0}$ (resp., $S \succ \mathbf{0}$) implies that $B \succeq \mathbf{0}$ (resp., $B \succ \mathbf{0}$).

Proof: Since B is symmetric, we can write B as $B = Q\Lambda Q^\top$ with an orthogonal matrix $Q = [\,\mathbf{q}_1|\mathbf{q}_2|\cdots|\,\mathbf{q}_n\,]$ and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_i)$, consisting

[4] If $(\lambda, \mathbf{v})$ is an eigenpair of AB, that is, $AB\mathbf{v} = \lambda\mathbf{v}$, then $BAB\mathbf{v} = \lambda B\mathbf{v}$ holds and $(\lambda, B\mathbf{v})$ is an eigenpair of BA. AB and BA however do not necessarily share their eigenvectors.

of B's eigenvalues, according to the spectral decomposition theorem. By Theorem 10.1, $Q^\top SQ \succeq \mathbf{0}$ holds from $S \succeq \mathbf{0}$. Then the diagonal entries $(Q^\top SQ)_{ii}$ are non-negative. We can expand $Q^\top SQ$ as

$$\begin{aligned} Q^\top SQ &= Q^\top ABQ + Q^\top BAQ \\ &= Q^\top AQQ^\top BQ + Q^\top BQQ^\top AQ \\ &= Q^\top AQ\Lambda + \Lambda Q^\top AQ. \end{aligned}$$

Therefore, $(Q^\top SQ)_{ii} = 2\lambda_i (Q^\top AQ)_{ii} = 2\lambda_i \mathbf{q}_i^\top A\mathbf{q}_i$. Combining with $\mathbf{q}_i^\top A\mathbf{q}_i > 0$ from $A \succ \mathbf{0}$ and $(Q^\top SQ)_{ii} \geq 0$, we conclude that $\lambda_i \geq 0$ holds and, therefore, $B \succeq \mathbf{0}$.

We can prove the case of $S \succ \mathbf{0}$ similarly.

■

Example 10.4 [Matrix square-root preserves the PSD ordering] Assume $A \succ \mathbf{0}$ and $B \succ \mathbf{0}$. Because $A^{1/2} + B^{1/2} \succ \mathbf{0}$ and $A^{1/2} - B^{1/2} \in \mathbb{S}^n$, we can use Theorem 10.4 with

$$2(A - B) = \big(A^{1/2} + B^{1/2}\big)\big(A^{1/2} - B^{1/2}\big) + \big(A^{1/2} - B^{1/2}\big)\big(A^{1/2} + B^{1/2}\big),$$

and conclude that $A^{1/2} - B^{1/2} \succ \mathbf{0}$ once $A \succ B$. That is, $A \succ B \succ \mathbf{0}$ implies $A^{1/2} \succ B^{1/2}$. This is similar to how square root maintains the order of positive real numbers.

The converse however does not hold in general. Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1.2 & 1 \\ 1 & 0.9 \end{bmatrix}$. Then, $A \succ \mathbf{0}$, $B \succ \mathbf{0}$ and $A \succ B$. However, $A^2 \not\succ B^2$.

▮

These results sometimes allow us to derive new expressions or conditions, when we use them with block matrices. Let us consider the positive definiteness of block diagonal matrices. First, we see that the positive definiteness of a block diagonal matrix is determined by the positive definiteness of each individual diagonal block. Mathematically, for $M = \begin{bmatrix} A & \mathbf{0}_{n,m} \\ \mathbf{0}_{m,n} & B \end{bmatrix}$,

$$M \succeq \mathbf{0}\,(\text{resp.}, M \succ \mathbf{0}) \quad \Leftrightarrow \quad A \succeq \mathbf{0}, B \succeq \mathbf{0}\,(\text{resp.}, A \succ \mathbf{0}, B \succ \mathbf{0}).^5$$

We can refine this result in the case of a symmetric block diagonal matrix, as follow.

[5] We leave it for you to prove this.

Fact 10.6 (Schur Complement) *Let $A \in \mathbb{S}^n$, $B \in \mathbb{S}^m$ and $C \in \mathbb{R}^{n,m}$ with $B \succ \mathbf{0}$. Consider a symmetric block matrix*

$$M = \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix},$$

and consider the Schur complement $S = A - CB^{-1}C^\top$ of A with respect to M. Then,

$$M \succeq \mathbf{0} \, (resp., M \succ \mathbf{0}) \quad \Leftrightarrow \quad S \succeq \mathbf{0} \, (resp., S \succ \mathbf{0}).$$

Proof: We get a block diagonal matrix after performing Gaussian elimination on the block matrix M, as follows: For a lower triangular matrix $L = \begin{bmatrix} I_n & \mathbf{0} \\ -B^{-1}C^\top & I_m \end{bmatrix}$,

$$\begin{aligned} L^\top M L &= \begin{bmatrix} I_n & -CB^{-1} \\ \mathbf{0} & I_m \end{bmatrix} \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -B^{-1}C^\top & I_m \end{bmatrix} \\ &= \begin{bmatrix} S & \mathbf{0} \\ C^\top & B \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -B^{-1}C^\top & I_m \end{bmatrix} \\ &= \begin{bmatrix} S & \mathbf{0}_{n,m} \\ \mathbf{0}_{m,n} & B \end{bmatrix} = D. \end{aligned}$$

Because $B \succ \mathbf{0}$, the positive definiteness of D is equivalent to that of S. The positive definiteness of M is equivalent to that of D due to the third property in Theorem 10.1, since L is an invertible lower triangular matrix with all diagonal entries set to 1. Therefore, the positive definiteness of M is equivalent to that of S.

■

10.4 Schur Triangularization

We present and prove Schur triangularization which is useful not only for proving further results later in this book but also in many real-world problems where the repeated product of a matrix is needed. Before doing so, we introduce a unitary matrix which is a complex version of the orthogonal matrix. A matrix Q is unitary when $Q^{-1} = Q^{\mathsf{H}}$, that is, $QQ^{\mathsf{H}} = Q^{\mathsf{H}}Q = I$. If $A = Q\,B\,Q^{\mathsf{H}}$ for a unitary matrix Q, $A^n = QB^nQ^{\mathsf{H}}$ for any matrix B, which is a useful property when computing the power of a matrix.

Theorem 10.5 (Schur Triangularization) *Let the eigenvalues of $n \times n$ matrix A be arranged in any given order $\lambda_1, \lambda_2, \ldots, \lambda_n$ (including multiplicities), and let $(\lambda_1, \mathbf{x})$ be an eigenpair of A, in which $\mathbf{x}$ is a unit vector. Then,*

(a) *There is an $n \times n$ unitary matrix $Q = [\mathbf{x}|Q_2]$ such that $A = QUQ^{\mathsf{H}}$, where $U = (u_{ij})$ is upper triangular and has diagonal entries $u_{ii} = \lambda_i$ for $i = 1, 2, \ldots, n$.*

(b) *If A, each eigenvalue and $\mathbf{x}$ are all real, then there is an $n \times n$ real orthogonal $Q = [\mathbf{x}|Q_2]$ such that $A = QUQ^{\top}$, where $U = (u_{ij})$ is upper triangular and has diagonal entries $u_{ii} = \lambda_i$ for $i = 1, 2, \ldots, n$.*

Proof: We use mathematical induction to prove this theorem. First, (a) trivially holds when $n = 1$. Now assume (a) holds up to $n - 1$. We construct an $n \times n$ unitary matrix $\hat{Q} = [\mathbf{x}|V]$ from n vectors including $\mathbf{x}$, which can be obtained using for instance the Gram-Schmidt procedure. V here denotes an $n \times (n-1)$ matrix such that $V^{\mathsf{H}}\mathbf{x} = \mathbf{0}$. Then,

$$\hat{Q}^{\mathsf{H}} A \hat{Q} = \hat{Q}^{\mathsf{H}} [\lambda_1 \mathbf{x} \,|\, AV] = \begin{bmatrix} \mathbf{x}^{\mathsf{H}} \\ V^{\mathsf{H}} \end{bmatrix} [\lambda_1 \mathbf{x} \,|\, AV] = \begin{bmatrix} \lambda_1 \mathbf{x}^{\mathsf{H}}\mathbf{x} & \mathbf{x}^{\mathsf{H}} AV \\ \lambda_1 V^{\mathsf{H}}\mathbf{x} & V^{\mathsf{H}} AV \end{bmatrix} = \begin{bmatrix} \lambda_1 & \mathbf{x}^{\mathsf{H}} AV \\ \mathbf{0} & V^{\mathsf{H}} AV \end{bmatrix},$$

because

$$A\hat{Q} = [A\mathbf{x}|AV] = [\lambda_1 \mathbf{x}|AV].$$

The eigenvalues of the upper triangular block matrix on the right-hand side consist of the eigenvalues of individual blocks, that is, λ_1 and the eigenvalues of $V^{\mathsf{H}}AV$. Because A and $\hat{Q}^{\mathsf{H}} A \hat{Q}$ are similar and thus have the same set of eigenvalues, the eigenvalues of $V^{\mathsf{H}}AV$ are $\lambda_2, \ldots, \lambda_n$. Due to the assumption, it holds that $V^{\mathsf{H}}AV = Q_{n-1} U_{n-1} Q_{n-1}^{\mathsf{H}}$, where Q_{n-1} is an $(n-1) \times (n-1)$ unitary matrix, and U_{n-1} is an upper triangular matrix whose diagonal entries are $\lambda_2, \ldots, \lambda_n$. Using an appropriate choice of an $(n-1)$-dimensional $\mathbf{a}$, we can rewrite $\hat{Q}^{\mathsf{H}} A \hat{Q}$ as

$$\hat{Q}^{\mathsf{H}} A \hat{Q} = \begin{bmatrix} \lambda_1 & \mathbf{x}^{\mathsf{H}} AV \\ \mathbf{0} & Q_{n-1} U_{n-1} Q_{n-1}^{\mathsf{H}} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_{n-1} \end{bmatrix} \begin{bmatrix} \lambda_1 & \mathbf{a}^{\mathsf{H}} \\ \mathbf{0} & U_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_{n-1} \end{bmatrix}^{\mathsf{H}}.$$

Since Q_{n-1} is unitary, both $\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_{n-1} \end{bmatrix}$ and $Q = \hat{Q}\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_{n-1} \end{bmatrix}$ are unitary as well. $U = \begin{bmatrix} \lambda_1 & \mathbf{a}^{\mathsf{H}} \\ \mathbf{0} & U_{n-1} \end{bmatrix}$ is an upper triangular matrix with $\lambda_1, \lambda_2, \ldots, \lambda_n$ on its diagonal. With Q and U, we see that $A = QUQ^{\mathsf{H}}$, which proves (a). We can prove (b) similarly however with conjugation in (a) replaced with transpose.

■

With Schur triangularization, we can easily prove earlier results on computing the trace and determinant of a matrix using its eigenvalues.

Corollary 10.4 *Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of an $n \times n$ matrix A. Then,*

$$\operatorname{trace} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n \quad \textit{and} \quad \det A = \lambda_1 \lambda_2 \cdots \lambda_n \,.$$

Proof: According to Theorem 10.5, we can decompose A as $A = QUQ^{\mathsf{H}}$. Then,

$$\operatorname{trace} A = \operatorname{trace}(QUQ^{\mathsf{H}}) = \operatorname{trace}(UQ^{\mathsf{H}}Q) = \operatorname{trace} U = \lambda_1 + \lambda_2 + \cdots + \lambda_n,$$

and

$$\det A = \det(QUQ^{\mathsf{H}}) = \det Q \det U \det Q^{\mathsf{H}} = \det(QQ^{\mathsf{H}}) \det U = \det U = \lambda_1 \lambda_2 \cdots \lambda_n.$$

■

10.5 Perron-Frobenius Theorem

Let an $n \times n$ matrix $A = (a_{ij})$ satisfy $a_{ij} \geq 0$ and $\sum_{j=1}^{n} a_{ij} = 1$, that is, the sum of each row is 1, and all the entries are greater than or equal to 0. We call such a matrix a Markov matrix and use it to describe a probabilistic transition in a dynamical system known as a Markov chain. In this case, a_{ij} denotes the probability of the system transitioning from the i-th state to the j-th state. Because every row sums to 1, the Markov matrix always has $(1, \mathbf{1})$ as its eigenpair. All the other eigenvalues are less than or equal to 1.

Fact 10.7 *$\rho(A) = 1$ if A is a Markov matrix.*

Proof: Let $(\lambda, \mathbf{v})$ be an eigenpair of A. $|v_k| > 0$ if $k = \operatorname{argmax}_{1 \leq i \leq n} |v_i|$. Let us consider the k-th equation in $A\mathbf{v} = \lambda\mathbf{v}$, which is $\sum_{j=1}^{n} a_{kj} v_j = \lambda v_k$. Since A

is a Markov matrix, $a_{kj} \geq 0$ and $\sum_{j=1}^{n} a_{kj} = 1$. Then, $|\lambda| \leq 1$, because

$$|\lambda||v_k| = |\lambda v_k| = \left|\sum_{j=1}^{n} a_{kj} v_j\right| \leq \sum_{j=1}^{n} a_{kj}|v_j| \leq \sum_{j=1}^{n} a_{kj}|v_k| = |v_k|.$$

Since $(1, \mathbf{1})$ is an eigenpair, we conclude that $\rho(A) = 1$.

■

We call a vector $\mathbf{p}$ a probability vector if $\mathbf{p} \geq \mathbf{0}$ and $\mathbf{p}^\top \mathbf{1} = 1$. When $\mathbf{p}$ represents the distribution over the states of a system at a time, $\mathbf{p}^\top A$ is the distribution after one step of transition happened. Then, $\mathbf{p}^\top A = \mathbf{p}^\top$ implies that the distribution over the system's states does not evolve even after transition according to A. We call such a probability vector $\mathbf{p}$ a stationary distribution or an equilibrium distribution.

We present the Perron-Frobenius theorem for a matrix with positive entries only.

Theorem 10.6 (Perron-Frobenius) *Let $A = (a_{ij})$ be an $n \times n$ positive matrix, i.e., $a_{ij} > 0$ for all i and j. Then, there exists a positive eigenvalue of the spectral radius of A, and its associated eigenvector with at least one positive element is unique up to scaling by a positive constant. This eigenvector in fact has only positive components.*

Proof: Consider $\lambda_0 = \max\{\lambda : \text{there exists } \mathbf{x} \neq \mathbf{0}, \mathbf{x} \geq \mathbf{0} \text{ such that } A\mathbf{x} \geq \lambda \mathbf{x}\}$. Because $A > \mathbf{0}$, $A\mathbf{1} > \mathbf{0}$. From this, we get that $A\mathbf{1} \geq \lambda' \mathbf{1}$ and hence $\lambda_0 \geq \lambda' > 0$, where λ' is the smallest element of vector $A\mathbf{1}$.[6] For this maximal λ_0, let $\mathbf{x}_0 \geq \mathbf{0}$ satisfy $A\mathbf{x}_0 \geq \lambda_0 \mathbf{x}_0$. If $A\mathbf{x}_0 \neq \lambda_0 \mathbf{x}_0$, $\mathbf{y} = A\mathbf{x}_0 - \lambda_0 \mathbf{x}_0 \geq \mathbf{0} \neq \mathbf{0}$. Since $A > \mathbf{0}$, we know that $A\mathbf{y} > \mathbf{0}$ and that $A(A\mathbf{x}_0) > \lambda_0(A\mathbf{x}_0)$, because $AA\mathbf{x}_0 - \lambda_0 A\mathbf{x}_0 = A\mathbf{y} > \mathbf{0}$. This means that we can find λ greater than λ_0 for the vector $A\mathbf{x}_0$ that satisfies $A(A\mathbf{x}_0) \geq \lambda(A\mathbf{x}_0)$, which is contradictory to the definition of λ_0. Therefore, $A\mathbf{x}_0 = \lambda_0 \mathbf{x}_0$, and $\mathbf{x}_0 > \mathbf{0}$ since $A\mathbf{x}_0 > \mathbf{0}$.

Let $(\lambda, \mathbf{y})$ be an eigenpair with $\lambda \neq \lambda_0$. For $\mathbf{y} = (y_j)$, let $\hat{\mathbf{y}}$ be a vector with $|y_j|$ as its elements. If we take the absolute values of both sides of $\sum_{j=1}^{n} a_{ij} y_j = \lambda y_i$, which is the i-th row of $A\mathbf{y} = \lambda \mathbf{y}$, we get

$$|\lambda|\hat{y}_i = |\lambda||y_i| = |\lambda y_i| = \left|\sum_{j=1}^{n} a_{ij} y_j\right| \leq \sum_{j=1}^{n} a_{ij}|y_j| = \sum_{j=1}^{n} a_{ij}\hat{y}_j.$$

[6] we can also show that $\lambda_0 \leq \max_{i,j} a_{ij}$.

In vectors, this is equivalent to $A\hat{\mathbf{y}} \geq |\lambda|\hat{\mathbf{y}}$. Together with the definition of λ_0, $|\lambda| \leq \lambda_0$, which implies $\rho(A) = \lambda_0$.

Finally, consider $\mathbf{x}_1 > \mathbf{0}$ that is linearly independent of $\mathbf{x}_0$ and satisfies $A\mathbf{x}_1 = \lambda_0\mathbf{x}_1$. Then, $\mathbf{w} = \alpha\mathbf{x}_0 + \mathbf{x}_1$ is an eigenvector of λ_0 for any α, but there exists a negative value for α such that $\mathbf{w} \geq \mathbf{0}$ but $\mathbf{w} \not> \mathbf{0}$. This is contradictory, as we already showed above that the eigenvector of λ_0 is positive. Therefore, There is a unique eigenvector associated with λ_0 up to scaling. ■

When every entry of a Markov matrix A is positive, it has a positive left-eigenvector associated with the eigenvalue of 1 according to Theorem 10.6 applied to $A^\top$. Once we normalize this vector (to sum to 1), we get the stationary distribution. If A has at least one zero, Theorem 10.6 does not apply. In the following section, we however use another result to show that this matrix also has the stationary distribution.

10.6 Eigenvalue Adjustments and Applications

When we add a rank-one matrix $\mathbf{v}\mathbf{w}^\mathsf{H}$ to a matrix, which has an eigenpair $(\lambda, \mathbf{v})$, only λ changes to $\lambda + \mathbf{w}^\mathsf{H}\mathbf{v}$, while all the other eigenvalues are maintained. This is a useful result when we want to adjust only one particular eigenvalue, and this had been used to improve the convergence rate of Google's PageRank. We refer to this result, stated below, as the Brauer theorem.

Theorem 10.7 (Brauer) *Let $(\lambda, \mathbf{v})$ be an eigenpair of an $n \times n$ matrix A and $\lambda, \lambda_2, \ldots, \lambda_n$ its eigenvalues. For any $\mathbf{w} \in \mathbb{C}^n$, the eigenvalues of $A + \mathbf{v}\mathbf{w}^\mathsf{H}$ are $\lambda + \mathbf{w}^\mathsf{H}\mathbf{v}, \lambda_2, \ldots, \lambda_n$, and $(\lambda + \mathbf{w}^\mathsf{H}\mathbf{v}, \mathbf{v})$ is an eigenpair of $A + \mathbf{v}\mathbf{w}^\mathsf{H}$.*

Proof: $\mathbf{u} = \frac{1}{|\mathbf{v}|}\mathbf{v}$ is a unit vector constructed from $\mathbf{v}$. Then, $(\lambda, \mathbf{u})$ is also an eigenpair of A. By Schur triangularization from Theorem 10.5, we know the following holds with a unitary matrix $Q = [\mathbf{u} \,|\, Q_2]$: $Q^\mathsf{H}AQ = U = \begin{bmatrix} \lambda & \mathbf{a}^\mathsf{H} \\ \mathbf{0} & U_{n-1} \end{bmatrix}$, where U_{n-1} is an upper triangular matrix with $\lambda_2, \ldots, \lambda_n$ on its diagonal. Because

$$Q^\mathsf{H}\mathbf{v}\mathbf{w}^\mathsf{H}Q \quad = \quad (Q^\mathsf{H}\mathbf{v})(\mathbf{w}^\mathsf{H}Q) = \begin{bmatrix} \mathbf{u}^\mathsf{H}\mathbf{v} \\ Q_2^\mathsf{H}\mathbf{v} \end{bmatrix} [\mathbf{w}^\mathsf{H}\mathbf{u} \,|\, \mathbf{w}^\mathsf{H}Q_2]$$

$$= \begin{bmatrix} |\mathbf{v}| \\ \mathbf{0} \end{bmatrix} [\mathbf{w}^{\mathsf{H}}\mathbf{u} \,|\, \mathbf{w}^{\mathsf{H}}Q_2] = \begin{bmatrix} \mathbf{w}^{\mathsf{H}}\mathbf{v} & \mathbf{b}^{\mathsf{H}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \text{ for some vector } \mathbf{b},$$

we arrive at

$$Q^{\mathsf{H}}(A + \mathbf{v}\mathbf{w}^{\mathsf{H}})Q = \begin{bmatrix} \lambda & \mathbf{a}^{\mathsf{H}} \\ \mathbf{0} & U_{n-1} \end{bmatrix} + \begin{bmatrix} \mathbf{w}^{\mathsf{H}}\mathbf{v} & \mathbf{b}^{\mathsf{H}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \lambda + \mathbf{w}^{\mathsf{H}}\mathbf{v} & (\mathbf{a}+\mathbf{b})^{\mathsf{H}} \\ \mathbf{0} & U_{n-1} \end{bmatrix}.$$

The eigenvalues of the final matrix are $\lambda + \mathbf{w}^{\mathsf{H}}\mathbf{v}, \lambda_2, \ldots, \lambda_n$, and so are the eigenvalues of its similar matrix $A + \mathbf{v}\mathbf{w}^{\mathsf{H}}$. Furthermore, $(A + \mathbf{v}\mathbf{w}^{\mathsf{H}})\mathbf{v} = A\mathbf{v} + \mathbf{v}\mathbf{w}^{\mathsf{H}}\mathbf{v} = (\lambda + \mathbf{w}^{\mathsf{H}}\mathbf{v})\mathbf{v}$, which completes the proof. ■

Using Theorem 10.7, we can also change the rest of the eigenvalue of a matrix.

Corollary 10.5 *Let $(\lambda, \mathbf{v})$ be an eigenpair of $n \times n$ matrix A and let $\lambda, \lambda_2, \ldots, \lambda_n$ be its eigenvalues. Let $\mathbf{w} \in \mathbb{C}^n$ be such that $\mathbf{w}^{\mathsf{H}}\mathbf{v} = 1$ and let $\tau \in \mathbb{C}$. Then the eigenvalues of $A_\tau = \tau A + (1-\tau)\lambda \mathbf{v}\mathbf{w}^{\mathsf{H}}$ are $\lambda, \tau\lambda_2, \ldots, \tau\lambda_n$.*

Proof: The eigenvalue of τA are $\tau\lambda_1, \tau\lambda_2, \ldots, \tau\lambda_n$. By Theorem 10.7, the eigenvalues of $A_\tau = \tau A + \mathbf{v}\big((1-\overline{\tau})\overline{\lambda}\mathbf{w}\big)^{\mathsf{H}}$ are $\tau\lambda + \big((1-\overline{\tau})\overline{\lambda}\mathbf{w}\big)^{\mathsf{H}}\mathbf{v} = \tau\lambda + (1-\tau)\lambda\mathbf{w}^{\mathsf{H}}\mathbf{v} = \tau\lambda + (1-\tau)\lambda = \lambda$ and $\tau\lambda_2, \ldots, \tau\lambda_n$. ■

Now we present the Perron-Frobenius theorem for a non-negative matrix, expanding Theorem 10.6 from earlier.

Theorem 10.8 (Perron-Frobenius) *Any Markov matrix has a stationary distribution.*

Proof: Let A be a Markov matrix. Fact 10.7 states that the eigenvalues of A are $1, \lambda_2, \ldots$ and λ_n with $|\lambda_i| \leq 1$ for all i. Furthermore, if $A > \mathbf{0}$, Theorem 10.6 shows that there is a unique stationary distribution. If A has zeroes, we cannot use Theorem 10.6. Instead, we convert A into a positive matrix using Theorem 10.7. If we apply Theorem 10.7 with $\mathbf{v} = \mathbf{1}$ and $\mathbf{w} = \epsilon\mathbf{1}$, where ϵ is a small positive number, the eigenvalues of the positive matrix $A + \epsilon\mathbf{1}\mathbf{1}^{\top}$ are $1 + n\epsilon$, $\lambda_2, \ldots, \lambda_n$. According to Theorem 10.6, there is a unique unit left-eigenvector $\mathbf{u}_\epsilon$ associated with the eigenvalue $1 + n\epsilon$ for the positive matrix $A + \epsilon\mathbf{1}\mathbf{1}^{\top}$ for each ϵ, and this eigenvector is positive. Since $\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}| = 1\}$ is compact, a

subsequence $\mathbf{u}_{\epsilon_k}$ of a sequence $\mathbf{u}_\epsilon$ on this compact sphere converges to a vector $\mathbf{u} = \lim_{k\to\infty} \mathbf{u}_{\epsilon_k}$ on the sphere as $k \to \infty$. Because

$$\mathbf{u}_{\epsilon_k}^\top \left(A + \epsilon_k \mathbf{1}\mathbf{1}^\top\right) = (1 + n\epsilon_k)\mathbf{u}_{\epsilon_k}^\top,$$

we get, in the limit of $k \to \infty$,

$$\mathbf{u}^\top A = \mathbf{u}^\top.$$

Furthermore, $\mathbf{u} \geq \mathbf{0}$ since $\mathbf{u}_{\epsilon_k} \geq \mathbf{0}$ for all k, which completes the proof. ■

In order to prove the theorem above, we rely on the topological concept of compactness. We refer readers to any textbook on mathematical analysis for more details, such as [1].

Example 10.5 [Google Matrix] Assume a large Markov matrix A that represents the connectivity among the web pages on the internet. A core idea behind Google Search is to rank these web pages according to the left-eigenvector of this matrix A. When a Markov matrix is large, it is usual in practice to use an iterative algorithm, such as power iteration from Section 9.7.

For faster convergence of the power iteration (9.7), we modify A such that the eigenvalue of the largest absolute value, which is by construction 1, is unique, and the absolute values of all the other eigenvalues are less than 1. We ensure the modified matrix continues to be a Markov matrix.

Let $\mathbf{v} = \mathbf{1}$, $\mathbf{w} = \frac{1}{n}\mathbf{1}$, $E = \mathbf{1}\mathbf{1}^\top$ and $0 < \tau < 1$. Note that $\mathbf{w}^\top\mathbf{v} = 1$. Compute the weighted sum of a rank-one matrix, $\mathbf{v}\mathbf{w}^\top$ to A with τ and $1 - \tau$ as their coefficients, respectively, and we get

$$A_\tau = \tau A + (1 - \tau)\mathbf{v}\mathbf{w}^\top = \tau A + \frac{1 - \tau}{n} E. \tag{10.5}$$

The eigenvalues of this matrix A_τ are $1, \tau\lambda_2, \ldots, \tau\lambda_n$ according to Corollary 10.5. In other words, all eigenvalues except for 1 have been shrunk by the factor of τ. This weighted sum maintains the row sum to be 1, because $A_\tau\mathbf{1} = \tau A\mathbf{1} + (1 - \tau)\mathbf{1}\frac{1}{n}\mathbf{1}^\top\mathbf{1} = \tau\mathbf{1} + (1 - \tau)\mathbf{1} = \mathbf{1}$, and all the entries of this matrix continue to be positive. We refer to A_τ a Google matrix, as it was used to rank web pages for Google Search in early years. The founders of Google used $\tau = 0.85$ originally.

▮

Chapter 11

Big Theorems in Linear Algebra

We can define a monomial of a square matrix A by replacing x in $p(x) = cx^n$ with A, such that $p(A) = cA^n$. More generally, we can think of replacing x with a square matrix A in a polynomial $p(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$. By defining A^0 to be an identity matrix and treating c_0 as $c_0 x^0$, we get $p(A) = c_n A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I$. Let us use $p_A(x)$ to denote the characteristic polynomial of a matrix A, which may have complex coefficients. The two most abstract and general results in linear algebra are the Cayley-Hamilton theorem and the Jordan normal form theorem, both of which work with a characteristic polynomial of a matrix. According to the Cayley-Hamilton theorem, $p_A(A) = \mathbf{0}$ for any A, and using the Jordan normal form theorem, we can show that any matrix is similar to either a diagonal matrix or a Jordan form matrix, where a Jordan form matrix closely resembles a diagonal matrix.

11.1 The First Big Theorem: Cayley-Hamilton Theorem

In this section, we present the Cayley-Hamilton theorem which states that $p_A(A) = \mathbf{0}$ for any square matrix A when $p_A(x)$ is its characteristic polynomial. We use Schur triangularization from Theorem 10.5 to prove it.

Theorem 11.1 (Cayley-Hamilton) *Let*

$$p_A(x) = \det(A - xI) = (-1)^n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$$

be the characteristic polynomial of an $n \times n$ matrix A. Then

$$p_A(A) = (-1)^n A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_n = \mathbf{0}\,.$$

Proof: Let the eigenvalues of A be $\lambda_1, \lambda_2, \ldots, \lambda_n$. We can then write its characteristic polynomial as

$$p_A(x) = (\lambda_1 - x)(\lambda_2 - x)\cdots(\lambda_n - x).$$

According to the Schur triangularization (Theorem 10.5), we can rewrite A as $A = QUQ^{\mathsf{H}}$ with a unitary matrix Q and an upper triangular matrix U with $\lambda_1, \lambda_2, \ldots, \lambda_n$ on its diagonal. Since $p_A(A) = Q\,p_A(U)\,Q^{\mathsf{H}}$ from $Q^{\mathsf{H}}Q = I$, all we need is to show that $p_A(U) = (-1)^n(U - \lambda_1 I)(U - \lambda_2 I)\cdots(U - \lambda_n I) = \mathbf{0}$.

For $j = 1, 2, \ldots, n$, set

$$U_j = (U - \lambda_1 I)(U - \lambda_2 I)\cdots(U - \lambda_j I),$$

so that $p_A(U) = (-1)^n U_n$ holds. $U - \lambda_i I$ is an upper triangular matrix with its null i-th diagonal, that is, $(U - \lambda_i I)_{ii} = 0$. For instance, the first column of $U_1 = U - \lambda_1 I$ is an all-zero vector. Similarly, the first two columns of $U_2 = (U - \lambda_1 I)(U - \lambda_2 I)$ are all zeros, which can be checked without difficulty.

Based on this observation, assume the first $j - 1$ columns of U_{j-1} are all zeros. Given $U_j = U_{j-1}(U - \lambda_j I)$, each element in the j-th column of U_j is determined as the inner product between the corresponding row of U_{j-1} and the j-th column of $U - \lambda_j I$. Since the j-th element in the j-th column of $U - \lambda_j I$ is zero, only the first $j - 1$ entries of the j-th column can be non-zero. By the assumption, the first $j - 1$ entries of every row of U_{j-1} are all zeros, and hence the inner products of the rows of U_{j-1} and the j-th column of $U - \lambda_j I$ result in all zeros. In other words, the j-th column of U_j is an all-zero vector. Therefore, the first j columns of U_j are all zeros, which implies that $p_A(U) = (-1)^n U_n = \mathbf{0}$. ■

Although the Cayley-Hamilton theorem is frequently used in theoretical derivations and thought experiments, it is not practically used due to the high computational complexity involved in identifying the coefficients of the characteristic polynomial. As an example of theoretical use cases, let us consider

one interesting consequence of this theorem on the polynomial expression of an inverse matrix. When the characteristic polynomial of an invertible matrix A is $(-1)^n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$, the Cayley-Hamilton theorem states that $(-1)^n A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I_n = \mathbf{0}$. With $c_0 \neq 0$, after multiplying both sides with A^{-1} and re-arranging terms, we arrive at

$$A^{-1} = -\frac{c_1}{c_0} I - \frac{c_2}{c_0} A - \cdots - \frac{c_{n-1}}{c_0} A^{n-2} - \frac{(-1)^n}{c_0} A^{n-1}.$$

This allows us to express the inverse of A as a linear combination of the powers of A, although it is a hurdle in practice to figure out c_i's.

11.2 Decomposition of Nilpotency into Cyclic Subspaces

Nilpotency of a Matrix

A square matrix A is nilpotent of degree r if $A^{r-1} \neq \mathbf{0}$ but $A^r = \mathbf{0}$ for a positive integer r. More generally, given a subspace $\mathbb{W}$, we say "A is nilpotent of degree r on $\mathbb{W}$", if $A^r \mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in \mathbb{W}$ but $A^{r-1}\mathbf{w} \neq \mathbf{0}$ for some $\mathbf{w} \in \mathbb{W}$. In this section, we construct and investigate an important subspace given a nilpotent matrix.

Let us construct the following subset $\mathbb{V}_A$ of an n-dimensional vector space $\mathbb{V}$ and an $n \times n$ matrix A:

$$\mathbb{V}_A = \{\mathbf{v} \in \mathbb{V} : A^j \mathbf{v} = \mathbf{0} \quad \text{for some} \quad j\}. \tag{11.1}$$

Note that $A^j \mathbf{v} = A^{j-j_1} A^{j_1} = \mathbf{0}$ for all $j \geq j_1$ if $A^{j_1}\mathbf{v} = \mathbf{0}$ for some j_1. This implies that if $A^{j_1}\mathbf{v}_1 = \mathbf{0}$ and $A^{j_2}\mathbf{v}_2 = \mathbf{0}$, then, $A^{\max\{j_1,j_2\}}(\mathbf{v}_1 + \mathbf{v}_2) = A^{\max\{j_1,j_2\}}\mathbf{v}_1 + A^{\max\{j_1,j_2\}}\mathbf{v}_2 = \mathbf{0}$. That is, $\mathbf{v}_1 + \mathbf{v}_2 \in \mathbb{V}_A$ for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}_A$, which says that $\mathbb{V}_A$ is a subspace of $\mathbb{V}$.

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis of $\mathbb{V}_A$. Each $\mathbf{v}_i$ is in $\mathbb{V}_A$ and there exists r_i with which $A^{r_i}\mathbf{v}_i = \mathbf{0}$ for all i. Set $r_0 = \max\{r_1, \ldots, r_k\}$. For any $r \geq r_0$, $A^r \mathbf{v}_i = \mathbf{0}$ for all i. Because we can express any vector $\mathbf{v}$ in $\mathbb{V}_A$ as $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k$, $A^r\mathbf{v} = x_1 A^r \mathbf{v}_1 + \cdots + x_k A^r \mathbf{v}_k = \mathbf{0}$, and hence $\mathbb{V}_A \subset \text{Null}\,(A^r)$. From the definition of $\mathbb{V}_A$ in (11.1), it is clear that $\mathbb{V}_A \supset \text{Null}\,(A^r)$, and therefore, we arrive at

$$\mathbb{V}_A = \text{Null}\,(A^r) \quad \text{for some positive integer} \quad r.$$

Assume $\mathbf{v} \in \text{Null}\,(A^r) \cap \text{Col}\,(A^r)$. Because $\mathbf{v} \in \text{Col}\,(A^r)$, there exists $\mathbf{w} \in \mathbb{V}$ such that $\mathbf{v} = A^r\mathbf{w}$. Also, because $\mathbf{v} \in \text{Null}\,(A^r)$, $A^r\mathbf{v} = A^{2r}\mathbf{w} = \mathbf{0}$. According to (11.1), $\mathbf{w} \in \mathbb{V}_A$ and $\mathbf{v} = A^r\mathbf{w} = \mathbf{0}$, which tells us that $\text{Null}\,(A^r) \cap \text{Col}\,(A^r) = \{\mathbf{0}\}$. In other words, two subspaces, $\text{Null}\,(A^r)$ and $\text{Col}\,(A^r)$, are mutually independent. That is, $x_1 = x_2 = 0$ when $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 = \mathbf{0}$ for $\mathbf{v}_1 \in \text{Null}\,(A^r)$ and $\mathbf{v}_2 \in \text{Col}\,(A^r)$, because $x_1\mathbf{v}_1 = -x_2\mathbf{v}_2 \in \text{Null}\,(A^r) \cap \text{Col}\,(A^r) = \{\mathbf{0}\}$. Since the bases of these two subspaces are linearly independent, $\dim\big(\text{Null}\,(A^r) + \text{Col}\,(A^r)\big) = \dim \text{Null}\,(A^r) + \dim \text{Col}\,(A^r)$, and $\dim\big(\text{Null}\,(A^r) + \text{Col}\,(A^r)\big) = n = \dim \mathbb{V}$ according to the rank-nullity theorem (see Theorem 3.5). There is hence a positive integer r given a square matrix A such that

$$\mathbb{V} = \text{Null}\,(A^r) \oplus \text{Col}\,(A^r), \tag{11.2}$$

where $\oplus$ is a direct sum introduced in Definition 3.3.

These two subspaces, $\text{Null}\,(A^r)$ and $\text{Col}\,(A^r)$, are invariant under A:

- For $\mathbf{v} \in \text{Null}\,(A^r)$, $A\mathbf{v} \in \text{Null}\,(A^r)$ since $A^r(A\mathbf{v}) = A^{r+1}\mathbf{v} = A(A^r\mathbf{v}) = \mathbf{0}$;
- For $\mathbf{v} \in \text{Col}\,(A^r)$, $A\mathbf{v} \in \text{Col}\,(A^r)$ since $\mathbf{v} = A^r\mathbf{w}$ for some $\mathbf{w} \in \mathbb{V}$ and thus $A\mathbf{v} = A^r(A\mathbf{w})$.

Furthermore, it is easy to see that these two subspaces are invariant under the addition of αI for a scalar α. That is, $\text{Null}\,(A^r)$ and $\text{Col}\,(A^r)$ are invariant under $A + \alpha I$. Combining these observations, we arrive at the following theorem.

Theorem 11.2 *Let A be an $n \times n$ matrix and $\mathbb{V}$ be an n-dimensional vector space. Let*

$$\mathbb{V}_A = \{\mathbf{v} \in \mathbb{V} : A^j\mathbf{v} = \mathbf{0} \quad \textit{for some} \quad j\}.$$

Then there exists r_0 such that for any $r \geq r_0$

$$\mathbb{V}_A = \text{Null}\,(A^r)$$

and

$$\mathbb{V} = \text{Null}\,(A^r) \oplus \text{Col}\,(A^r),$$

where $\text{Null}\,(A^r)$ *and* $\text{Col}\,(A^r)$ *are invariant under $A + \alpha I$ for any scalar α.*

Let us think of a procedure to find linearly independent vectors in $\mathbb{V}_A$. Consider a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ recursively generated starting from a vector $\mathbf{v}_1 \in \mathbb{V}_A$ such that $\mathbf{v}_2 = A\mathbf{v}_1 \neq \mathbf{0}, \ldots, \mathbf{v}_k = A\mathbf{v}_{k-1} = A^{k-1}\mathbf{v}_1 \neq \mathbf{0}$ and $A^k\mathbf{v}_1 = A\mathbf{v}_k = \mathbf{0}$. In order to check their linear independence, assume $x_1\mathbf{v}_1 + \cdots + x_k\mathbf{v}_k =$

$\mathbf{0}$. For $\mathbf{v}_i$ with $i \geq 2$, $A^{k-1}\mathbf{v}_i = A^{k-1}A^{i-1}\mathbf{v}_1 = A^{i-2}(A^k\mathbf{v}_1) = \mathbf{0}$. Therefore, if we multiply both sides with A^{k-1}, leaving only $x_1A^{k-1}\mathbf{v}_1 = x_1\mathbf{v}_k = \mathbf{0}$. x_1 is thus 0. We repeat the same procedure by multiplying both sides of $x_2\mathbf{v}_2+\cdots+x_k\mathbf{v}_k = \mathbf{0}$ with A^{k-2}, and we see that $x_2 = 0$. By repeatedly applying this procedure, we get $x_1 = \cdots = x_k = 0$, and therefore, $\mathcal{B} = \{\mathbf{v}_1, A\mathbf{v}_1, \ldots, A^{k-1}\mathbf{v}_1\}$ is linearly independent.

Lemma 11.1 *Assume that a nonzero vector* $\mathbf{v} \in \mathbb{V}_A$ *satisfies* $A\mathbf{v} \neq \mathbf{0}, \ldots, A^{k-1}\mathbf{v} \neq \mathbf{0}$ *and* $A^k\mathbf{v} = \mathbf{0}$. *Then,* $\{\mathbf{v}, A\mathbf{v}, \ldots, A^{k-1}\mathbf{v}\}$ *is linearly independent.*

We say that the basic vectors in Lemma 11.1 exhibit a cyclic structure, and this structure plays a crucial role in decomposing a matrix into a Jordan form later. This lemma also says that the nilpotent degree r can not exceed n, the dimension of the underlying vector spaces, $\mathbb{R}^n$ or $\mathbb{C}^n$ as well as that we can set $r_0 \leq n$ in Theorem 11.2.

Direct Sum Decomposition of the Null Space of Nilpotent Matrices

We will analyze a nilpotent matrix A of degree r on $\mathbb{W} = \text{Null } A^r$. Since Null $A^k \subset$ Null A^{k+1} for any k,

$$\text{Null } A \subset \text{Null } A^2 \subset \cdots \subset \text{Null } A^{r-1} \subset \text{Null } A^r = \mathbb{W}. \tag{11.3}$$

By the definition of nilpotency, there exists at least one vector $\mathbf{w} \in \mathbb{W}$ such that $A^{r-1}\mathbf{w} \neq \mathbf{0}$. For this $\mathbf{w}$, $A^{r-k}\mathbf{w} \in \text{Null } A^k \setminus \text{Null } A^{k-1}$, and hence the subset inclusions in (11.3) are strict. We now use this nilpotency structure of A to decompose the subspace $\mathbb{W} = \text{Null } A^r$ as direct sums.

We first extend the notion of linear independence.

Definition 11.1 *Vectors* $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *are linearly independent of a subspace* $\mathbb{W}$ *if* $a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k \in \mathbb{W}$ *implies* $a_1 = \cdots = a_k = 0$.

When $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent of $\mathbb{W}$, they are also linearly independent since $\mathbf{0} \in \mathbb{W}$. We use this extended notion of linear independence to form a basis of Null A^r by finding vectors that are linearly independent of Null A^{k-1} and are included in Null $A^k \setminus$ Null A^{k-1}. With this extended definition of linear independence, we can derive a result critical for us to arrive at the Jordan form representation later in this chapter.

Theorem 11.3 *Let A be a square nilpotent matrix of degree r on* Null A^r *for $r \geq 1$. Then, there exist m and vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in$ Null A^r such that the non-zero vectors $A^j\mathbf{v}_\ell$, for $j \geq 0$ and $1 \leq \ell \leq m$, form a basis for* Null A^r. *Any vector linearly independent of* Null A^{r-1} *can be included in $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$.*

Proof: If a matrix A is nilpotent of degree r on Null A^r, then there exists a vector $\mathbf{v}$ such that $A^r\mathbf{v} = \mathbf{0}$ as well as $A^j\mathbf{v} \neq \mathbf{0}$ for $j < r$. The vector $\mathbf{w} = A\mathbf{v}$ satisfies $A^{r-1}\mathbf{w} = A^r\mathbf{v} = \mathbf{0}$ and $A^j\mathbf{w} = A^{j+1}\mathbf{v} \neq \mathbf{0}$ for $j < r-1$, which implies that A is nilpotent of degree $r-1$ on Null A^{r-1}. This observation enables us to use mathematical induction to prove this theorem:

- When $r = 1$, $\dim$ Null $A \geq 1$. We simply find a basis of Null A.
- Assume the theorem holds up to $r-1$. Let A be a nilpotent matrix of degree r on Null A^r. Let non-zero vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k \in$ Null A^r be linearly independent of Null A^{r-1}. These vectors are obtained maximally such that adding any other vectors in Null $A^r \setminus$ Null A^{r-1} to $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ causes linear dependence.[1] Therefore, $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ complemented by any basis of Null A^{r-1} forms a basis of Null A^r.

  Because these vectors $\mathbf{w}_i$ are in Null A^r, $A\mathbf{w}_i \in$ Null A^{r-1}. If we suppose $a_1A\mathbf{w}_1 + \cdots + a_kA\mathbf{w}_k = A(a_1\mathbf{w}_1 + \cdots + a_k\mathbf{w}_k) \in$ Null A^{r-2}, then $a_1\mathbf{w}_1 + \cdots + a_k\mathbf{w}_k \in$ Null A^{r-1} and it must be $a_1 = \cdots = a_k = 0$ since $\mathbf{w}_i$'s are linearly independent of Null A^{r-1}. Therefore, $A\mathbf{w}_1, \ldots, A\mathbf{w}_k$ are linearly independent of Null A^{r-2}.

  Since A is a nilpotent matrix of degree $r-1$ on Null A^{r-1}, by the induction hypothesis, there exist $\mathbf{v}_1, \ldots, \mathbf{v}_m$ in Null A^{r-1}, including all k vectors of $\{A\mathbf{w}_1, \ldots, A\mathbf{w}_k\}$, so that non-zero $A^j\mathbf{v}_\ell$'s form a basis of Null A^{r-1}. We can then form a basis of Null A^r by combining this basis of Null A^{r-1} consisting of $A^j\mathbf{v}_\ell$'s and $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$. Recall that our starting vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ can contain any linearly independent vectors of Null A^{r-1} since we did not impose any other condition on $\mathbf{w}_i$'s except the linear independence from Null A^{r-1}.

This completes the proof for all r. ■

Let us inspect the basis obtained by Theorem 11.3 more carefully. For each $\mathbf{v}_\ell$, we define r_ℓ to satisfy $A^{r_\ell}\mathbf{v}_\ell = \mathbf{0}$ and $A^{r_\ell-1}\mathbf{v}_\ell \neq \mathbf{0}$. With r_ℓ's, we

[1] As a specific example, we can find $\mathbf{w}_1, \ldots, \mathbf{w}_k$ in Null A^r by applying the Gram-Schmidt procedure on the basis of Null A^{r-1}.

can rewrite $\{A^j\mathbf{v}_\ell : j \geq 0, 1 \leq \ell \leq m\} = \cup_{\ell=1}^m \{\mathbf{v}_\ell, \ldots, A^{r_\ell-1}\mathbf{v}_\ell\}$. If we set $\mathbb{V}_\ell = \mathrm{span}\{\mathbf{v}_\ell, \ldots, A^{r_\ell-1}\mathbf{v}_\ell\}$, we can decompose Null A^r as

$$\text{Null } A^r = \mathbb{V}_1 \oplus \cdots \oplus \mathbb{V}_m\,, \quad r_\ell = \dim \mathbb{V}_\ell. \tag{11.4}$$

Furthermore, each $\mathbb{V}_\ell$ is invariant to A, since $\mathbf{v} \in \mathbb{V}_\ell$ implies $A\mathbf{v} \in \mathbb{V}_\ell$. Even though $\mathbb{V}_\ell$ is not determined uniquely as $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are not unique, the dimensions of $\mathbb{V}_\ell$'s are uniquely determined up to the order of $\mathbb{V}_\ell$'s in the direct sum. See Table 11.1 for an example.

Set $d_k = \dim \text{Null } A^k - \dim \text{Null } A^{k-1}$ for $k = 1, \ldots, r$. $d_k \geq 1$ since Null $A^k \setminus$ Null $A^{k-1} \neq \emptyset$. Then, among the basic vectors in a basis $\{A^j\mathbf{v}_\ell : j \geq 0, 1 \leq \ell \leq m\}$ of Null A^r, d_k of them are included in Null $A^k \setminus$ Null A^{k-1}. If $A^j\mathbf{v}_\ell$ is one of d_k basic vectors in Null $A^k \setminus$ Null A^{k-1}, the basis of $\mathbb{V}_\ell$ must contain $A^{k-1+j}\mathbf{v}_\ell$, which implies $r_\ell \geq k + j$. The dimension of $\mathbb{V}_\ell$ is thus at least k. Therefore, we can see that there are d_k-many $\mathbb{V}_\ell$'s of dimension at least k in the decomposition in (11.4). It is also easy to see $d_{k+1} \leq d_k$. Then, with $d_{r+1} = 0$, there are $(d_k - d_{k+1})$-many $\mathbb{V}_\ell$'s of dimension k. The sequence $(d_1, \ldots, d_r)$ is uniquely determined for a matrix A, and so are the number and dimensions of the subspaces in the decomposition (11.4). We often call $\mathbb{V}_\ell$ a cyclic subspace in order to emphasize the cyclic structure of $\mathrm{span}\{\mathbf{v}_\ell, \ldots, A^{r_\ell-1}\mathbf{v}_\ell\}$. We summarize this observation in the following theorem.

Theorem 11.4 *Let A be a nilpotent square matrix of degree r on* Null A^r. *Then, there exists a unique number m and r_ℓ's such that*

$$\text{Null } A^r = \mathbb{V}_1 \oplus \cdots \oplus \mathbb{V}_m$$

where $\mathbb{V}_\ell = \mathrm{span}\{\mathbf{v}_\ell, \ldots, A^{r_\ell-1}\mathbf{v}_\ell\}$. Furthermore, the number of summands $\mathbb{V}_\ell$ of dimension at least k is $\dim \text{Null } A^k - \dim \text{Null } A^{k-1}$. *Therefore, the decomposition is unique up to the number of summands and their dimensions.*

Consider Table 11.1 demonstrating this decomposition of a nilpotent matrix of degree 5. Each column in this table corresponds to the basis of $\mathbb{V}_\ell$, and if we combine vectors in the bottom-k rows, we get a basis of Null A^k. For instance, we form a basis of Null A^3 by combining vectors in the rows of $k = 1, 2$, and 3. Specifically, vectors in the row of $k = 1$ are the eigenvectors associated with the eigenvalue 0 of A as well as constitute a basis of Null A.

We can apply this result on nilpotency to $A - \lambda I$ where λ is an eigenvalue of A. In the next section, we show that $A - \lambda I$ is a nilpotent matrix of degree r on $\text{Null}\,(A - \lambda I)^r$ for some $r \leq n$. We then obtain a so-called Jordan block using Theorem 11.3 and 11.4.

Table 11.1: Demonstration of a Decomposition of Nilpotent Matrix

k	$\mathbb{V}_1$	$\mathbb{V}_2$	$\mathbb{V}_3$	$\mathbb{V}_4$	$\mathbb{V}_5$	$\mathbb{V}_6$	$\mathbb{V}_7$	$\dim \text{Null } A^k$	d_k
5	$\mathbf{v}_1$	$\mathbf{v}_2$						25	2
4	$A\mathbf{v}_1$	$A\mathbf{v}_2$	$\mathbf{v}_3$	$\mathbf{v}_4$	$\mathbf{v}_5$			23	5
3	$A^2\mathbf{v}_1$	$A^2\mathbf{v}_2$	$A\mathbf{v}_3$	$A\mathbf{v}_4$	$A\mathbf{v}_5$			18	5
2	$A^3\mathbf{v}_1$	$A^3\mathbf{v}_2$	$A^2\mathbf{v}_3$	$A^2\mathbf{v}_4$	$A^2\mathbf{v}_5$	$\mathbf{v}_6$		13	6
1	$A^4\mathbf{v}_1$	$A^4\mathbf{v}_2$	$A^3\mathbf{v}_3$	$A^3\mathbf{v}_4$	$A^3\mathbf{v}_5$	$A\mathbf{v}_6$	$\mathbf{v}_7$	7	7

11.3 Nilpotency of Eigenspace

Generalized Eigenvectors

Given an eigenvalue λ of an $n \times n$ matrix A, all the vectors in $\text{Null}\,(A-\lambda I)$ are its eigenvectors, and the dimension of this null space is the geometric multiplicity. If the geometric multiplicity is smaller than the algebraic multiplicity of λ, we run into an issue when diagonalizing A. Before studying it further, we derive the following corollary by applying Theorem 11.2 to $A - \lambda I$.

Corollary 11.1 *Let A be an $n \times n$ complex matrix with an eigenvalue λ. Then, there exists r_λ such that for any $r \geq r_\lambda$*

$$\{\mathbf{v} \in \mathbb{V} : (A - \lambda I)^j \mathbf{v} = \mathbf{0} \quad \textit{for some } j\} = \text{Null}\,(A - \lambda I)^r$$

and

$$\mathbb{C}^n = \text{Null}\,(A - \lambda I)^r \oplus \text{Col}\,(A - \lambda I)^r,$$

where $\text{Null}\,(A-\lambda I)^r$ *and* $\text{Col}\,(A-\lambda I)^r$ *are invariant under $A-cI$ for any scalar c.*

We can take it as granted $r_\lambda \leq n$ since the number of basic vectors in the basis of $\text{Null}\,(A-\lambda I)^{r_\lambda}$ is at most n. Therefore, $\text{Null}\,(A-\lambda I)^{r_\lambda} = \text{Null}\,(A-\lambda I)^n$ holds and we investigate the nilpotent structure of $A - \lambda I$ through the null space of $(A - \lambda I)^n$. As a work-around when the geometric multiplicity is smaller than the algebraic multiplicity, we call the vectors in $\text{Null}\,(A-\lambda I)^n$ generalized eigenvectors.[2] We then search for a matrix that exhibits a simple pattern of non-zero entries, while being similar to the original matrix, under these generalized eigenvectors as a basis.

[2] A generalized eigenvector may or may not be an eigenvector.

Interestingly, the only eigenvalue of A on $\mathrm{Null}\,(A-\lambda I)^n$ is λ itself. Let $(\mu, \mathbf{v})$ with $\mathbf{v} \in \mathrm{Null}\,(A-\lambda I)^n$ is an eigenpair of A. Then, $(A-\lambda I)\mathbf{v} = A\mathbf{v} - \lambda\mathbf{v} = (\mu-\lambda)\mathbf{v}$ and $\mathbf{0} = (A-\lambda I)^n\mathbf{v} = (\mu-\lambda)^n\mathbf{v}$, which implies $\mu = \lambda$. Furthermore, the generalized eigenvectors corresponding to different eigenvalues are linearly independent as so are the eigenvectors corresponding to different eigenvalues. In other words, for different eigenvalues $\lambda_1 \neq \lambda_2$,

$$\mathrm{Null}\,(A-\lambda_1 I)^n \cap \mathrm{Null}\,(A-\lambda_2 I)^n = \{\mathbf{0}\}. \tag{11.5}$$

Suppose that $\mathbf{0} \neq \mathbf{v} \in \mathrm{Null}\,(A-\lambda_1 I)^n \cap \mathrm{Null}\,(A-\lambda_2 I)^n$. Then, there exists $k \geq 0$ such that $(A-\lambda_2 I)^{k+1}\mathbf{v} = \mathbf{0}$ and $(A-\lambda_2 I)^k\mathbf{v} \neq \mathbf{0}$ since $\mathbf{v} \in \mathrm{Null}\,(A-\lambda_2 I)^n$. Set $\mathbf{w} = (A-\lambda_2 I)^k\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{w} = \lambda_2\mathbf{w}$. Since the product of $A-\lambda_1 I$ and $A-\lambda_2 I$ commutes,

$$(A-\lambda_1 I)^n\mathbf{w} = (A-\lambda_1 I)^n(A-\lambda_2)^k\mathbf{v} = (A-\lambda_2)^k(A-\lambda_1 I)^n\mathbf{v} = \mathbf{0}\,,$$

which implies $\mathbf{w} \in \mathrm{Null}\,(A-\lambda_1 I)^n$. Therefore $(\lambda_2, \mathbf{w})$ is an eigenpair of A on $\mathrm{Null}\,(A-\lambda_1 I)^n$, which contradict to the uniqueness of the eigenvalue of A on $\mathrm{Null}\,(A-\lambda_1 I)^n$, and (11.5) holds.

An $n \times n$ complex matrix A has n eigenvalues, including multiple roots, and is similar to an upper triangular matrix with the eigenvalues on its diagonal, according to the Schur triangularization, Theorem 10.5. That is, $A = QUQ^{\mathsf{H}}$ for some upper triangular matrix U and unitary matrix Q. Assume m distinct eigenvalues, $\lambda_1, \ldots, \lambda_m$, and the same eigenvalues are located adjacently on diagonals of U. If the algebraic multiplicity of λ_i is k_i, the k_i diagonal entries in the i-th diagonal block of U are all λ_i. Hence, the i-th diagonal block of $U-\lambda_i I$ is a $(k_i \times k_i)$ upper triangular matrix with all zero diagonal entries. In addition, for any k, $(A-\lambda_i I)^k = Q(U-\lambda_i I)^k Q^{\mathsf{H}}$ since

$$A - \lambda_i I = QUQ^{\mathsf{H}} - \lambda_1 QQ^{\mathsf{H}} = Q(U-\lambda_i I)Q^{\mathsf{H}}.$$

According to Fact 2.2, all entries of the i-th $(k_i \times k_i)$ diagonal block of $(U-\lambda_i I)^{k_i}$ are all zeros and the remaining diagonal entries of $(U-\lambda_i I)^{k_i}$ are $(\lambda_j-\lambda_i)^{k_i} \neq 0$. The rank of $(U-\lambda_i I)^{k_i}$ is thus $n-k_i$, and so is that of $(A-\lambda_i I)^{k_i}$, since Q is invertible. This results in $\dim \mathrm{Null}\,(A-\lambda_i I)^{k_i} = \dim \mathrm{Null}\,(U-\lambda_i I)^{k_i} = k_i$. Since k_i's are the multiplicities of multiple roots of the n-th order characteristic equation and $\mathrm{Null}\,(A-\lambda_i I)^{k_i} \subset \mathrm{Null}\,(A-\lambda_i I)^n$ holds,

$$\dim(\mathbb{C}^n) \;=\; n \;=\; k_1 + \cdots + k_m$$

$$\begin{aligned} &= \dim \mathrm{Null}\,(A-\lambda_1 I)^{k_1} + \cdots + \dim \mathrm{Null}\,(A-\lambda_m I)^{k_m} \\ &\leq \dim \mathrm{Null}\,(A-\lambda_1 I)^{n} + \cdots + \dim \mathrm{Null}\,(A-\lambda_m I)^{n}. \end{aligned}$$

On the other hand, each $\mathrm{Null}\,(A-\lambda_i)^n \subset \mathbb{C}^n$ implies

$$\mathbb{C}^n \supset \mathrm{Null}\,(A-\lambda_1 I)^n + \cdots + \mathrm{Null}\,(A-\lambda_m I)^n\,.$$

Combined with (11.5), we get an inclusion in terms of direct sums as

$$\mathbb{C}^n \supset \mathrm{Null}\,(A-\lambda_1 I)^n \oplus \cdots \oplus \mathrm{Null}\,(A-\lambda_m I)^n\,,$$

which implies $n \geq \dim \mathrm{Null}\,(A-\lambda_1 I)^n + \cdots + \dim \mathrm{Null}\,(A-\lambda_m I)^n$. Combining these inequalities, we get

$$n = k_1 + \cdots + k_m \leq \dim \mathrm{Null}\,(A-\lambda_1 I)^n + \cdots + \dim \mathrm{Null}\,(A-\lambda_m I)^n \leq n\,.$$

Then, $k_i \leq \dim \mathrm{Null}\,(A-\lambda_i I)^n$ implies $k_i = \dim \mathrm{Null}\,(A-\lambda_i I)^n$, and the above set inclusion is an equality in fact. That is, we can decompose $\mathbb{C}^n$ as

$$\mathbb{C}^n \quad = \quad \mathrm{Null}\,(A-\lambda_1 I)^n \oplus \cdots \oplus \mathrm{Null}\,(A-\lambda_m I)^n$$

We summarize this observation into the following theorem.

Theorem 11.5 *Let A be an $n \times n$ matrix, and $\lambda_1, \ldots, \lambda_m$ its eigenvalues. Then,*

$$\mathbb{C}^n = \mathrm{Null}\,(A-\lambda_1 I)^n \oplus \cdots \oplus \mathrm{Null}\,(A-\lambda_m I)^n\,. \tag{11.6}$$

Once we obtain a basis of each subspace $\mathrm{Null}\,(A-\lambda_i I)^n$, we form a basis of $\mathbb{C}^n$ by combining them.

11.4 The Second Big Theorem: Jordan Normal Form Theorem

Assume λ_i is an eigenvalue of an $n \times n$ complex matrix A with its algebraic multiplicity k_i. From Corollary 11.1, $A - \lambda_i I$ is a nilpotent matrix on $\mathrm{Null}\,(A - \lambda_i I)^n$, which allows us to apply Theorem 11.3 and 11.4 to the nilpotent matrix $A - \lambda_i I$. $\mathrm{Null}\,(A-\lambda_i I)^n$ is decomposed as the direct sum of the subspaces with bases $\{\mathbf{v}_\ell, \ldots, (A-\lambda_i I)^{r_\ell - 1}\mathbf{v}_\ell\}$, according to Theorem 11.3. With $\mathbf{w}_k = (A-\lambda_i I)^{k-1}\mathbf{v}_\ell$, the basis is $\{\mathbf{w}_1, \ldots, \mathbf{w}_{r_\ell}\}$, and the basic vectors are related to each other by

$$\mathbf{w}_{k+1} = (A-\lambda_i I)\mathbf{w}_k \quad \text{for } k = 1, \ldots, r_\ell - 1\,, \quad (A-\lambda_i I)\mathbf{w}_{r_\ell} = \mathbf{0}.$$

We can rewrite it as

$$A\mathbf{w}_k = \lambda_i \mathbf{w}_k + \mathbf{w}_{k+1} \quad \text{for } k = 1, \ldots, r_\ell - 1\,, \quad A\mathbf{w}_{r_\ell} = \lambda_i \mathbf{w}_{r_\ell}.$$

Let us define a Jordan block J_ℓ as the following $r_\ell \times r_\ell$ matrix:

$$J_\ell = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda_i & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_i \end{bmatrix} = \lambda_i I + \begin{bmatrix} \mathbf{0} \,|\, \mathbf{e}_1 \,|\, \cdots \,|\, \mathbf{e}_{r_\ell - 1} \end{bmatrix}.$$

When $W_\ell = \begin{bmatrix} \mathbf{w}_{r_\ell} \,|\, \cdots \,|\, \mathbf{w}_1 \end{bmatrix}$,[3]

$$AW_\ell = W_\ell J_\ell. \tag{11.7}$$

In other words, linear transformation, defined by A, in a subspace with a basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_{r_\ell}\}$, can be expressed using a Jordan block J_ℓ. For each $\mathbf{v}_\ell$, there exists a Jordan block J_ℓ that satisfies (11.7). Hence, for $W_{\lambda_i} = \begin{bmatrix} W_1 \,|\, \cdots \,|\, W_{r_\ell} \end{bmatrix}$ and $J_{\lambda_i} = \operatorname{diag}(J_1, \ldots, J_{r_\ell})$,

$$AW_{\lambda_i} = W_{\lambda_i} J_{\lambda_i}, \tag{11.8}$$

where W_{λ_i} and J_{λ_i} are $n \times k_i$ and $k_i \times k_i$ matrices, respectively.

The sizes and number of Jordan blocks that constitute J_{λ_i} are uniquely determined by Theorem 11.4. There exist W_{λ_i} and J_{λ_i} that satisfy (11.8), for each eigenvalue λ_i for A. If we let $W = \begin{bmatrix} W_{\lambda_1} \,|\, \cdots \,|\, W_{\lambda_m} \end{bmatrix}$ and $J = \operatorname{diag}(J_{\lambda_1}, \ldots, J_{\lambda_m})$,

$$AW = WJ, \tag{11.9}$$

and we call J a Jordan (normal) form. Since W, which takes basic vectors as its columns, is invertible, $J = W^{-1}AW$, meaning that A and J are similar.

We summarize this result as the following theorem.

Theorem 11.6 (Jordan Normal Form Theorem) *Any $n \times n$ matrix is similar to a Jordan normal form. The Jordan form is unique up to the number and sizes of Jordan blocks.*

[3]Be careful with the column indices, as they are flipped, because it is a convention to define a Jordan block as an upper triangular matrix with the super-diagonal set to 1.

A Jordan form J is not a diagonal matrix but closely resembles it. Using this special structure, we can often compute J^k more efficiently than A^k, by replacing the diagonal matrix in power iteration, from Section 9.7, with such a Jordan block. It is less efficient than using a diagonal matrix but is often more efficient than using the original matrix directly.

Example 11.1 Let us find the Jordan form of a 4×4 matrix A, where

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

Try to compute the eigenvalues of A by hand. It should be non-trivial and a time-consuming hand calculation is required. Or, you can give up after a lengthy process of simplifying its characteristic equation and use your favorite numerical linear algebra software. Such a algebra engine would tell you that

$$P = \begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \text{ with } P^{-1} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Because

$$B = P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix},$$

we see that B is a Jordan form of A. From the Jordan form B which is similar to the original matrix A, we recognize that 1,2, and 4 are eigenvalues of A and the 4 has a geometric multiplicity of 1 even though its algebraic multiplicity is 2.

Let us further investigate the cyclic subspace of

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

Since the last two columns of

$$(A-4I)^2 = \begin{bmatrix} 0 & -9 & -5 & -5 \\ 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 4 \end{bmatrix}$$

match exactly, we get

$$\text{Null}\,(A-4I)^2 = \text{span}\{(0,0,1,-1)^\top, (1,0,0,0)^\top\} = \{(a,0,b,-b)^\top : a,b \in \mathbb{R}\}.$$

For any non-zero vector $(a,0,b,-b)^\top$ with $a+b \neq 0$,

$$(A-4I)(a,0,b,-b)^\top = (a+b)(1,0,-1,1)^\top \neq \mathbf{0}.$$

Say $\mathbf{v} = (1,0,1,-1)^\top$ and $(A-4I)\mathbf{v} = \mathbf{w} = 2(1,0,-1,1)^\top$. Then, $(A-4I)\mathbf{w} = \mathbf{0}$ and $(4,\mathbf{w})$ is an eigenpair of A. Therefore, $\{\mathbf{v},\mathbf{w}\}$ is a basis for the cyclic subspace of $A-4I$. ∎

Chapter 12

Homework Assignments

Chapter 1

1. Compute $\begin{bmatrix} 1 \\ 0 \\ 2 \\ -3 \end{bmatrix} + \begin{bmatrix} \sqrt{2} \\ \pi \\ -2 \\ 0 \end{bmatrix}$ and $\sqrt{3}\begin{bmatrix} \sqrt{2} \\ \pi \\ -2 \\ 0 \end{bmatrix}$.

2. Compute $\begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 9 \\ 2 & 6 & 5 \\ 3 & 5 & 8 \end{bmatrix} + 2\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$.

3. Let $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. Compute $A\mathbf{v}$.

Chapter 2

1. Prove the following properties of symmetric matrices.

   (a) Any symmetric matrix is square.

   (b) Every diagonal matrix is symmetric.

   (c) For any matrix A, both $A^\top A$ and $AA^\top$ are symmetric.

   (d) $ADA^\top$ and $A^\top DA$ are symmetric when D is a diagonal matrix.

2. Consider the following simultaneous linear equations in 3 unknowns, $\mathbf{x} = (u, v, w)^\top$:

$$\begin{cases} u & -4v & +7w & = & -9 \\ 2u & -6v & +9w & = & -10 \\ u & -2v & +5w & = & -7 \end{cases}$$

   (a) Convert the equations in matrix-vector form as $A\mathbf{x} = \mathbf{b}$. What are A and $\mathbf{b}$?

   (b) Decompose A in the form of LDU where L and U are lower and upper triangular with unit diagonals, and D is a diagonal matrix. What are L, D, and U ?

   (c) Find $\mathbf{x}$ by solving $DU\mathbf{x} = L^{-1}\mathbf{b}$.

   (d) Compute A^{-1}.

3. Repeat 2 with randomly generated 10×10 matrix A and $\mathbf{b} \in \mathbb{R}^{10}$.

4. Prove that matrix multiplication is associative $(AB)C = A(BC)$ and distributive $A(B + C) = AB + AC$, $(B + C)D = BD + CD$.

5. Find an example showing that matrix multiplication is not commutative, that is, find two matrices A and B such that $AB \neq BA$

6. Show that $AI_n = A = I_m A$ for any $m \times n$ matrix A.

7. Check the following multiplication rule for two block matrices:

$$AB = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

8. Compute A^2, A^3, and A^4 where

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

Chapter 3

1. Show that $\{\mathbf{0}\}$, $\{c\mathbf{x} : c \in \mathbb{R}\}$ for $\mathbf{x} \in \mathbb{V}$, and $\{c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n : c_1, \ldots, c_n \in \mathbb{R}\}$ for $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{V}$ are subspaces of a vector space $\mathbb{V}$.

2. Show that $\{(x, y) : x \geq 0, y \geq 0\}$ is not a subspace of $\mathbb{R}^2$.

3. Show that the set of all $n \times n$ lower triangular matrices is a vector space and a subspace of the vector space consisting of all $n \times n$ matrices. Do the same work for the set of all $n \times n$ symmetric matrices.

4. Recall the definition of Cartesian product of two sets, $A \times B = \{(a, b) : a \in A, b \in B\}$ and $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. Show that $\mathbb{R}^2 = (\mathbb{R} \times \{0\}) \oplus (\{0\} \times \mathbb{R})$.

5. Let U be the row echelon form of A. Show that the vectors, in the null space of U, obtained by setting 1 for a single free variable and other free variables as zeros constitute a basis for the null space of A.

6. Consider the following 4×5 matrix A:

$$A = \begin{bmatrix} 2 & 2 & 1 & -6 & 4 \\ 4 & 4 & 1 & 10 & 13 \\ 6 & 6 & 0 & 20 & 19 \\ 8 & 8 & 1 & 14 & 23 \end{bmatrix}.$$

   (a) Compute the row echelon form and reduced row echelon form of A.

   (b) What is the rank of A?

   (c) Characterize the null space of A.

   (d) What is the dimension of $\text{Null}\,(A)$?

   (e) Find a maximally independent set of column vectors of A.

   (f) What is the dimension of $\text{Col}\,(A)$?

7. Let a set $\mathbb{V}$ include all polynomials of degree n or less, $\mathbb{V} = \{a_0 + a_1 t + \cdots + a_n t^n : a_0, a_1, \ldots, a_n \in \mathbb{R}\}$. Let T be a transform that maps a polynomial $f(t)$ to the polynomial $\int_0^t f(s)ds$. Note that the n is a fixed integer in this question.

   (a) Show that $\mathbb{V}$ is a vector space over the multiplication scalar field $\mathbb{R}$.

   (b) What is the dimension of $\mathbb{V}$?

   (c) To make T be a map from $\mathbb{V}$ into $\mathbb{W}$, what should be $\mathbb{W}$ if $\mathbb{W}$ is a vector space?

   (d) Is T a linear map?

   (e) Find bases $\mathcal{B}_{\mathbb{V}}$ and $\mathcal{B}_{\mathbb{W}}$ of $\mathbb{V}$ and $\mathbb{W}$.

(f) Find the transform matrix A representing T with respect to the bases $\mathcal{B}_{\mathbb{V}}$ and $\mathcal{B}_{\mathbb{W}}$ chosen above.

8. In a finite-dimensional vector space, show that any linearly independent set of vectors can be extended to a basis.

Chapter 4

1. Let a vector space $\mathbb{V}$ include all polynomials of degrees 2 or less, $\mathbb{V} = \{a_0 + a_1 t + a_2 t^2 : a_0, a_1, a_2 \in \mathbb{R}\}$. For two polynomials $f(t)$ and $g(t)$ in $\mathbb{V}$, we define an inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(t) g(t) dt \,.$$

(a) Compute $|f|$, $|g|$, and $|f - g|$ for $f(t) = t$ and $g(t) = t^2$.

(b) Show that $\{1, t, t^2\}$ is linearly independent.

(c) For $\mathbb{W} = \text{span}\{1, t^2\}$, compute $\mathbf{P}_{\mathbb{W}}(t)$.

(d) Show that $\mathcal{B} = \{1, t, t^2\}$ is a basis for $\mathbb{V}$.

(e) Find a matrix representation of the inner product $\langle f, g \rangle$ with respect to the basis $\mathcal{B}$.

2. Consider the Euclidean vector space $\mathbb{R}^4$. For two vectors $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top$ and $\mathbf{y} = (y_1, y_2, y_3, y_4)^\top$ in $\mathbb{R}^4$, we define a *non-standard* inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = 2x_1 y_1 + \frac{2}{3}(x_1 y_3 + x_2 y_2 + x_3 y_1) + \frac{2}{5}(x_2 y_4 + x_3 y_3 + x_4 y_2) + \frac{2}{7} x_4 y_4 \,,$$

or, in matrix form as

$$\langle \mathbf{x}, \mathbf{y} \rangle = [x_1, x_2, x_3, x_4] \begin{bmatrix} 2 & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{2}{5} \\ \frac{2}{3} & 0 & \frac{2}{5} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{2}{7} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} .$$

(a) Compute $|\mathbf{x}|$, $|\mathbf{y}|$, and $|\mathbf{x} - \mathbf{y}|$ for $\mathbf{x} = (1, 0, 0, 0)^\top$ and $\mathbf{y} = (0, 0, 1, 0)^\top$.

(b) Find an orthonormal basis of $\text{span}\{(1, 0, 0, 0)^\top, (0, 0, 1, 0)^\top\}$.

(c) Let $\mathbb{W} = \text{span}\{(1, 0, 0, 0)^\top, (0, 0, 1, 0)^\top\}$. Find an orthonormal basis of $\mathbb{W}^\perp$.

3. Let a vector space $\mathbb{V}$ include all polynomials of degree 3 or less, $\mathbb{V} = \{a_0 + a_1 t + a_2 t^2 + a_3 t^3 : a_0, a_1, a_2, a_3 \in \mathbb{R}\}$. For two polynomials $f(t)$, $g(t) \in \mathbb{V}$, we define an inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(t) g(t) dt \,.$$

   (a) Compute $|f|$, $|g|$, and $|f - g|$ for $f(t) = 1$ and $g(t) = t^2$.

   (b) Find an orthonormal basis of $\text{span}\{1, t^2\}$.

   (c) Let $\mathbb{W} = \text{span}\{1, t^2\}$. Find an orthonormal basis of $\mathbb{W}^\perp$.

   (d) Find the matrix representation of the inner product $\langle f, g \rangle$ with respect to the basis $\{1, t, t^2, t^3\}$ of $\mathbb{V}$.

4. Compare questions 2 and 3.

Chapter 5

1. Find singular values and singular vectors of the following rotation matrix:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

2. Let a 4×4 matrix A is given as

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

   (a) Find singular values and right-/left-singular vectors of A.

   (b) Find $\|A\|_2$ and $\|A\|_F$.

   (c) Find the pseudoinverse A^+ of A.

   (d) Find a 4×4 matrix B of rank 2 that minimizes $\|A - B\|_2$.

   (e) Can you find eigenvalues and eigenvectors of A by hand calculation?

3. Let a 4×4 matrix A is given as

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

(a) Find singular values and right-/left-singular vectors of A.

(b) Can you find eigenvalues and eigenvectors of A by hand calculation?

(c) Find rank A.

4. Suppose A is a 3×4 matrix given as $A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 6 & 3 & -3 & 3 \\ -2 & -1 & 1 & -1 \end{bmatrix}$.

(a) Find eigenvalues and eigenvectors of $A^\top A$.

(b) Find eigenvalues and eigenvectors of $AA^\top$.

(c) Is there a 3×4 matrix B of rank 2 that minimizes $\|A - B\|_2$?

Chapter 7

1. Let A and B be two $n \times n$ positive definite matrices. Consider the following sum of two quadratic forms

$$f(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^\top A(\mathbf{x} - \mathbf{a}) + (\mathbf{x} - \mathbf{b})^\top B(\mathbf{x} - \mathbf{b})$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Re-formulate the sum as a single quadratic form: that is, find C, $\mathbf{d}$, and r such that

$$f(\mathbf{x}) = (\mathbf{x} - \mathbf{d})^\top C(\mathbf{x} - \mathbf{d}) + r\,.$$

2. Let A be an $n \times n$ symmetric positive definite matrix. Denote its eigenvalues as $\lambda_1, \ldots, \lambda_n$. Let B be the square root of A, that is, $A = B^2$.

(a) Let $A = U^\top U$ be the Cholesky factorization of A. Find $|\det U|$.

(b) Find the eigenvalues of B.

(c) Let $B = QR$ be the QR-decomposition of B. Find $|\det R|$.

(d) Does there exist the square root C of B? If yes, what are the eigenvalues of C?

3. Let A be an $n \times n$ symmetric positive definite matrix whose spectral decomposition is described as $A = V\Lambda V^\top$ where V is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ for $\lambda_i > 0$. We also define $\Lambda^{1/k} = \mathrm{diag}(\lambda_1^{1/k}, \ldots, \lambda_n^{1/k})$.

(a) Characterize a symmetric positive definite matrix B_2 satisfying $A = B_2^2$ in terms of V and Λ.

(b) Characterize a symmetric positive definite matrix B_k satisfying $A = B_k^k$ in terms of V and Λ for every positive integer k.

(c) What would be $\lim_{k\to\infty} B_k$? Validate your answer as reasonably as possible.

Chapter 8

1. Compute the determinants of A, U, $U^\top$, U^{-1}, and M where

$$A = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix} \begin{bmatrix} 3 & -2 & 2 \end{bmatrix}, \quad U = \begin{bmatrix} 4 & 4 & 8 & 7 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 3 \end{bmatrix}, \quad M = \begin{bmatrix} 0 & 0 & 0 & 3 \\ 0 & 0 & 2 & 6 \\ 0 & 1 & 2 & 2 \\ 4 & 4 & 8 & 7 \end{bmatrix}.$$

2. Let A be an $n \times n$ tridiagonal matrix of

$$A = \begin{bmatrix} 1 & -1 & & & & \\ 1 & 1 & -1 & & & \\ & 1 & 1 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 1 & -1 \\ & & & & 1 & 1 \end{bmatrix}.$$

Find $\det A$.

3. Find the determinant of the following Vandermonde matrix:

$$V_3 = \begin{bmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{bmatrix}.$$

4. Find the determinant of the following rotation matrix:

$$\begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{bmatrix}.$$

Chapter 9

1. Show that the set of eigenvectors associated with a single eigenvalue is a subspace if we add the origin to the set.

2. Find all eigenvalues of A, B, U, U^{-1}, and T where

$$A = \begin{bmatrix}1\\2\\1\end{bmatrix}\begin{bmatrix}1 & 2 & 1\end{bmatrix}, \quad B = \begin{bmatrix}1\\2\\1\end{bmatrix}\begin{bmatrix}2 & 1 & 2\end{bmatrix}, \quad U = \begin{bmatrix}4 & 4 & 8\\0 & 1 & 2\\0 & 0 & 2\end{bmatrix},$$

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

3. Find all eigenvectors of A in 2.

4. Find all eigenvalues and eigenvectors of the following rotation matrix:

$$\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}.$$

Chapter 13

Problems

13.1 Problems for Chapter 1 $\sim$ 4

Problem Set 1

1. Suppose A is a 3×4 matrix and $\mathbf{b}$ is a vector in $\mathbb{R}^3$.

   (a) Give an example of the matrix A with rank 2. No entry of A is allowed to be zero and PA admits an LU decomposition with $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ whereas A itself does not allow LU decomposition.

   (b) For the matrix A in (a), give two examples of $\mathbf{b}$, one of which allows a solution of $A\mathbf{x} = \mathbf{b}$ and the other one doesn't.

   (c) Find two vectors in $\mathrm{Col}\,(A)$ closest to the two vectors in (b), respectively.

2. Suppose A and B are $n \times m$ and $m \times n$ matrices, respectively.

   (a) Recalling that each column of AB is a linear combination of the columns of A, prove that $\mathrm{rank}(AB) \leq \mathrm{rank}(A)$. Can we conclude $\mathrm{rank}(AB) \leq \mathrm{rank}(B)$ by straightforwardly applying the above statement? If yes, why?

   (b) Assume $AB = I_n$ where I_n is the identity matrix. What is the rank of A? If $m = n$, what is BA?

(c) Assume $AB = I_n$ where I is the identity matrix and $m > n$. Does $A\mathbf{x} = \mathbf{b}$ have a solution for any $\mathbf{b} \in \mathbb{R}^n$? Does $B\mathbf{y} = \mathbf{c}$ have a solution for any $\mathbf{c} \in \mathbb{R}^m$?

3. Let a vector space $\mathbb{V}$ include all polynomials of degree 3 or less, $\mathbb{V} = \{a_0 + a_1 t + a_2 t^2 + a_3 t^3 : a_0, a_1, a_2, a_3 \in \mathbb{R}\}$. Let T be a transformation that maps a polynomial $f(t)$ to the derivative $f'(t)$.

   (a) To make T a map from $\mathbb{V}$ into $\mathbb{W}$, what should be $\mathbb{W}$ if $\mathbb{W}$ is a vector space?

   (b) Find bases $\mathcal{B}_\mathbb{V}$ and $\mathcal{B}_\mathbb{W}$ of $\mathbb{V}$ and $\mathbb{W}$.

   (c) Find the matrix A representing the transform T with respect to the bases $\mathcal{B}_\mathbb{V}$ and $\mathcal{B}_\mathbb{W}$ chosen above.

4. Let a vector space $\mathbb{V}$ include all polynomials of degree 3 or less, $\mathbb{V} = \{a_0 + a_1 t + a_2 t^2 + a_3 t^3 : a_0, a_1, a_2, a_3 \in \mathbb{R}\}$. For two polynomials $f(t)$, $g(t) \in \mathbb{V}$, we define an inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(t) g(t) dt \,.$$

   (a) Compute $|f|$, $|g|$, and $|f - g|$ for $f(t) = t$ and $g(t) = t^2$.

   (b) Find an orthonormal basis of $\text{span}\{1, t^2\}$.

   (c) Let $\mathbb{W} = \text{span}\{t^2\}$. Find a basis of $\mathbb{W}^\perp$.

5. Consider the Euclidean vector space $\mathbb{R}^4$. For two vectors $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top$ and $\mathbf{y} = (y_1, y_2, y_3, y_4)^\top$ in $\mathbb{R}^4$, we define a *non-standard* inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = 2x_1 y_1 + \frac{2}{3}(x_1 y_3 + x_2 y_2 + x_3 y_1) + \frac{2}{5}(x_2 y_4 + x_3 y_3 + x_4 y_2) + \frac{2}{7} x_4 y_4 \,,$$

   or, in matrix form as

$$\langle \mathbf{x}, \mathbf{y} \rangle = [x_1, x_2, x_3, x_4] \begin{bmatrix} 2 & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{2}{5} \\ \frac{2}{3} & 0 & \frac{2}{5} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{2}{7} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} .$$

   (a) Compute $|\mathbf{x}|$, $|\mathbf{y}|$, and $|\mathbf{x} - \mathbf{y}|$ for $\mathbf{x} = (0, 1, 0, 0)^\top$ and $\mathbf{y} = (0, 0, 1, 0)^\top$.

   (b) Find an orthonormal basis of $\text{span}\{(1, 0, 0, 0)^\top, (0, 0, 1, 0)^\top\}$.

   (c) Let $\mathbb{W} = \text{span}\{(0, 0, 1, 0)^\top\}$. Find a basis of $\mathbb{W}^\perp$.

6. (a) Let a subspace $\mathbb{W}$ of $\mathbb{R}^3$ be spanned by two vectors $(-1,0,1)^\top$ and $(1,1,0)^\top$. Find the projection of $(1,1,1)^\top$.

   (b) Consider a linear model $z = ax + by$. We have three observations of (x,y,z): $(-1,1,1)$, $(0,1,1)$, $(1,0,1)$. Compute the least square estimates of a and b.

   (c) Compare the above two questions.

7. True or False. No need to explain your guesses.

   (a) For an $m \times n$ $(m > n)$ matrix, the number of linearly independent rows equals the number of linearly independent columns.

   (b) For an $n \times n$ matrix A, a map T from the vector space of $n \times n$ matrix onto itself, defined as $T(X) = AX - XA$ is a linear transformation.

   (c) Suppose that a square matrix A is invertible. Then, the following block matrix B is invertible, and the inverse is given as

$$B = \begin{bmatrix} A & \mathbf{0} \\ C & I \end{bmatrix} \quad \text{and} \quad B^{-1} = \begin{bmatrix} A^{-1} & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix}$$

   where $\mathbf{0}$ is a matrix with 0's and I is an identity matrix in appropriate sizes, respectively.

   (d) A set of linearly independent vectors is orthogonal.

   (e) For a square matrix A, $\dim \mathrm{Null}\,(A) = \dim \mathrm{Null}\,(A^\top)$.

Problem Set 2

1. Suppose A is a 3×4 matrix and $\mathbf{b}$ is a vector in $\mathbb{R}^3$.

   (a) Give an example of the matrix A with rank 2. No entry of A is allowed to be zero and PA admits an LU decomposition with $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ whereas A itself does not allow LU decomposition.

   (b) For the matrix A in (a), give two examples of $\mathbf{b}$, one of which allows a solution of $A\mathbf{x} = \mathbf{b}$ and the other one doesn't.

   (c) Find vectors in $\mathrm{Col}\,(A)$ which are closest to the two vectors in (b), respectively.

2. Assume that $\mathbf{u}_1, \ldots, \mathbf{u}_k$ and $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are sets of orthonormal vectors in $\mathbb{R}^n$, respectively. Find the rank of $n \times n$ matrix

$$\sum_{j=1}^{k} j\mathbf{u}_j\mathbf{v}_j^\top .$$

3. Let a vector space $\mathbb{V}$ include all functions in the form of $f(t) = a_1e^{-t} + a_2e^{-2t} + a_3e^{-3t}$ where $a_1, a_2, a_3 \in \mathbb{R}$, that is, $\mathbb{V} = \{a_1e^{-t} + a_2e^{-2t} + a_3e^{-3t} : a_1, a_2, a_3 \in \mathbb{R}\}$. For two functions $f(t)$ and $g(t)$ in $\mathbb{V}$, we define an inner product

$$\langle f, g\rangle = \int_0^\infty f(t)g(t)dt .$$

   (a) Check that $\mathbb{V}$ is a vector space over the scalar $\mathbb{R}$.

   (b) Compute $|f|$, $|g|$, and $|f - g|$ for $f(t) = e^{-t}$ and $g(t) = e^{-2t}$.

   (c) Show that $\mathcal{B}_1 = \{e^{-t}, e^{-2t}, e^{-3t}\}$ is linearly independent.

   (d) Find the matrix representation of the inner product $\langle\cdot,\cdot\rangle$ with respect to the basis $\mathcal{B}_1$.

   (e) For $\mathbb{W} = \text{span}\{e^{-t}, e^{-3t}\}$, compute $\mathbf{P}_{\mathbb{W}}(e^{-2t})$.

   (f) Find an orthonormal basis $\mathcal{B}_2$ for $\mathbb{V}$.

   (g) Find the matrix representation of the inner product $\langle\cdot,\cdot\rangle$ with respect to the basis $\mathcal{B}_2$.

   (h) The density level of some chemical at time t is described by $a_1e^{-t} + a_2e^{-2t} + a_3e^{-3t} + \varepsilon$ for some fixed $a_1, a_2, a_3 \in \mathbb{R}$. At n time points $0 < t_1 < t_2 < \cdots < t_n$, the observed density levels are $y_1, y_2, \ldots, y_n$. For $\varepsilon_i = y_i - \left(a_1e^{-t_i} + a_2e^{-2t_i} + a_3e^{-3t_i}\right)$, characterize $\hat{a}_1, \hat{a}_2, \hat{a}_3$ that minimizes $\sum_{i=1}^n \varepsilon_i^2$.

4. Let $\mathbb{V}$ be a vector space and T be a linear transform representing a projection.

   (a) Show that $\mathbb{W} = \{T(\mathbf{v}) : \mathbf{v} \in \mathbb{V}\}$ be a subspace of $\mathbb{V}$.

   (b) Show that $I - T$ is also a projection where $(I - T)(\mathbf{v}) = \mathbf{v} - T(\mathbf{v})$.

   (c) What is the linear transform $T \circ (I - T)$?

   (d) Describe $\mathbb{W}^\perp$ in terms of T.

5. True or False. No need to explain your guesses.

(a) For an $n \times n$ matrix A, $I - A$ has rank $n - k$ if $\operatorname{rank} A = k$.

(b) For an $n \times n$ projection matrix P, $I - P$ has rank $n - k$ if $\operatorname{rank} P = k$.

(c) For an $n \times n$ matrix A, a map T from the vector space of $n \times n$ matrix onto itself, defined as $T(X) = A^\top X - X^\top A$ is a linear transformation.

(d) Suppose that U and V are $n \times k$ matrices. Then

$$\begin{bmatrix} I_n & \mathbf{0} \\ V^\top & I_k \end{bmatrix} \begin{bmatrix} I_n + UV^\top & U \\ \mathbf{0} & I_k \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -V^\top & I_k \end{bmatrix} = \begin{bmatrix} I_n & U \\ \mathbf{0} & I_k - V^\top U \end{bmatrix}.$$

(e) A set of orthogonal vectors is linearly independent.

13.2 Problems for Chapter 5 ∼ 9

Problem Set 1

1. A 3×3 square matrix has a QR-type decomposition (not an exact QR-decomposition) as $A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1/2 & -1 \\ 1 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$.

   (a) Find the Q and R factors in the QR-decomposition of A.

   (b) Compute the volume of a parallelopiped whose six faces are parallelograms formed by the column vectors of A.

   (c) Compute A^{-1}. You may describe the inverse in a decomposed form.

2. Suppose A is a 3×4 matrix. We computed its singular value decomposition (SVD) using a computer program. Unfortunately, there was a problem with the screen, and we could not recognize some figures of the SVD results. We have U, V, and Σ such that $A = U\Sigma V^\top$ where

$$U = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & * & * \\ -1/\sqrt{2} & * & * \end{bmatrix}, V = \begin{bmatrix} 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 0 & * \\ 1/2 & * & * \\ 1/2 & 0 & * \end{bmatrix}, \Sigma = \begin{bmatrix} 3 & 0 & * \\ 0 & 2 & * \\ 0 & * & 1 \end{bmatrix},$$

   and $*$ means missing figure on the screen.

   (a) Fill out V.

   (b) Find the largest eigenvalue and corresponding eigenvectors of $A^\top A$.

   (c) Find a rank 2 matrix B that minimizes $\|A - B\|_2$.

3. Suppose A is a 3×4 matrix given as $A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 4 & 2 & -2 & 2 \\ -2 & -1 & 1 & -1 \end{bmatrix}$.

   (a) Find a singular value decomposition of A.

   (b) Find the pseudoinverse A^+ of A.

   (c) Find $\|A\|_2$.

4. Assume an $m \times n$ matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ has non-zero columns $\mathbf{a}_i \in \mathbb{R}^m$, $i = 1, \ldots, n$ that are orthogonal to each other: $\mathbf{a}_i^\top \mathbf{a}_j = 0$ for $i \neq j$. Find an SVD for A, in terms of $\mathbf{a}_i$'s. Be as explicit as you can.

5. Assume an $m \times n$ matrix A with $m \geq n$, have singular values $\sigma_1, \ldots, \sigma_n$. Set an $(m+n) \times n$ matrix $\tilde{A} = \begin{bmatrix} A \\ I_n \end{bmatrix}$.

   (a) Find the singular values of $\tilde{A}$.

   (b) Find an SVD of the matrix $\tilde{A}$ in terms of the SVD for A.

6. $A = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ where $\mathbf{v}_i$'s are orthonormal, $\lambda_1 > 0$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

   (a) Compute $\det A$.

   (b) Find all eigenvalues and corresponding eigenvectors of A.

   (c) For A to be positive semi-definite, what conditions on λ_i do we need?

   (d) If A is positive semi-definite, characterize an SVD of A.

   (e) If A is positive semi-definite, characterize the pseudoinverse A^+ of A.

   (f) If A is positive definite, characterize the inverse A^{-1} of A.

7. Let A be an $n \times n$ symmetric positive definite matrix whose spectral decomposition is described as $A = V \Lambda V^\top$ where V is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ for $\lambda_i > 0$. We also define $\Lambda^{1/k} = \mathrm{diag}(\lambda_1^{1/k}, \ldots, \lambda_n^{1/k})$.

   (a) Characterize a symmetric positive definite matrix B_2 satisfying $A = B_2^2$ in terms of V and Λ.

   (b) Characterize a symmetric positive definite matrix B_k satisfying $A = B_k^k$ in terms of V and Λ for every positive integer k.

(c) What would be $\lim_{k\to\infty} B_k$? Validate your answer as reasonably as possible.

8. True or False. No need to explain your guesses.

   (a) For an orthogonal matrix Q, $\det Q = 1$.

   (b) For $n \times n$ symmetric matrices A and B, $\lambda_k(A+B) > \lambda_k(A)$ if B is positive definite.

   (c) Suppose that B and C are square matrices. Then, the following block matrix $A = \begin{bmatrix} B & \mathbf{0} \\ \mathbf{0} & C \end{bmatrix}$ is positive definite if and only if both B and C are positive definite.

   (d) For an $m \times n$ matrix A, $A^+ = (A^\top A)^{-1}A$ if $\text{rank}(A) = m$.

   (e) Every projection matrix has 0 as its determinant.

Problem Set 2

1. Let a matrix A be $m \times n$ and a matrix B be $n \times m$. Consider a block matrix

$$M = \begin{bmatrix} \mathbf{0} & A \\ -B & I_n \end{bmatrix}.$$

   Find $\det M$.

2. Assume that a matrix B has eigenvalues 1, 2, 3, a matrix C has eigenvalues 4, 0, -4, and a matrix D has eigenvalues -1, -2, -3. Set a 6×6 matrix $A = \begin{bmatrix} B & C \\ \mathbf{0} & D \end{bmatrix}$ where B, C, and D are 3×3 matrices. We also define a linear transformation $\ell : \mathbb{R}^6 \to \mathbb{R}^6$ as $\ell(\mathbf{x}) = A\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^6$.

   (a) What are the eigenvalues of the 6×6 matrix A?

   (b) Consider the unit cube Q in $\mathbb{R}^6$, that is, $Q = \{(x_1, \ldots, x_6)^\top : 0 \le x_i \le 1, i = 1, \ldots, 6\} \subset \mathbb{R}^6$. Find the volume of $\ell(Q) = \{\ell(\mathbf{x}) : \mathbf{x} \in Q\}$.

3. Let a 4×4 matrix A is given as

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

(a) Find eigenvalues and eigenvectors of A.

(b) Find singular values and right-/left-singular vectors of A.

(c) Find $\|A\|_2$ and $\|A\|_F$.

(d) Find the pseudoinverse A^+ of A.

(e) Find a 4×4 matrix B of rank 2 that minimizes $\|A - B\|_2$.

4. Suppose A is a 3×4 matrix given as $A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 4 & 2 & -2 & 2 \\ -2 & -1 & 1 & -1 \end{bmatrix}$.

(a) Find eigenvalues and eigenvectors of $A^\top A$.

(b) Find eigenvalues and eigenvectors of $AA^\top$.

(c) Is there a 3×4 matrix B of rank 2 that minimizes $\|A - B\|_2$?

5. Let A be an $m \times n$ matrix. Define its symmetrization $s(A)$ as

$$s(A) = \begin{bmatrix} \mathbf{0} & A \\ A^\top & \mathbf{0} \end{bmatrix},$$

which is an $(m+n) \times (m+n)$ symmetric matrix as the name suggests.

(a) Let $(\sigma, \mathbf{v}, \mathbf{u})$ be a singular triplet of A. Then, σ and $-\sigma$ are eigenvalues of $s(A)$. Find the eigenvectors of $s(A)$ associated with the eigenvalues σ and $-\sigma$, respectively.

(b) Let $(\lambda, \mathbf{w})$ be an eigenpair of $s(A)$ where $\lambda \neq 0$. Then, $-\lambda$ is also an eigenvalue of $s(A)$. Find the eigenvector of $s(A)$ associated with the eigenvalue $-\lambda$.

(c) Let $\lambda < 0$ and $(\lambda, \mathbf{w})$ be an eigenpair of $s(A)$. Find right-/left-singular vectors of A associated with the singular value $-\lambda$.

6. Let A be an $m \times n$ matrix with at least one non-zero entry. What are the projection $\mathbf{P}_{\mathrm{Col}\,(A)}$ onto $\mathrm{Col}\,(A)$ and the pseudoinverse A^+ of A if

(a) A consists of a single column $\mathbf{v}$, that is, $n = 1$;

(b) the columns of A are orthonormal;

(c) the columns of A are linearly independent (that is $\mathrm{rank}\, A = n$);

(d) the columns of A are possibly linearly dependent, and one of its compact SVD is $U\Sigma V^\top$.

7. Let A be an $n \times n$ symmetric positive definite matrix. Denote its eigenvalues as $\lambda_1, \ldots, \lambda_n$. Let B be the square root of A, that is, $A = B^2$.

   (a) Let $A = U^\top U$ be the Cholesky factorization of A. Find $|\det U|$.

   (b) Find the eigenvalues of B.

   (c) Let $B = QR$ be the QR-decomposition of B. Find $|\det R|$.

   (d) Does there exist the square root C of B? If yes, what are the eigenvalues of C?

Bibliography

[1] Ash, R. B. (2009) *Real Variables with Basic Metric Space Topology,* Republished by Dover Publications, USA.

[2] Bisgard, J. (2019) *Analysis and Linear Algebra: The Singular Value Decomposition and Applications,* American Mathematical Society, USA.

[3] Blum, A., Hopcroft, J., and Kannan, R. (2020) *Foundations of Data Science,* Cambridge University Press, UK.

[4] Garcia, S. R., and Horn, R. A. (2017) *A Second Course in Linear Algebra,* Cambridge University Press, UK.

[5] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008) Kernel Methods in Machine Learning. *The Annals of Statistics* 36(3), 1171–1220.

[6] Langville, A. N., and Meyer, C. D. (2006) *Google's Page Rank and Beyond: The Science of Search Engine Rankings,* Princeton University Press, USA.

[7] Strang, G. (2006) *Linear Algebra and Its Applications,* Fourth Edition, Brooks/Cole, Cengage Learning, USA.

[8] Trefethen, L. N., Bau, D. (1997) *Numerical Linear Algebra,* SIAM: Society for Industrial and Applied Mathematics, USA.

Appendix A

Convexity

Many important and interesting mathematical observations are based on a shared set of special structures and properties underlying target mathematical objects. One such example is the convexity of a set or a function, which we discuss further here.

Convexity of a Set

Consider a subset C of a vector space. We say C is convex if C contains an interpolated vector, $\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2$, given a pair of vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, in C and a positive scalar, $0 < \lambda < 1$.

Definition A.1 *A subset C of a vector space is convex if*

$$\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2 \in C \quad \textit{for any } \mathbf{v}_1, \mathbf{v}_2 \in C \textit{ and } 0 < \lambda < 1.$$

This definition applies to a subset of a vector space. It is thus natural to check whether the scalar multiplication and vector addition preserve the convexity.

Fact A.1 *Let C_1 and C_2 be convex sets and $\alpha > 0$. Then, both αC_1 and $C_1 + C_2$ are convex.*
Recall that $\alpha C_1 = \{\alpha \mathbf{v} : \mathbf{v} \in C_1\}$ and $C_1 + C_2 = \{\mathbf{v}_1 + \mathbf{v}_2 : \mathbf{v}_1 \in C_1, \mathbf{v}_2 \in C_2\}$.

Proof: Let $\mathbf{v}_1, \mathbf{v}_2 \in C_1$. Then, for any $0 < \lambda < 1$, $\lambda(\alpha \mathbf{v}_1) + (1-\lambda)(\alpha \mathbf{v}_2) = \alpha\big(\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2\big) \in \alpha C_1$, since $\lambda \mathbf{v}_1 + (1-\lambda)\mathbf{v}_2 \in C_1$ from the convexity of C_1. Let both $\mathbf{v}_1 + \mathbf{v}_2$ and $\mathbf{w}_1 + \mathbf{w}_2$ be in $C_1 + C_2$ where $\mathbf{v}_1, \mathbf{w}_1 \in C_1$ and $\mathbf{v}_2, \mathbf{w}_2 \in C_2$.

For any $0 < \lambda < 1$, $\lambda(\mathbf{v}_1 + \mathbf{v}_2) + (1-\lambda)(\mathbf{w}_1 + \mathbf{w}_2) = \big(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{w}_1\big) + \big(\lambda\mathbf{v}_2 + (1-\lambda)\mathbf{w}_2\big) \in C_1 + C_2$, since $\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{w}_1 \in C_1$ and $\lambda\mathbf{v}_2 + (1-\lambda)\mathbf{w}_2 \in C_2$ from the convexity of C_1 and C_2. ■

Among the set operations, intersection preserves the convexity. You may easily find an example where union does not preserve the convexity.

Fact A.2 *Let C_1 and C_2 be convex sets. Then, $C_1 \cap C_2$ is also convex.*

Proof: For any $\mathbf{v}_1, \mathbf{v}_2 \in C_1 \cap C_2$ and $0 < \lambda < 1$, $\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2 \in C_1$ and $\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2 \in C_2$ at the same time from the convexity of C_1 and C_2. Therefore, the fact holds. ■

Convexity of a Function

A real-valued function f defined on a convex set C in a vector space $\mathbb{V}$ is called a convex function if a set $E = \big\{\big(\mathbf{v}, r\big) : \mathbf{v} \in C, r \geq f(\mathbf{v})\big\} \subset \mathbf{V} \times \mathbb{R}$ is convex. The set E is called an epigraph of f and represents a space above the graph of f. This definition of convex function in terms of convex set can be translated into an equivalent but more practical description as follows:

Definition A.2 *A real-valued function f defined on a convex set C is convex if*

$$f\big(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2\big) \leq \lambda f(\mathbf{v}_1) + (1-\lambda) f(\mathbf{v}_2) \quad \textit{for any } \mathbf{v}_1, \mathbf{v}_2 \in C \textit{ and } 0 < \lambda < 1.$$

As the convexity of a function can be characterized in terms of the convexity of its epigraph, Fact A.1 can be translated to apply to functions, that is, addition and scalar multiplication of functions preserve the convexity.

Fact A.3 *Let f_1 and f_2 be convex functions and $\alpha > 0$. Then, both αf_1 and $f_1 + f_2$ are convex.*

Proof: Let $\mathbf{v}_1, \mathbf{v}_2 \in C_1$ and $0 < \lambda < 1$. $(\alpha f_1)(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) \leq \alpha\big(\lambda f_1(\mathbf{v}_1) + (1-\lambda) f_1(\mathbf{v}_2)\big) = \lambda(\alpha f_1)(\mathbf{v}_1) + (1-\lambda)(\alpha f_1)(\mathbf{v}_2)$. $(f_1 + f_2)(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) = f_1(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) + f_2(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) \leq \lambda f_1(\mathbf{v}_1) + (1-\lambda) f_1(\mathbf{v}_2) + \lambda f_2(\mathbf{v}_1) + (1-\lambda) f_2(\mathbf{v}_2) = \lambda(f_1 + f_2)(\mathbf{v}_1) + (1-\lambda)(f_1 + f_2)(\mathbf{v}_2)$. ■

Similarly, we can translate Fact A.2 to show that the maximum of two functions, of which epigraph corresponds to the intersection of the epigraphs of two original functions, is also convex. This can be stated as follows:

Fact A.4 *Let f_1 and f_2 be convex functions. Then, $g(\mathbf{v}) = \max\{f_1(\mathbf{v}), f_2(\mathbf{v})\}$ is convex.*

Proof: For two vectors $\mathbf{v}_1, \mathbf{v}_2$ and $0 < \lambda < 1$, since $f_1(\mathbf{v}) \le g(\mathbf{v})$ and $f_2(\mathbf{v}) \le g(\mathbf{v})$,

$$\begin{aligned} g(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) &= \max\big\{f_1(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2),\ f_2(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2)\big\} \\ &\le \max\big\{\lambda f_1(\mathbf{v}_1) + (1-\lambda)f_1(\mathbf{v}_2),\ \lambda f_2(\mathbf{v}_1) + (1-\lambda)f_2(\mathbf{v}_2)\big\} \\ &\le \max\big\{\lambda g(\mathbf{v}_1) + (1-\lambda)g(\mathbf{v}_2),\ \lambda g(\mathbf{v}_1) + (1-\lambda)g(\mathbf{v}_2)\big\} \\ &= \lambda g(\mathbf{v}_1) + (1-\lambda)g(\mathbf{v}_2)\,. \end{aligned}$$

■

We are often faced with a composition of convex and linear functions. The resulting function from such a composition turned out to be convex as well.

Fact A.5 *Let f be a convex function on a vector space $\mathbb{V}$ and ℓ be a linear transformation from a vector space $\mathbb{W}$ into $\mathbb{V}$. Then, $f \circ \ell$ is a convex function on $\mathbb{W}$.*

Proof: Let $\mathbf{v}_1, \mathbf{v}_2 \in C_1$ and $0 < \lambda < 1$. $(f \circ \ell)(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2) = f\big(\ell(\lambda\mathbf{v}_1 + (1-\lambda)\mathbf{v}_2)\big) = f\big(\lambda\ell(\mathbf{v}_1) + (1-\lambda)\ell(\mathbf{v}_2)\big) \le \lambda f\big(\ell(\mathbf{v}_1)\big) + (1-\lambda)f\big(\ell(\mathbf{v}_2)\big) = \lambda(f \circ \ell)(\mathbf{v}_1) + (1-\lambda)(f \circ \ell)(\mathbf{v}_2)$.

■

The square function is a basic building block of non-linear convex functions. We can show its convexity following a few steps of arithmetics.

Example A.1 Consider $f(x) = x^2$ on $\mathbb{R}$. Then, for $x_1, x_2 \in \mathbb{R}$ and $0 < \lambda < 1$,

$$\begin{aligned} &\lambda f(x_1) + (1-\lambda)f(x_2) - f(\lambda x_1 + (1-\lambda)x_2) \\ &= \lambda x_1^2 + (1-\lambda)x_2^2 - \big(\lambda x_1 + (1-\lambda)x_2\big)^2 \\ &= \lambda x_1^2 + (1-\lambda)x_2^2 - \big(\lambda^2 x_1^2 + 2\lambda(1-\lambda)x_1x_2 + (1-\lambda)^2 x_2^2\big) \\ &= \big(\lambda x_1^2 - \lambda^2 x_1^2 - \lambda(1-\lambda)x_1x_2\big) + \big((1-\lambda)x_2^2 - \lambda(1-\lambda)x_1x_2 - (1-\lambda)^2 x_2^2\big) \\ &= \lambda x_1\big(x_1 - \lambda x_1 - (1-\lambda)x_2\big) + (1-\lambda)x_2\big(x_2 - \lambda x_1 - (1-\lambda)x_2\big) \\ &= \lambda(1-\lambda)x_1(x_1 - x_2) + \lambda(1-\lambda)x_2(x_2 - x_1) \\ &= \lambda(1-\lambda)(x_1 - x_2)^2 \\ &\ge 0. \end{aligned}$$

This shows that f is convex. Furthermore, $g(\mathbf{x}) = g(x_1, \ldots, x_n) = x_i^2$ is also a convex function. ∎

Convexity of a Quadratic Form

Assume that $f : \mathbb{R}^m \to \mathbb{R}$ is convex. For an $m \times n$ matrix A, if we compose f and a linear transformation $\ell(\mathbf{x}) = A\mathbf{x}$, $f \circ \ell(\mathbf{x}) = f(A\mathbf{x})$ is convex by Fact A.5. In the special case of a quadratic function, we relate the convexity to the positive definiteness of A, as below.

Theorem A.1 *If A is a positive semi-definite matrix, then $\mathbf{x}^\top A\mathbf{x}$ is a convex function in $\mathbf{x}$.*

Proof: Without losing generality, we may assume that A is symmetric. Then, by item 3 of Fact 7.1, we get

$$A = B^\top B$$

for some $m \times n$ matrix. Consider a function $f(\mathbf{y}) = \mathbf{y}^\top \mathbf{y} = \sum_{i=1}^m y_i^2$. We know that each y_i^2 is convex from Example A.1, and the sum of convex functions is also convex by Fact A.3. Hence, f is convex. Then, $f(B\mathbf{x}) = \mathbf{x}^\top B^\top B\mathbf{x} = \mathbf{x}^\top A\mathbf{x}$ is convex by Fact A.5. ■

Appendix B

Permutation and its Matrix Representation

A permutation σ is a bijective function from and onto $\{1, \ldots, n\}$, that is, $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $\sigma(i) \neq \sigma(j)$ if $i \neq j$. You may think that a permutation shuffles the order of $1, \ldots, n$. Therefore, its inverse function σ^{-1} always exists, and the inverse function is a permutation, too. Because $(\sigma^{-1})^{-1} = \sigma$, $\{\sigma : \sigma \text{ is a permutation on } \{1, \ldots, n\}\} = \{\sigma^{-1} : \sigma \text{ is a permutation on } \{1, \ldots, n\}\}$. We can also define a $n \times n$ matrix associated with a permutation σ so that $(i, \sigma(i))$-entryof the matrix is 1 and all other entries are zero in the i-th row. Since $(i, \sigma(i))$ equals $(\sigma^{-1}(j), j)$ if we set $j = \sigma(i)$, each column of the matrix also has only one non-zero element. This matrix is called a permutation matrix.

Let us consider an example of size 4, $\sigma(1) = 3, \sigma(2) = 2, \sigma(3) = 4$, and $\sigma(4) = 1$. Then, its permutation matrix Q is

$$Q = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

You may regard a permutation matrix as a row- or column- shuffled identity matrix. Its inverse permutation is $\sigma^{-1}(1) = 4, \sigma^{-1}(2) = 2, \sigma^{-1}(3) = 1$, and

$\sigma^{-1}(4) = 3$, whose permutation matrix Q' is

$$Q' = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

You can see that $Q' = Q^\top$ from this example. That is, the matrix of inverse permutation is the transpose of the original permutation matrix. Furthermore, it is easy to see $QQ' = I$ and $Q'Q = I$, which implies that Q' is the inverse Q^{-1} of Q. That is, the matrix of inverse permutation is the inverse of the original permutation matrix. Combining these two relations, every permutation matrix is invertible and its transpose is its inverse matrix, that is, every permutation matrix is orthogonal.

This conclusion holds not only for this small example. To generalize these results, it is convenient to borrow a summation representation of rank-one matrices if you are exposed to the rank-one matrix. For a permutation σ, its permutation matrix Q can be compactly expressed as

$$Q = \sum_{i=1}^{n} \mathbf{e}_i \mathbf{e}_{\sigma(i)}^\top \tag{B.1}$$

and

$$Q^\top = \big(\sum_{i=1}^{n} \mathbf{e}_i \mathbf{e}_{\sigma(i)}^\top\big)^\top = \sum_{i=1}^{n} \big(\mathbf{e}_i \mathbf{e}_{\sigma(i)}^\top\big)^\top = \sum_{i=1}^{n} \mathbf{e}_{\sigma(i)} \mathbf{e}_i^\top. \tag{B.2}$$

Then,

$$\begin{aligned} QQ^\top &= \Big(\sum_{i=1}^{n} \mathbf{e}_i \mathbf{e}_{\sigma(i)}^\top\Big)\Big(\sum_{j=1}^{n} \mathbf{e}_{\sigma(j)} \mathbf{e}_j^\top\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{e}_i \mathbf{e}_{\sigma(i)}^\top \mathbf{e}_{\sigma(j)} \mathbf{e}_j^\top \\ &= \sum_{i=1}^{n} \mathbf{e}_i \big(\mathbf{e}_{\sigma(i)}^\top \mathbf{e}_{\sigma(i)}\big) \mathbf{e}_i^\top \\ &= \sum_{i=1}^{n} \mathbf{e}_i \mathbf{e}_i^\top \\ &= I\,. \end{aligned}$$

That is, for a permutation matrix Q,

$$QQ^\top = I\,, \tag{B.3}$$

which also implies

$$Q^{-1} = Q^{\top} . \tag{B.4}$$

Therefore, any permutation matrix is an orthogonal matrix.

Appendix C

Existence of Optimal Solutions

Let $B = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| \leq 1\}$ be a unit sphere in the d-dimensional Euclidean space.

Consider the following optimization problem given $f : B \to \mathbb{R}$:

$$\max_{\mathbf{x} \in B} f(\mathbf{x}).$$

If f is continuous and has an upper bound, we can show that there exists at least one $\mathbf{x}^* \in B$ such that $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ for all $\mathbf{x} \in B$. That is, there exists a vector in B that takes the maximum value of the function f.[1] If we knew already topological statements such as "a continuous image of a compact set is compact" and "a compact set in $\mathbb{R}$ is closed and bounded", from advanced calculus or topology, it is a trivial result. Here, we instead provide a proof based on basic techniques from calculus. First, we start with the following lemma.

Lemma C.1 *If a real number α is an upper bound of a continuous function $f : B \to \mathbb{R}$,[2] then either*

- *there exists $\mathbf{x}^* \in B$ such that $\alpha = f(\mathbf{x}^*)$, or;*

[1] This is not always the case, since for instance there is no element within $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ that takes the maximum value of $f(x) = x$. Although the maximum does not exist, the supremum of f is at $x = 1$.

[2] A real number α is an upper bound of a function $f : X \to \mathbb{R}$ when $f(\mathbf{x}) \leq \alpha$ for all $\mathbf{x} \in X$.

- *for any* $\mathbf{x} \in B$, *either*

  1. *there exists* $\tilde{\mathbf{x}} \in B$ *such that* $f(\mathbf{x}) < f(\tilde{\mathbf{x}}) < \alpha$ *and* $\alpha - f(\tilde{\mathbf{x}}) = \frac{1}{2}\big(\alpha - f(\mathbf{x})\big)$, *or;*
  2. *there exists an upper bound* β *of* f *that satisfies* $\beta - f(\mathbf{x}) = \frac{1}{2}\big(\alpha - f(\mathbf{x})\big)$.

Proof: Let $\mathbf{x}^*$ be a solution of $f(\mathbf{x}) = \alpha$ if it exists. Otherwise, the condition means $f(\mathbf{x}) < \alpha$ for all $\mathbf{x} \in B$. In the latter case, for any $\hat{\mathbf{x}} \in B$, consider the mid-point between α and $f(\hat{\mathbf{x}})$, $\beta = f(\hat{\mathbf{x}}) + \frac{1}{2}\big(\alpha - f(\hat{\mathbf{x}})\big) < \alpha$, and the corresponding equation $f(\mathbf{x}) = \beta$.

1. If a solution exists, let it be $\tilde{\mathbf{x}}$. Then, $f(\tilde{\mathbf{x}}) = \beta$, which implies that

$$\alpha - f(\tilde{\mathbf{x}}) = \alpha - \beta = \alpha - f(\hat{\mathbf{x}}) - \frac{1}{2}\big(\alpha - f(\hat{\mathbf{x}})\big) = \frac{1}{2}\big(\alpha - f(\hat{\mathbf{x}})\big).$$

   We also see that $f(\tilde{\mathbf{x}}) > f(\hat{\mathbf{x}})$.

2. If there is no solution of $f(\mathbf{x}) = \beta$, β is another upper bound of f, as f is a continuous function, and

$$\beta - f(\hat{\mathbf{x}}) = \frac{1}{2}\big(\alpha - f(\hat{\mathbf{x}})\big).$$

■

Assume that α is an upper bound of a continuous function f over B. Starting from $\mathbf{x}_0 = \mathbf{0} \in B$ and $\alpha_0 = \alpha$, we can iteratively find $\mathbf{x}_k \in B$ and upper bounds α_k, for $k = 1, 2, \ldots$ that satisfy

$$\alpha_k - f(\mathbf{x}_k) = \frac{1}{2}\big(\alpha_{k-1} - f(\mathbf{x}_{k-1})\big) \tag{C.1}$$

by repeatedly applying Lemma C.1.

If we found $\mathbf{x}^* \in B$ such that $f(\mathbf{x}^*) = \alpha_k$ for some k, $\mathbf{x}^*$ touches an upper bound and is a solution to $f(\mathbf{x}^*) = \max_{\mathbf{x} \in B} f(\mathbf{x})$. We instead assume that there is no solution to $f(\mathbf{x}) = \alpha_k$ for all k among $\mathbf{x} \in B$. We update along the second bullet of Lemma C.1 as follows:

- case 1: keep the upper bound $\alpha_k = \alpha_{k-1}$, and update $\mathbf{x}_k$ as $\tilde{\mathbf{x}}$;
- case 2: keep $\mathbf{x}_k = \mathbf{x}_{k-1}$, and update upper bound α_k as β.

If we do not obtain the solution in finite k, by (C.1), the upper bound sequence satisfies

$$0 \le \alpha_{k-1} - \alpha_k \le \frac{1}{2^k}\big(\alpha - f(\mathbf{0})\big)\,.$$

Then, $\lim_{k\to\infty}\alpha_k = \alpha + \sum_{k=1}^{\infty}(\alpha_k - \alpha_{k-1}) = \ell$ exists by the comparison test in calculus course. For any $\mathbf{x} \in B$, $f(\mathbf{x}) \le \alpha_k$ implies that $f(\mathbf{x}) \le \lim_{k\to\infty}\alpha_k = \ell$, which means that ℓ is also an upper bound. Furthermore, $\lim_{k\to\infty}\big(\alpha_k - f(\mathbf{x}_k)\big) = 0$ holds from (C.1). Combining two limits together, we get $\lim_{k\to\infty} f(\mathbf{x}_k) = \ell$. That is, we get a sequence $\mathbf{x}_k \in B$ whose function values converging to an upper bound ℓ. But, we don't know yet whether $\mathbf{x}_k$ itself converges to a vector in B.

To analyze $\{\mathbf{x}_k : k = 1, 2, \ldots\}$, consider the d-dimensional cube $C_0 = [-1, 1]^d \subset \mathbb{R}^d$ containing the unit sphere B. Bisecting each edge of C_0 generates 2^d smaller cube consisting of edges of length 1, whose union recovers C_0. Since the infinite sequence $\mathbf{x}_k$'s lie in the 2^d small cubes, at least one cube contains infinite terms of the sequence. Say this cube as C_1. We re-index $\mathbf{x}_k$'s in C_1 as a subsequence $\mathbf{x}_i^{(1)}$, $i = 1, 2, \ldots$ while preserving the order of the original sequence. To C_1, we repeat the procedure again to get a cube C_2 of 1/2-edge length containing infinitely many $\mathbf{x}_i^{(1)}$'s. Re-index $\mathbf{x}_i^{(1)}$'s in C_2 as $\mathbf{x}_i^{(2)}$, $i = 1, 2, \ldots$. If we repeat this procedure infinitely many times, we get a sequence of cubes $C_0 \supset C_1 \supset \cdots \supset C_j \supset \cdots$ where each cube C_j contains infinitely many $\mathbf{x}_k$'s under the name of $\mathbf{x}_i^{(j)}$'s such that $\{\mathbf{x}_i^{(j)}\}_{i=1}^{\infty} \subset \{\mathbf{x}_i^{(j-1)}\}_{i=1}^{\infty}$.

We choose diagonal terms $\{\mathbf{x}_i^{(i)}\}_{i=1}^{\infty}$, the i-th term in the i-th subsequence. It is important to notice $\mathbf{x}_i^{(i)} \in C_j$ for all $i \ge j$ from the monotone inclusion of cubes. Take the first coordinate of $\mathbf{x}_i^{(i)} \in \mathbb{R}^d$ and call y_i. Observe that $|y_{j+1} - y_j| \le \left(\frac{1}{2}\right)^{j-1}$ since both $\mathbf{x}_{j+1}^{(j+1)}$ and $\mathbf{x}_j^{(j)}$ belong to C_j whose edges have length of $\left(\frac{1}{2}\right)^{j-1}$. If we apply the comparison test in calculus course, the sequence $y_j = y_1 + \sum_{i=1}^{j-1}(y_{i+1} - y_i)$ converges. Following the same arguments to other coordinates, we can conclude that there exists $\mathbf{x}^* \in \mathbb{R}^d$ such that $|\mathbf{x}_j^{(j)} - \mathbf{x}^*| \to 0$. In addition, from $|\mathbf{x}_j^{(j)}| \le 1$, $|\mathbf{x}^*| \le |\mathbf{x}^* - \mathbf{x}_j^{(j)}| + |\mathbf{x}_j^{(j)}| \le |\mathbf{x}^* - \mathbf{x}_j^{(j)}| + 1$ holds for all j. Taking a limit on j, we narrow the location of limit vectors $\mathbf{x}^*$ within B. On the other hand, since $f(\mathbf{x}_j^{(j)})$ is a subsequence of $f(\mathbf{x}_k)$ converging to ℓ, we observe that $\ell = \lim_{k\to\infty} f(\mathbf{x}_k) = \lim_{j\to\infty} f(\mathbf{x}_j^{(j)})$. Furthermore, $\lim_{j\to\infty} f(\mathbf{x}_j^{(j)}) = f(\mathbf{x}^*)$ since f is a continuous function, which implies $\ell = f(\mathbf{x}^*)$.

Lemma C.2 *If a continuous function $f : B \to \mathbb{R}$ has an upper bound, there*

exists an $\mathbf{x}^* \in B$ *such that*

$$f(\mathbf{x}^*) = \max_{\mathbf{x}\in B} f(\mathbf{x})\,.$$

If we apply Lemma C.2 to a optimization problem to maximize $-f$ where f is a continuous function with a lower bound, then we can find a minimizer of f on B. In addition, it is very useful fact in dealing with optimization formulations that, for any sets B and C,

$$\text{“if } \mathbf{x}^* \in B \subset C,\ f(\mathbf{x}^*) \le \max_{\mathbf{x}\in B} f(\mathbf{x}) \le \max_{\mathbf{x}\in C} f(\mathbf{x}) \text{ holds.”}$$

We borrow this fact at many places without any explicit mentions.

Let us apply Lemma C.2 to a maximization of the norms of linearly transformed vectors.

Lemma C.3 *Let* A *be an* $n \times d$ *matrix. Then there exists a unit vector* $\mathbf{x}^*$ *such that*

$$|A\mathbf{x}^*| = \max_{|\mathbf{x}|\le 1} |A\mathbf{x}| = \max_{|\mathbf{x}|=1} |A\mathbf{x}|\,.$$

Proof: If $A = \mathbf{0}$, the equalities hold trivially. Therefore, let us assume that $A \neq \mathbf{0}$. Then the maximum value is positive.

Consider the first equality. Since the function $|A\mathbf{x}|$ is continuous, we can apply Lemma C.2 once we show that the function $|A\mathbf{x}|$ has an upper bound on B. Let $|\mathbf{x}| \le 1$ and denote i-th row of A as $\mathbf{a}_i^\top$. Then, by the Cauchy-Schwartz inequality,

$$|A\mathbf{x}|^2 = \sum_{i=1}^{n} (\mathbf{a}_i^\top \mathbf{x})^2 \le \sum_{i=1}^{n} |\mathbf{a}_i|^2 |\mathbf{x}|^2 \le \sum_{i=1}^{n} |\mathbf{a}_i|^2 = \|A\|_F^2$$

and we can conclude that the function $|A\mathbf{x}|$ has an upper bound $\|A\|_F$ on $\{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| \le 1\}$.

For the second equality, let $\mathbf{x}^*$ be the vector satisfying the first equality. If $|\mathbf{x}^*| = 0$, $|A\mathbf{x}^*| = 0$ contradicts to $A \neq \mathbf{0}$. If $|\mathbf{x}^*| < 1$, then the unit vector $\mathbf{y} = \frac{1}{|\mathbf{x}^*|}\mathbf{x}^*$ provides a bigger value $|A\mathbf{y}| > |A\mathbf{x}^*|$, which is a contradiction. Hence $|\mathbf{x}^*| = 1$ and this implies the second equality. ■

We extend Lemma C.3 further to an optimization problem with orthonormal conditions.

Lemma C.4 *Let A be an $n \times d$ matrix and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be orthonormal. Then there exists a feasible $\mathbf{x}^*$ satisfying $|\mathbf{x}^*| = 1, \mathbf{v}_1 \perp \mathbf{x}^*, \ldots, \mathbf{v}_k \perp \mathbf{x}^*$ such that*

$$\begin{aligned} |A\mathbf{x}^*| &= \max\{|A\mathbf{x}| : |\mathbf{x}| \le 1, \mathbf{v}_1 \perp \mathbf{x}, \ldots, \mathbf{v}_k \perp \mathbf{x}\} \\ &= \max\{|A\mathbf{x}| : |\mathbf{x}| = 1, \mathbf{v}_1 \perp \mathbf{x}, \ldots, \mathbf{v}_k \perp \mathbf{x}\}. \end{aligned}$$

Proof: Assume $k < d$. Expand the k orthonormal vectors such that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1} \ldots, \mathbf{v}_d\}$ to be an orthonormal basis by Gram-Schmidt procedure. Denote $V = [\,\mathbf{v}_1 \,|\, \ldots, |\, \mathbf{v}_d\,]$, the matrix whose columns are $\mathbf{v}_k$'s. If $\mathbf{y}$ is the coordinate vector with respect to the new basis, $\mathbf{x} = V\mathbf{y}$ holds. Note that $\mathbf{v}_i \perp \mathbf{x}$ is equivalent to $y_i = 0$ since $\mathbf{v}_i^\top \mathbf{x} = \mathbf{v}_i^\top V\mathbf{y} = y_i$. Furthermore, $|\mathbf{x}| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\mathbf{y}^\top V^\top V\mathbf{y}} = \sqrt{\mathbf{y}^\top \mathbf{y}} = |\mathbf{y}|$. Set $\tilde{A}$ to be the last $(d-k)$ columns of AV and $\tilde{\mathbf{y}} = (y_{k+1}, \ldots, y_d)^\top$. Then, the following equalities

$$\begin{aligned} \{|A\mathbf{x}| : |\mathbf{x}| \le 1, \mathbf{v}_1 \perp \mathbf{x}, \ldots, \mathbf{v}_k \perp \mathbf{x}\} &= \{|AV\mathbf{y}| : |\mathbf{y}| \le 1, y_1 = 0, \ldots, y_k = 0\} \\ &= \{|\tilde{A}\tilde{\mathbf{y}}| : |\tilde{\mathbf{y}}| \le 1\} \end{aligned}$$

imply

$$\max\{|A\mathbf{x}| : |\mathbf{x}| \le 1, \mathbf{v}_1 \perp \mathbf{x}, \ldots, \mathbf{v}_k \perp \mathbf{x}\} = \max\{|\tilde{A}\tilde{\mathbf{y}}| : |\tilde{\mathbf{y}}| \le 1\}.$$

By applying Lemma C.3 to $\max\{|\tilde{A}\tilde{\mathbf{y}}| : |\tilde{\mathbf{y}}| \le 1\}$, we get a unit vector $\tilde{\mathbf{y}}^*$ and $\tilde{\mathbf{y}}^{**} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{y}}^* \end{bmatrix} \in \mathbb{R}^d$ satisfying

$$\max\{|\tilde{A}\tilde{\mathbf{y}}| : |\tilde{\mathbf{y}}| \le 1\} = |\tilde{A}\tilde{\mathbf{y}}^*| = |AV\tilde{\mathbf{y}}^{**}| = |A\mathbf{x}^*|$$

where $\mathbf{x}^* = V\tilde{\mathbf{y}}^*$ with $|\mathbf{x}^*| = |\tilde{\mathbf{y}}^*| = 1$ and $\mathbf{v}_1 \perp \mathbf{x}^*, \ldots, \mathbf{v}_k \perp \mathbf{x}^*$. This proves the first equality. Since $|\mathbf{x}^*| = 1$, the second equality holds. ■

Appendix D

Covariance Matrices

A vector whose elements are random variables is called a random vector. That is, a d-dimensional random vector is denoted by $\mathbf{X} = (X_1, \ldots, X_d)^\top$ where each X_i is a random variable. The expectation of a random vector is computed element-wise as

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}, \; \mathbb{E}\big[\mathbf{X}\big] = \mathbb{E}\begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix} = \big(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]\big)^\top \in \mathbb{R}^d .$$

Similarly, a matrix consisting of random variables is called a random matrix. The expectation of a random matrix is defined as a matrix of the same size whose random variables are replaced by their expectations.

A variance as well as a mean are basic statistics of a random variable. To measure the random deviation of a random vector from the mean vector, we extend the definition of the variance of a random variable to the covariance matrix of a random vector. The covariance matrix consists of covariances of all possible combinations of entries, which are random variables themselves, of a random vector.

We start by defining the covariance between two random variables X and Y as

$$\mathbb{C}\mathsf{ov}(X, Y) = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big].$$

Under this definition, $\mathbb{C}\mathsf{ov}(X, X) = \mathbb{V}\mathsf{ar}(X)$. The (i, j)-th entry of the covariance matrix of a random vector $\mathbf{X}$ is then the covariance $\mathbb{C}\mathsf{ov}(X_i, X_j)$. We can conveniently represent a covariance matrix as the expectation of the following

$d \times d$ random matrix:

$$(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top = \Big((X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j]) \Big).$$

Because $\big((\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^\top\big)^\top = (\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^\top = \big(\mathbf{X}-\mathbb{E}[\mathbf{X}]\big)\big(\mathbf{X}^\top - \mathbb{E}\big[\mathbf{X}^\top\big]\big)$, this matrix is of rank-one and symmetric.

The symmetry of this matrix is preserved under the expectation. On the other hand, the rank of the covariance matrix generally increases, as the covariance matrix can be thought of as the average of many statistically independent rank-one matrices.

In short, we can write a covariance matrix as[1]

$$\begin{aligned} \Sigma &= \mathbb{C}\mathsf{ov}(\mathbf{X}, \mathbf{X}) = \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\big] \\ &= \Big(\mathbb{E}\big[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\big]\Big). \end{aligned}$$

Due to the linearity of expectation,

$$\Sigma = \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top = \Big(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\Big).$$

this expression further simplifies to

$$\Sigma = \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big],$$

if $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. It is thus convenient to shift a random vector by its mean vector as

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbb{E}[\mathbf{X}] = \big(X_1 - \mathbb{E}[X_1], \ldots, X_d - \mathbb{E}[X_d]\big)^\top,$$

and work with $\tilde{\mathbf{X}}$ instead, since $\mathbb{C}\mathsf{ov}(\mathbf{X}, \mathbf{X}) = \mathbb{C}\mathsf{ov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}})$.

To get familiar with covariance matrix, it is helpful to write down the entries of a covariance matrix as the expectation of a random matrix.

$$\begin{aligned} \Sigma &= \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\big] = \mathbb{E}\big[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top\big] \\ &= \mathbb{E}\begin{bmatrix} \tilde{X}_1^2 & \tilde{X}_1\tilde{X}_2 & \cdots & \tilde{X}_1\tilde{X}_d \\ \tilde{X}_2\tilde{X}_1 & \tilde{X}_2^2 & \cdots & \tilde{X}_2\tilde{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_d\tilde{X}_1 & \tilde{X}_d\tilde{X}_2 & \cdots & \tilde{X}_d^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}\big[\tilde{X}_1^2\big] & \mathbb{E}\big[\tilde{X}_1\tilde{X}_2\big] & \cdots & \mathbb{E}\big[\tilde{X}_1\tilde{X}_d\big] \\ \mathbb{E}\big[\tilde{X}_2\tilde{X}_1\big] & \mathbb{E}\big[\tilde{X}_2^2\big] & \cdots & \mathbb{E}\big[\tilde{X}_2\tilde{X}_d\big] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\big[\tilde{X}_d\tilde{X}_1\big] & \mathbb{E}\big[\tilde{X}_d\tilde{X}_2\big] & \cdots & \mathbb{E}\big[\tilde{X}_d^2\big] \end{bmatrix} \end{aligned}$$

[1] As you may have guessed already, we can define a covariance matrix between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ as$\mathbb{C}\mathsf{ov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top\big]$.

$$= \begin{bmatrix} \mathbb{V}\mathsf{ar}(X_1) & \mathbb{C}\mathsf{ov}(X_1, X_2) & \cdots & \mathbb{C}\mathsf{ov}(X_1, X_d) \\ \mathbb{C}\mathsf{ov}(X_2, X_1) & \mathbb{V}\mathsf{ar}(X_2) & \cdots & \mathbb{C}\mathsf{ov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathsf{ov}(X_d, X_1) & \mathbb{C}\mathsf{ov}(X_d, X_2) & \cdots & \mathbb{V}\mathsf{ar}(X_d) \end{bmatrix}.$$

D.1 Positive Definiteness of Covariance Matrices

For a d-dimensional random vector $\mathbf{X}$ and an arbitrary $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v}^\top \mathbf{X}$ is a random variable. Note that $\mathbf{v}^\top \mathbf{X} = \mathbf{X}^\top \mathbf{v}$. Its square has alternative expressions of

$$(\mathbf{v}^\top \mathbf{X})^2 = (\mathbf{v}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{v}) = \mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{v}.$$

Using this, the quadratic form induced from Σ satisfies

$$\begin{aligned} \mathbf{v}^\top \Sigma \mathbf{v} &= \mathbf{v}^\top \big(\mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top \big) \mathbf{v} \\ &= \mathbf{v}^\top \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] \mathbf{v} - \mathbf{v}^\top \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top \mathbf{v} \\ &= \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{v}\big] - \mathbb{E}\big[\mathbf{v}^\top \mathbf{X}\big]\mathbb{E}\big[\mathbf{X}^\top \mathbf{v}\big] \\ &= \mathbb{E}\big[(\mathbf{v}^\top \mathbf{X})^2\big] - \big(\mathbb{E}[\mathbf{v}^\top \mathbf{X}]\big)^2 \\ &= \mathbb{V}\mathsf{ar}\big((\mathbf{v}^\top \mathbf{X})^2\big) \geq 0, \end{aligned}$$

which shows that Σ is positive semi-definite. If Σ is not positive definite, there exists some $\mathbf{v}$ such that $\mathbb{E}[(\mathbf{v}^\top \mathbf{X})^2] = 0$ and hence $\mathbf{v}^\top \mathbf{X} = 0$ with probability 1. That is, random vector $\mathbf{X}$ is linearly dependent and $\mathbf{X}$ can not admit a d-dimensional probability density function.

D.2 A Useful Quadratic Identity

Consider a block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ {A_{12}}^\top & A_{22} \end{bmatrix}$$

consisting of matrices A_{11}, A_{12}, and A_{22} of sizes $n_1 \times n_1$, $n_1 \times n_2$, and $n_2 \times n_2$, respectively. Assume that A_{11} and A_{22} are symmetric and invertible. Then A is symmetric.

$\mathbf{u}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{u}_2 \in \mathbb{R}^{n_2}$ are also given. For $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$, let us compute the following quadratic difference

$$\mathbf{u}^\top A^{-1} \mathbf{u} - \mathbf{u}_2 A_{22}^{-1} \mathbf{u}_2 \,.$$

Assuming the invertibility of the Schur complement $S_{11} = A_{11} - A_{12}A_{22}^{-1}A_{12}^\top$ as in (2.11), we obtain

$$A^{-1} = \begin{bmatrix} S_{11}^{-1} & -S_{11}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}^\top S_{11}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{12}^\top S_{11}^{-1}A_{12}A_{22}^{-1} \end{bmatrix}.$$

We can develop the following quadratic form using this inverse representation as

$$\begin{aligned} \mathbf{u}^\top A^{-1}\mathbf{u} &= \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}^\top \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^\top & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \\ &= \mathbf{u}_1^\top S_{11}^{-1}\mathbf{u}_1 - 2\mathbf{u}_1^\top S_{11}^{-1}A_{12}A_{22}^{-1}\mathbf{u}_2 + \mathbf{u}_2^\top \big(A_{22}^{-1} + A_{22}^{-1}A_{12}^\top S_{11}^{-1}A_{12}A_{22}^{-1}\big)\mathbf{u}_2 \\ &= \mathbf{u}_1^\top S_{11}^{-1}\mathbf{u}_1 - 2\mathbf{u}_1^\top S_{11}^{-1}A_{12}A_{22}^{-1}\mathbf{u}_2 + \mathbf{u}_2^\top A_{22}^{-1}\mathbf{u}_2 + \mathbf{u}_2^\top A_{22}^{-1}A_{12}^\top S_{11}^{-1}A_{12}A_{22}^{-1}\mathbf{u}_2 . \end{aligned}$$

Then the quadratic difference becomes as simple as

$$\begin{aligned} &\mathbf{u}^\top A^{-1}\mathbf{u} - \mathbf{u}_2^\top A_{22}^{-1}\mathbf{u}_2 \\ &\quad = \mathbf{u}_1^\top S_{11}^{-1}\mathbf{u}_1 - 2\mathbf{u}_1^\top S_{11}^{-1}A_{12}A_{22}^{-1}\mathbf{u}_2 + \mathbf{u}_2^\top A_{22}^{-1}A_{12}^\top S_{11}^{-1}A_{12}A_{22}^{-1}\mathbf{u}_2 \\ &\quad = \big(\mathbf{u}_1 - A_{12}A_{22}^{-1}\mathbf{u}_2\big)^\top S_{11}^{-1}\big(\mathbf{u}_1 - A_{12}A_{22}^{-1}\mathbf{u}_2\big) . \end{aligned}$$

We re-arrange it as a quadratic identity of

$$\mathbf{u}^\top A^{-1}\mathbf{u} = \mathbf{u}_2^\top A_{22}^{-1}\mathbf{u}_2 + \big(\mathbf{u}_1 - A_{12}A_{22}^{-1}\mathbf{u}_2\big)^\top S_{11}^{-1}\big(\mathbf{u}_1 - A_{12}A_{22}^{-1}\mathbf{u}_2\big) , \qquad \text{(D.1)}$$

which is a key to obtain a conditional density for a multivariate Gaussian distribution. (Refer Appendix D.4.)

D.3 Multivariate Gaussian Distribution

Let $\mathbf{X}$ be a random vector. $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ and $\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top]$ are its mean vector and covariance matrix, respectively. Assume that a multivariate probability density function $f(\mathbf{x})$ to describe the likelihood of a random vector $\mathbf{X}$ is given as a function[2] of

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) .$$

For an appropriate function $g(\cdot)$,

$$f(\mathbf{x}) = g\big((\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\big)$$

[2] As introduced in Section 7.5, $\sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}$ is a Mahalanobis distance defined by a positive definite matrix Σ and a center point $\boldsymbol{\mu}$. The Mahalanobis distance can be interpreted as a statistical distance denominated by the standard deviation.

is a density of a so-called elliptical distribution, the family of which includes the multivariate Gaussian distribution and the multivariate t-distribution. If the function g is a decreasing function, the level set, $\{\mathbf{x} \in \mathbb{R}^d : f(x) \geq a\}$ of the density function is characterized by the following ellipsoid

$$\{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \alpha\}$$

for some positive α (Refer Section 7.5). This simple ellipsoidal geometry of level sets provides various ideas in the analysis of high-dimensional data.

For a $d \times d$ positive definite matrix Σ, the following definite integral is well-known:

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x} = \sqrt{(2\pi)^d \det \Sigma}\,.$$

Once normalizing the left hand side with the constant on the right hand side, we obtain a multivariate density function $f_{\boldsymbol{\mu},\Sigma}$ on $\mathbb{R}^d$ given by

$$f_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

which satisfies

$$f_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) > 0 \text{ and } \int_{\mathbb{R}^d} f_{\boldsymbol{\mu},\Sigma}(\mathbf{x}) d\mathbf{x} = 1\,.$$

This distribution is called multivariate Gaussian distribution. It can be shown that the mean vector and covariance matrix of $f_{\boldsymbol{\mu},\Sigma}$ are $\boldsymbol{\mu}$ and Σ, respectively. Since the Gaussian distribution is fully characterized by the mean and covariance, we simply denote it as $N(\boldsymbol{\mu}, \Sigma)$. It is often necessary to know the conditional density function of multivariate Gaussian in many applications, the derivation of which is non-trivial. It is provided by combining results on block matrices and determinants in Appendix D.4. In addition, Appendix D.6 explains how to generate random samples of the multivariate Gaussian distribution based on the Cholesky decomposition.

D.4 Conditional Multivariate Gaussian Distribution

Let $\mathbf{X}_1$ be an n_1-dimensional random vector, $\mathbf{X}_2$ an n_2-dimensional random vector, and $\boldsymbol{\mu}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^{n_2}$ their mean vectors, respectively. The covariances are given as $\Sigma_{11} = \mathbb{E}\big[(\mathbb{X}_1 - \boldsymbol{\mu}_1)(\mathbb{X}_1 - \boldsymbol{\mu}_1)^\top\big]$, $\Sigma_{12} = \mathbb{E}\big[(\mathbb{X}_1 - \boldsymbol{\mu}_1)(\mathbb{X}_2 -$

$\boldsymbol{\mu}_2)^\top\big]$, and $\Sigma_{22} = \mathbb{E}\big[(\mathbb{X}_2-\boldsymbol{\mu}_2)(\mathbb{X}_2-\boldsymbol{\mu}_2)^\top\big]$. Define the augmented random vector as $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}$ and its mean vector $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}$. The covariance of $\mathbf{X}$ is

$$\Sigma = \mathbb{E}\big[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top\big] = \mathbb{E}\left[\begin{bmatrix}\mathbf{X}_1-\boldsymbol{\mu}_1\\ \mathbf{X}_2-\boldsymbol{\mu}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1-\boldsymbol{\mu}_1\\ \mathbf{X}_2-\boldsymbol{\mu}_2\end{bmatrix}^\top\right] = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

where $\Sigma_{21} = {\Sigma_{12}}^\top$. If we set $\mathbf{u}_1 = \mathbf{X}_1 - \boldsymbol{\mu}_1$, $\mathbf{u}_2 = \mathbf{X}_2 - \boldsymbol{\mu}_2$, and $A_{ij} = \Sigma_{ij}$ and plug them into (D.1) with a notation

$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top,$$

we obtain a decomposition

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}_2-\boldsymbol{\mu}_2)^\top\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\ &+\big(\mathbf{x}_1-\big(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\big)\big)^\top\hat{\Sigma}^{-1}\big(\mathbf{x}_1-\big(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\big)\big).\end{aligned}$$

To find a conditional density of the multivariate Gaussian, we have to simplify $\frac{f_{\boldsymbol{\mu},\Sigma}(\mathbf{x})}{f_{\boldsymbol{\mu}_2,\Sigma_{22}}(\mathbf{x}_2)}$. The above decomposition of quadratic terms helps us simplify the exponents of the conditional density. Formula (8.4) lets us know $\det\Sigma = \det\Sigma_{22}\det\hat{\Sigma}$. Therefore, the conditional multivariate Gaussian density is given by

$$\begin{aligned}\frac{f_{\boldsymbol{\mu},\Sigma}(\mathbf{x})}{f_{\boldsymbol{\mu}_2,\Sigma_{22}}(\mathbf{x}_2)} &= \frac{\sqrt{(2\pi)^{n_2}\det\Sigma_{22}}}{\sqrt{(2\pi)^{n_1+n_2}\det\Sigma}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})+\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)}\\ &= \frac{1}{\sqrt{(2\pi)^{n_1}\det\hat{\Sigma}}}\\ &\quad\times e^{-\frac{1}{2}(\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)))^\top\hat{\Sigma}^{-1}(\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)))}.\end{aligned}$$

As we can notice from the density function, this conditional distribution itself is again a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)$ and covariance $\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top$. It is usual to call $\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ a conditional mean and $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top$ a conditional covariance. The conditional covariance is in a form of the Schur complement.

D.5 Ill-conditioned Sample Covariance Matrices

Inversion of sample covariance matrices is a popular task in Statistics and machine learning. From d-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, unbiased estimator of

their mean vector and covariance matrix are

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} \big(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\big)\big(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\big)^\top,$$

respectively. $\hat{\Sigma}$ is desired to be positive definite. In practice, the inversions often fail in some reasons. One of them is that $\hat{\Sigma}$ is not guaranteed to be positive definite, but semi-definite. Another possibility is that numerical errors in the inversion operations destroy the invertibility of positive definite matrices if they have very small positive eigenvalues. In any cases, we perturb the estimator by a small positive numberϵ, and use $\hat{\Sigma} + \epsilon I$ instead. A direct check using the definition of eigenpair or Fact 7.7 confirm that every eigenvalue of $\hat{\Sigma}$ increases by ϵ through this perturbation.

D.6 Gaussian Sampling using Cholesky Decomposition

It has a long history to generate random samples from statistical distributions. One of the simplest random sampling is to choose an integer uniformly among $\{0, 1, 2, \ldots, p-1\}$ for a prime number p. If we divide the generated numbers by p, we get approximate random samples uniformly on $[0, 1]$. As a random sample on $[0, 1]$, bigger prime p results in higher quality samples on $[0, 1]$. Once we have random samples uniformly on $[0, 1]$, we can generate random samples from a distribution with a cumulative distribution function F by composing the $[0, 1]$-sample with F^{-1}. The exponential distribution is a well-working example of this approach. However, it is difficult for many distributions to find the inverse functions in the form of easy evaluation. The dependence structures for multivariate distributions are another source of difficulty for random sampling. It is not easy to get random samples from the multivariate Gaussian as well.

However, there are many libraries to provide high quality random samples from $N(0, 1)$ or equivalently $N(\mathbf{0}, I)$ because of its popularity. Hence, we start with independent random samples from $N(\mathbf{0}, I)$ to obtain random samples from $N(\boldsymbol{\mu}, \Sigma)$. Assume that the covariance Σ has a Cholesky decomposition of $\Sigma = R^\top R$. For this factor R, we define a random vector by

$$\mathbf{X} = R^\top \mathbf{Z}$$

where the random vector $\mathbf{Z}$ follows $N(\mathbf{0}, I)$, the d-dimensional standard multivariate Gaussian distribution. Then,

- its mean vector is $\mathbb{E}[\mathbf{X}] = R^\top \mathbb{E}[\mathbf{Z}] = R^\top \mathbf{0} = \mathbf{0}$;
- its covariance is

$$\mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big] = \mathbb{E}\big[R^\top \mathbf{Z}\mathbf{Z}^\top R\big] = R^\top \mathbb{E}\big[\mathbf{Z}\mathbf{Z}^\top\big] R = R^\top I R = \Sigma$$

  since $\mathbb{E}\big[\mathbf{Z}\mathbf{Z}^\top\big] = I$.

Hence, the random vector $\mathbf{X}$ shares its mean and covariance with $N(\boldsymbol{\mu}, \Sigma)$, but we do not know what is the distribution of $\mathbf{X}$ yet. A key property for this characterization is that a linear combination of Gaussian random variables is again Gaussian. Therefore, each element of $R^\top \mathbf{Z}$ is also Gaussian, which in turn implies that the random sample $\mathbf{X}$ follows $N(\boldsymbol{\mu}, \Sigma)$.

Appendix E

Complex Numbers and Matrices

A complex number is defined by two real numbers. For this definition, we need a special complex number $\mathtt{i}$ called an imaginary unit satisfying $\mathtt{i}^2 = -1$. For two real numbers a and b, $a + \mathtt{i}b$ is a complex number. So, $\mathtt{i}$ is a complex number corresponding to a pair of 0 and 1. We denote the set of complex numbers as $\mathbb{C}$ and the set of vectors with n complex entries as $\mathbb{C}^n$. In arithmetic of complex numbers, $\mathtt{i}$ may be regarded as a symbol like a real variable with a notational convention: $a + \mathtt{i}(-b) = a - \mathtt{i}b$ and $a + \mathtt{i}(1) = a + \mathtt{i}$. The complex conjugate of $a + \mathtt{i}b$ is $a - \mathtt{i}b$ and also denoted by $\overline{a + \mathtt{i}b}$. When $b = 0$, $a + \mathtt{i}b = a$ is a real number. Basic arithmetic of complex numbers are listed.

- $(a_1 + \mathtt{i}b_1) \pm (a_2 + \mathtt{i}b_2) = (a_1 \pm a_2) + \mathtt{i}(b_1 \pm b_2)$;
- $(a_1 + \mathtt{i}b_1) \times (a_2 + \mathtt{i}b_2) = (a_1a_2 - b_1b_2) + \mathtt{i}(a_1b_2 + b_1a_2)$;
- $|a_1 + \mathtt{i}b_1|^2 = (a_1 + \mathtt{i}b_1) \times (\overline{a_1 + \mathtt{i}b_1}) = (a_1 + \mathtt{i}b_1) \times (a_1 - \mathtt{i}b_1) = a_1^2 + b_1^2 \geq 0$ and $|a_1 + \mathtt{i}b_1| = \sqrt{a_1^2 + b_1^2}$;
- $(a_1+\mathtt{i}b_1)^{-1} = \frac{1}{(a_1+\mathtt{i}b_1)\times(a_1-\mathtt{i}b_1)}(a_1-\mathtt{i}b_1) = \frac{a_1}{(a_1^2+b_1^2)} - \mathtt{i}\frac{b_1}{(a_1^2+b_1^2)}$ if $|a+\mathtt{i}b| \neq 0$.

Note that the conjugation of real numbers does not alter the real numbers. The operational order of arithmetic and conjugation can be reversed. For $z_1, z_2 \in \mathbb{C}$,

- $\overline{(a_1 + \mathtt{i}b_1) \pm (a_2 + \mathtt{i}b_2)} = (\overline{a_1 + \mathtt{i}b_1}) \pm (\overline{a_2 + \mathtt{i}b_2})$, that is, $\overline{z_1 \pm z_2} = \overline{z_1} \pm \overline{z_2}$;

- $\overline{(a_1 + \mathtt{i}b_1) \times (a_2 + \mathtt{i}b_2)} = (\overline{a_1 + \mathtt{i}b_1}) \times (\overline{a_2 + \mathtt{i}b_2})$, that is, $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$.

A complex matrix is a matrix with complex numbers as its entries. For complex numbers, a conjugate transpose corresponds to the transpose of real matrices, defined as $A^{\mathsf{H}} = (\overline{a_{ji}})$ for $A = (a_{ij})$. Similarly to the symmetry $A^\top = A$, A is a Hermitian matrix if $A^{\mathsf{H}} = A$. Notice that $A^{\mathsf{H}} = A^\top$ for real matrix A. On the other hand, $z^{\mathsf{H}} = \overline{z}$ if we regard $z \in \mathbb{C}$ as a 1×1 matrix. For $z \in \mathbb{C}$, $\mathbf{z} \in \mathbb{C}^n$, complex matrices A and B,

- $\overline{\overline{z}} = z$, $\overline{\overline{\mathbf{z}}} = \mathbf{z}$, $\overline{\overline{A}} = A$;
- $\overline{A + B} = \overline{A} + \overline{B}$, $\overline{AB} = \overline{A}\,\overline{B}$;
- $(A + B)^{\mathsf{H}} = A^{\mathsf{H}} + B^{\mathsf{H}}$, $(AB)^{\mathsf{H}} = B^{\mathsf{H}} A^{\mathsf{H}}$

where we assume that the matrices are well-sized such that arithmetic is well-operated.

Few useful facts are ready.

Fact E.1 *A complex number z is a real number if and only if $z = z^{\mathsf{H}}$.*

Proof: For $z = a + \mathtt{i}b$, $\overline{z} = a - \mathtt{i}b$. Then, $b = 0$ is equivalent to $z = \overline{z}$. ■

Fact E.2 *If $A = A^{\mathsf{H}}$, then for all complex vectors $\mathbf{x}$, $\mathbf{x}^{\mathsf{H}} A \mathbf{x}$ is real.*

Proof: Using Fact E.1, $(\mathbf{x}^{\mathsf{H}} A \mathbf{x})^{\mathsf{H}} = \mathbf{x}^{\mathsf{H}} A^{\mathsf{H}} (\mathbf{x}^{\mathsf{H}})^{\mathsf{H}} = \mathbf{x}^{\mathsf{H}} A \mathbf{x}$ implies that $\mathbf{x}^{\mathsf{H}} A \mathbf{x}$ is real. ■

The standard inner product in the vector space $\mathbb{C}^n$ is defined as $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^{\mathsf{H}} \mathbf{v}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$. Most properties of the standard inner product in $\mathbb{R}^n$ including bilinearity are preserved except $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$.

Appendix F

Alternative Proof of Spectral Decomposition Theorem

We provide an alternative proof of Theorem 5.2 without relying on SVD. This proof instead starts from the assumption that a real symmetric matrix has at least one eigenvalue, which is due to the fundamental theorem of algebra. There is a strong degree of resemblance between this proof and that of Theorem 10.5 (the Schur triangularization theorem).

(The Real Spectral Decomposition) *Let A be a real symmetric matrix. Then, A is orthogonally diagonalizable. That is,*

$$A = V\Lambda V^\top = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top ,$$

where V is an orthogonal matrix with orthonormal columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, $|\mathbf{v}_i| = 1$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

Proof: Let $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top$. Because A is symmetric, there is at least one eigenpair $(\lambda, \mathbf{v})$ where λ is real and $|\mathbf{v}| = 1$. For such an eigenpair, there exists a real orthogonal matrix Q that satisfies $Q\mathbf{v} = \mathbf{e}_1$ and $Q^{-1} = Q^\top$.[1] Since

[1] We use the Gram-Schmidt process to begin from $\mathbf{v}$ and successively produce orthonormal vectors, $\mathbf{v}_2, \ldots, \mathbf{v}_n$. Then, $Q^\top = \begin{bmatrix} \mathbf{v} \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n \end{bmatrix}$ satisfies $Q\mathbf{v} = \mathbf{e}_1$ and $QQ^\top = I$.

$A\mathbf{v} = \lambda\mathbf{v}$,

$$QAQ^\top \mathbf{e}_1 = QA\mathbf{v} = \lambda Q\mathbf{v} = \lambda\mathbf{e}_1.$$

The first column of $QAQ^\top$ is $\begin{bmatrix}\lambda\\ \mathbf{0}\end{bmatrix}$, and $QAQ^\top$ is symmetric. We can thus express it as

$$QAQ^\top = \begin{bmatrix}\lambda & \mathbf{0}^\top\\ \mathbf{0} & A_{n-1}\end{bmatrix},$$

where A_{n-1} is an appropriately chosen $(n-1)\times(n-1)$ symmetric matrix.

Assume there exists an $(n-1)\times(n-1)$ real orthogonal matrix Q_{n-1} such that $Q_{n-1}A_{n-1}Q_{n-1}^\top = \Lambda_{n-1}$, for this $(n-1)\times(n-1)$ symmetric matrix A_{n-1}. Let $Q_n = \begin{bmatrix}1 & \mathbf{0}^\top\\ \mathbf{0} & Q_{n-1}\end{bmatrix}$. Then, Q_n is an $n\times n$ real orthogonal matrix, and satisfies $Q_n\begin{bmatrix}\lambda & \mathbf{0}^\top\\ \mathbf{0} & A_{n-1}\end{bmatrix}Q_n^\top = \begin{bmatrix}\lambda & \mathbf{0}^\top\\ \mathbf{0} & \Lambda_{n-1}\end{bmatrix}$. Thus, $Q_nQAQ^\top Q_n^\top$ is diagonal, as

$$Q_nQAQ^\top Q_n^\top = (Q_nQ)A(Q_nQ)^\top = \begin{bmatrix}\lambda & \mathbf{0}^\top\\ \mathbf{0} & \Lambda_{n-1}\end{bmatrix} = \Lambda.$$

Q_nQ is orthogonal because both Q_n and Q are orthogonal. If we let $V = (Q_nQ)^\top$, we now see that $A = V\Lambda V^\top$. Finally, by (3.8) in Corollary 3.1, we get

$$V\Lambda V^\top = \sum_{i=1}^{n}\lambda_i\mathbf{v}_i\mathbf{v}_i^\top.$$

■

Index