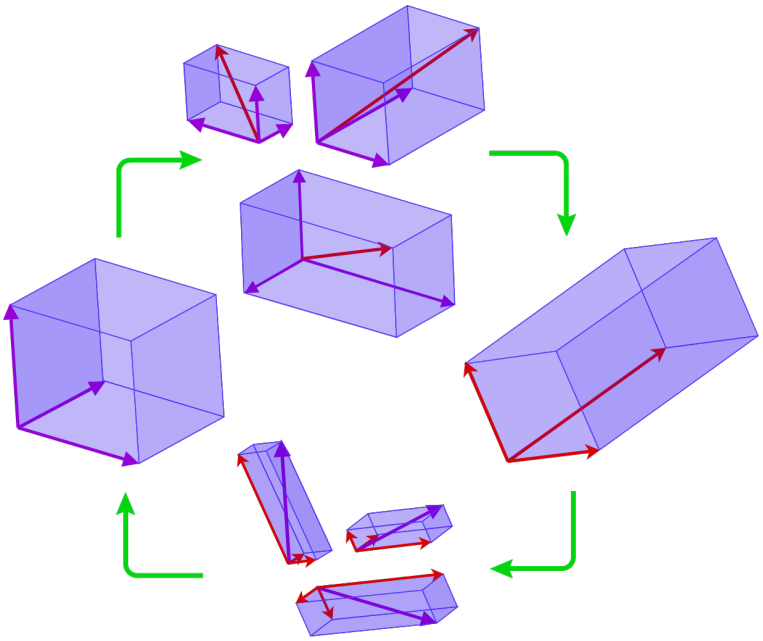
Applied Linear Algebra

MAT 3341

Spring/Summer 2019



Alistair Savage

Department of Mathematics and Statistics

University of Ottawa

This work is licensed under a
Creative Commons Attribution-ShareAlike 4.0 International License

Contents

Preface 4

1 Matrix algebra 5
- 1.1 Conventions and notation . . . 5
- 1.2 Matrix arithmetic . . . 6
- 1.3 Matrices and linear transformations . . . 11
- 1.4 Gaussian elimination . . . 13
- 1.5 Matrix inverses . . . 17
- 1.6 LU factorization . . . 25

2 Matrix norms, sensitivity, and conditioning 36
- 2.1 Motivation . . . 36
- 2.2 Normed vector spaces . . . 37
- 2.3 Matrix norms . . . 40
- 2.4 Conditioning . . . 44

3 Orthogonality 48
- 3.1 Orthogonal complements and projections . . . 48
- 3.2 Diagonalization . . . 53
- 3.3 Hermitian and unitary matrices . . . 56
- 3.4 The spectral theorem . . . 60
- 3.5 Positive definite matrices . . . 65
- 3.6 QR factorization . . . 71
- 3.7 Computing eigenvalues . . . 77

4 Generalized diagonalization 85
- 4.1 Singular value decomposition . . . 85
- 4.2 Fundamental subspaces and principal components . . . 91
- 4.3 Pseudoinverses . . . 94
- 4.4 Jordan canonical form . . . 98
- 4.5 The matrix exponential . . . 102

5 Quadratic forms 106
- 5.1 Definitions . . . 106
- 5.2 Diagonalization of quadratic forms . . . 108

5.3 Rayleigh's principle and the min-max theorem 112

Index **118**

Preface

These are notes for the course *Applied Linear Algebra* (MAT 3341) at the University of Ottawa. This is a third course in linear algebra. The prerequisites are uOttawa courses MAT 1322 and (MAT 2141 or MAT 2342).

In this course we will explore aspects of linear algebra that are of particular use in concrete applications. For example, we will learn how to factor matrices in various ways that aid in solving linear systems. We will also learn how one can effectively compute estimates of eigenvalues when solving for precise ones is impractical. In addition, we will investigate the theory of quadratic forms. The course will involve a mixture of theory and computation. It is important to understand *why* our methods work (the theory) in addition to being able to apply the methods themselves (the computation).

Acknowledgements: I would like to thank Benoit Dionne, Monica Nevins, and Mike Newman for sharing with me their lecture notes for this course.

Alistair Savage

Course website: `https://alistairsavage.ca/mat3341`

Chapter 1

Matrix algebra

We begin this chapter by briefly recalling some matrix algebra that you learned in previous courses. In particular, we review matrix arithmetic (matrix addition, scalar multiplication, the transpose, and matrix multiplication), linear transformations, and gaussian elimination (row reduction). Next we discuss matrix inverses. Although you have seen the concept of a matrix inverse in previous courses, we delve into the topic in further detail. In particular, we will investigate the concept of *one-sided* inverses. We then conclude the chapter with a discussion of LU factorization, which is a very useful technique for solving linear systems.

1.1 Conventions and notation

We let $\mathbb{Z}$ denote the set of integers, and let $\mathbb{N} = \{0, 1, 2, \dots\}$ denote the set of natural numbers.

In this course we will work over the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers, unless otherwise specified. To handle both cases simultaneously, we will use the notation $\mathbb{F}$ to denote the field of scalars. So $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, unless otherwise specified. We call the elements of $\mathbb{F}$ *scalars*. We let $\mathbb{F}^\times = \mathbb{F} \setminus \{0\}$ denote the set of *nonzero* scalars.

We will use uppercase roman letters to denote matrices: A, B, C, M, N, etc. We will use the corresponding lowercase letters to denote the entries of a matrix. Thus, for instance, a_{ij} is the (i, j)-entry of the matrix A. We will sometimes write $A = [a_{ij}]$ to emphasize this. In some cases, we will separate the indices with a comma when there is some chance for confusion, e.g. $a_{i,i+1}$ versus a_{ii+1}.

Recall that a matrix A has *size* $m \times n$ if it has m rows and n columns:

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

We let $M_{m,n}(\mathbb{F})$ denote the set of all $m \times n$ matrices with entries in $\mathbb{F}$. We let $\mathrm{GL}(n, \mathbb{F})$ denote the set of all invertible $n \times n$ matrices with entries in $\mathbb{F}$. (Here 'GL' stands for *general*

linear group.) If $a_1, \ldots, a_n \in \mathbb{F}$, we define

$$\operatorname{diag}(a_1, \ldots, a_n) = \begin{bmatrix} a_1 & 0 & \cdots & \cdots & 0 \\ 0 & a_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & a_n \end{bmatrix}.$$

We will use boldface lowercase letters $\mathbf{a}$, $\mathbf{b}$, $\mathbf{x}$, $\mathbf{y}$, etc. to denote *vectors*. (In class, we will often write vectors as $\vec{a}$, $\vec{b}$, etc. since bold is hard to write on the blackboard.) Most of the time, our vectors will be elements of $\mathbb{F}^n$. (Although, in general, they can be elements of any vector space.) For vectors in $\mathbb{F}^n$, we denote their components with the corresponding non-bold letter with subscripts. We will write vectors $\mathbf{x} \in \mathbb{F}^n$ in column notation:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_1, x_2, \ldots, x_n \in \mathbb{F}.$$

Sometimes, to save space, we will also write this vector as

$$\mathbf{x} = (x_1, x_2, \ldots, x_n).$$

For $1 \leq i \leq n$, we let $\mathbf{e}_i$ denote the i-th *standard basis vector* of $\mathbb{F}^n$. This is the vector

$$\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0),$$

where the 1 is in the i-th position. Then $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$ is a basis for $\mathbb{F}^n$. Indeed, every $\mathbf{x} \in \mathbb{F}^n$ can be written uniquely as the linear combination

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

1.2 Matrix arithmetic

We now quickly review the basic matrix operations. Further detail on the material in this section can be found in [Nic, §§2.1–2.3].

1.2.1 Matrix addition and scalar multiplication

We add matrices of the same size componentwise:

$$A + B = [a_{ij} + b_{ij}].$$

If A and B are of different sizes, then the sum $A + B$ is *not defined*. We define the *negative* of a matrix A by

$$-A = [-a_{ij}]$$

Then the *difference* of matrices of the same size is defined by

$$A - B = A + (-B) = [a_{ij} - b_{ij}].$$

If $k \in \mathbb{F}$ is a scalar, then we define the *scalar multiple*

$$kA = [ka_{ij}].$$

We denote the *zero matrix* by 0. This is matrix with all entries equal to zero. Note that there is some possibility for confusion here since we will use 0 to denote the real (or complex) number zero, as well as the zero matrices of different sizes. The context should make it clear which zero we mean. The context should also make clear what size of zero matrix we are considering. For example, if $A \in M_{m,n}(\mathbb{F})$ and we write $A+0$, then 0 must denote the $m \times n$ zero matrix here.

The following theorem summarizes the important properties of matrix addition and scalar multiplication.

Proposition 1.2.1. *Let A, B, and C be $m \times n$ matrices and let $k, p \in \mathbb{F}$ be scalars. Then we have the following:*

(a) $A + B = B + A$ (*commutativity*)

(b) $A + (B + C) = (A + B) + C$ (*associativity*)

(c) $0 + A = A$ (0 *is an additive identity*)

(d) $A + (-A) = 0$ ($-A$ *is the additive inverse of* A)

(e) $k(A + B) = kA + kB$ (*scalar multiplication is distributive over matrix addition*)

(f) $(k + p)A = kA + pA$ (*scalar multiplication is distributive over scalar addition*)

(g) $(kp)A = k(pA)$

(h) $1A = A$

Remark 1.2.2. Proposition 1.2.1 can be summarized as stating that the set $M_{m,n}(\mathbb{F})$ is a *vector space* over the field $\mathbb{F}$ under the operations of matrix addition and scalar multiplication.

1.2.2 Transpose

The *transpose* of an $m \times n$ matrix A, written A^T, is the $n \times m$ matrix whose rows are the columns of A in the same order. In other words, the (i, j)-entry of A^T is the (j, i)-entry of A. So,

$$\text{if } A = [a_{ij}], \text{ then } A^T = [a_{ji}].$$

We say the matrix A is *symmetric* if $A^T = A$. Note that this implies that all symmetric matrices are *square*, that is, they are of size $n \times n$ for some n.

Example 1.2.3. We have

$$\begin{bmatrix} \pi & i & -1 \\ 5 & 7 & 3/2 \end{bmatrix}^T = \begin{bmatrix} \pi & 5 \\ i & 7 \\ -1 & 3/2 \end{bmatrix}.$$

The matrix

$$\begin{bmatrix} 1 & -5 & 7 \\ -5 & 0 & 8 \\ 7 & 8 & 9 \end{bmatrix}$$

is symmetric.

Proposition 1.2.4. *Let A and B denote matrices of the same size, and let $k \in \mathbb{F}$. Then we have the following:*

(a) $(A^T)^T = A$

(b) $(kA)^T = kA^T$

(c) $(A + B)^T = A^T + B^T$

1.2.3 Matrix-vector multiplication

Suppose A is an $m \times n$ matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{F}^m$:

$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}.$$

For $\mathbf{x} \in \mathbb{F}^n$, we define the *matrix-vector product*

$$A\mathbf{x} := x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n \in \mathbb{F}^m.$$

Example 1.2.5. If

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 3 & 1/2 & \pi \\ -2 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix},$$

then

$$A\mathbf{x} = -1 \begin{bmatrix} 2 \\ 3 \\ -2 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 1/2 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ \pi \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -3 \\ -5/2 + 2\pi \\ 5 \\ 0 \end{bmatrix}.$$

1.2.4 Matrix multiplication

Suppose $A \in M_{m,n}(\mathbb{F})$ and $B \in M_{n,k}(\mathbb{F})$. Let $\mathbf{b}_i$ denote the i-th column of B, so that

$$B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_k \end{bmatrix}.$$

We then define the *matrix product* AB to be the $m \times k$ matrix given by

$$AB := \begin{bmatrix} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_k \end{bmatrix}.$$

Recall that the *dot product* of $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ is defined to be

$$\mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n. \tag{1.1}$$

Then another way to compute the matrix product is as follows: The (i,j)-entry of AB is the dot product of the i-th row of A and the j-column of B. In other words

$$C = AB \iff c_{ij} = \sum_{\ell=1}^{n} a_{i\ell} b_{\ell j} \text{ for all } 1 \leq i \leq m,\ 1 \leq \ell \leq k.$$

Example 1.2.6. If

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

then

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

Recall that the $n \times n$ *identity matrix* is the matrix

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

Even though there is an $n \times n$ identity matrix for each n, the size of I should be clear from the context. For instance, if $A \in M_{m,n}(\mathbb{F})$ and we write AI, then I is the $n \times n$ identity matrix. If, on the other hand, we write IA, then I is the $m \times m$ identity matrix. In case we need to specify the size to avoid confusion, we will write I_n for the $n \times n$ identity matrix.

Proposition 1.2.7 (Properties of matrix multiplication)**.** *Suppose A, B, and C are matrices of sizes such that the indicated matrix products are defined. Furthermore, suppose a is a scalar. Then:*

(a) $IA = A = AI$ (*I is a multiplicative identity*)

(b) $A(BC) = (AB)C$ (*associativity*)

(c) $A(B + C) = AB + AC$ (*distributivity on the left*)

(d) $(B + C)A = BA + CA$ (*distributivity on the right*)

(e) $a(AB) = (aA)B = A(aB)$

(f) $(AB)^T = B^T A^T$

Note that matrix multiplication is *not* commutative in general. First of all, it is possible that the product AB is defined but BA is not. This the case when $A \in M_{m,n}(\mathbb{F})$ and $B \in M_{n,k}(\mathbb{F})$ with $m \neq k$. Now suppose $A \in M_{m,n}(\mathbb{F})$ and $B \in M_{n,m}(\mathbb{F})$. Then AB and BA are both defined, but they are different sizes when $m \neq n$. However, even if $m = n$, so that A and B are both square matrices, we can have $AB \neq BA$. For example, if

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

then

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

Of course, it is possible for $AB = BA$ for some specific matrices (e.g. the zero or identity matrices). In this case, we say that A and B *commute*. But since this does not hold in general, we say that matrix multiplication is not commutative.

1.2.5 Block form

It is often convenient to group the entries of a matrix into submatrices, called *blocks*. For instance, if

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

then we could write

$$A = \begin{bmatrix} B & C \\ D & \end{bmatrix}, \quad \text{where} \quad B = \begin{bmatrix} 1 & 2 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 3 & 5 \\ 0 & 4 & 7 \\ 2 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 3 \\ 5 & -4 \end{bmatrix}.$$

Similarly, if

$$X = \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 3 & -5 \\ -2 & 9 \end{bmatrix},$$

then

$$\begin{bmatrix} 0 & \mathbf{e}_2 & X \\ & & A \end{bmatrix} = \left[\begin{array}{c|c|cc} 0 & 0 & 2 & 1 \\ 0 & 1 & -1 & 0 \\ \hline 0 & 0 & 3 & -5 \\ 0 & 0 & -2 & 9 \end{array}\right],$$

where we have used horizontal and vertical lines to indicate the blocks. Note that we could infer from the sizes of X and A, that the 0 in the block matrix must be the zero vector in $\mathbb{F}^4$ and that $\mathbf{e}_2$ must be the second standard basis vector in $\mathbb{F}^4$.

Provided the sizes of the blocks match up, we can multiply matrices in block form using the usual rules for matrix multiplication. For example,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} AX + BY \\ CX + DY \end{bmatrix}$$

as long as the products AX, BY, CX, and DY are defined. That is, we need the number of columns of A to equal the number of rows of X, etc. (See [Nic, Th. 2.3.4].) Note that, since matrix multiplication is not commutative, the *order of the multiplication of the blocks is very important here.*

We can also compute transposes of matrices in block form. For example,

$$\begin{bmatrix} A & B & C \end{bmatrix}^T = \begin{bmatrix} A^T \\ B^T \\ C^T \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix}^T = \begin{bmatrix} A^T & C^T \\ B^T & D^T \end{bmatrix}.$$

In certain circumstance, we can also compute determinants in block form. Precisely, if A and B are square matrices, then

$$\det \begin{bmatrix} A & X \\ 0 & B \end{bmatrix} = (\det A)(\det B) \quad \text{and} \quad \det \begin{bmatrix} A & 0 \\ Y & B \end{bmatrix} = (\det A)(\det B). \tag{1.2}$$

(See [Nic, Th. 3.1.5].)

Exercises.

Recommended exercises: Exercises in [Nic, §§2.1–2.3].

1.3 Matrices and linear transformations

We briefly recall the connection between matrices and linear transformations. For further details, see [Nic, §2.6].

Recall that a map $T\colon V \to W$, where V and W are $\mathbb{F}$-vector spaces (e.g. $V = \mathbb{F}^n$, $W = \mathbb{F}^m$), is called a *linear transformation* if

(T1) $T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in V$, and

(T2) $T(a\mathbf{x}) = aT(\mathbf{x})$ for all $\mathbf{x} \in V$ and $a \in \mathbb{F}$.

We let $\mathcal{L}(V, W)$ denote the set of all linear transformations from V to W.

Multiplication by $A \in M_{m,n}(\mathbb{F})$ is a linear transformation

$$T_A\colon \mathbb{F}^n \to \mathbb{F}^m, \quad \mathbf{x} \mapsto A\mathbf{x}. \tag{1.3}$$

Conversely, *every* linear map $\mathbb{F}^m \to \mathbb{F}^n$ is given by multiplication by a some matrix. Indeed, suppose

$$T\colon \mathbb{F}^n \to \mathbb{F}^m$$

is a linear transformation. Then define the matrix

$$A = \begin{bmatrix} T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) \end{bmatrix} \in M_{m,n}(\mathbb{F}).$$

Proposition 1.3.1. *With A and T defined as above, we have $T = T_A$.*

Proof. For $\mathbf{x} \in \mathbb{F}^n$, we have

$$\begin{aligned} T\mathbf{x} &= T(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) \\ &= x_1T(\mathbf{e}_1) + \cdots + x_n(T\mathbf{e}_n) && \text{(since } T \text{ is linear)} \\ &= A\mathbf{x}. \end{aligned}$$

□

It follows from the above discussion that we have a one-to-one correspondence

$$M_{m,n}(\mathbb{F}) \to \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m), \quad A \mapsto T_A. \tag{1.4}$$

Now suppose $A \in M_{m,n}(\mathbb{F})$ and $B \in M_{n,k}(\mathbb{F})$. Then, for $\mathbf{x} \in \mathbb{F}^k$, we have

$$(T_A \circ T_B)(\mathbf{x}) = T_A(T_B(\mathbf{x})) = T_A(B\mathbf{x}) = A(B\mathbf{x}) = (AB)\mathbf{x} = T_{AB}\mathbf{x}.$$

Hence

$$T_A \circ T_B = T_{AB}.$$

So, under the bijection (1.4), matrix multiplication corresponds to composition of linear transformations. In fact, this is *why* we define matrix multiplication the way we do.

Recall the following definitions associated to a matrix $A \in M_{m,n}(\mathbb{F})$.

- The *column space* of A, denoted $\operatorname{col} A$, is the span of the columns of A. It is a subspace of $\mathbb{F}^m$ and equal to the image of T_A, denoted $\operatorname{im} T_A$.
- The *row space* of A, denoted $\operatorname{row} A$, is the span of the columns of A. It is a subspace of $\mathbb{F}^n$.
- The *rank* of A is

  $$\operatorname{rank} A = \dim(\operatorname{col} A) = \dim(\operatorname{im} T_A) = \dim(\operatorname{row} A).$$

  (The first equality is a definition, the second follows from the definition of T_A, and the third is a logical consequence that we will recall in Section 1.4.)
- The *null space* of A is

  $$\{\mathbf{x} \in \mathbb{F}^n : A\mathbf{x} = 0\}.$$
- The *nullity* of A, denoted $\operatorname{null} A$, is the dimension of the null space of A.

Recall also that the *kernel* of a linear transformation $T\colon V \to W$ is

$$\ker T = \{\mathbf{v} \in V : T\mathbf{v} = 0\}$$

It follows that $\ker T_A$ is equal to the null space of A, and so

$$\operatorname{null} A = \dim(\ker T_A). \tag{1.5}$$

The important *Rank-Nullity Theorem* (also known as the *Dimension Theoreom*) states that if $T\colon V \to W$ is a linear transformation, then

$$\dim V = \dim(\ker T) + \dim(\operatorname{im} T). \tag{1.6}$$

For a matrix $A \in M_{m,n}(\mathbb{F})$, applying the Rank-Nullity Theorem to $T_A\colon \mathbb{F}^n \to \mathbb{F}^m$ gives

$$n = \operatorname{null} A + \operatorname{rank} A. \tag{1.7}$$

Exercises.

Recommended exercises: Exercises in [Nic, §2.6].

1.4 Gaussian elimination

Recall that you learned in previous courses how to row reduce a matrix. This procedure is called *Gaussian elimination*. We briefly review this technique here. For further details, see [Nic, §§1.1, 1.2, 2.5].

To row reduce a matrix, you used the following operations:

Definition 1.4.1 (Elementary row operations)**.** The following are called *elementary row operations* on a matrix A with entries in $\mathbb{F}$.

- *Type I*: Interchange two rows of A.
- *Type II*: Multiply any row of A by a nonzero element of $\mathbb{F}$.
- *Type III*: Add a multiple of one row of A to another row of A.

Definition 1.4.2 (Elementary matrices)**.** An $n \times n$ *elementary matrix* is a matrix obtained by performing an elementary row operation on the identity matrix I_n. In particular, we define the following elementary matrices:

- For $1 \leq i, j \leq n$, $i \neq j$, we let $P_{i,j}$ be the elementary matrix obtained from I_n by interchanging the i-th and j-th rows.
- For $1 \leq i \leq n$ and $a \in \mathbb{F}^\times$, we let $M_i(a)$ be the elementary matrix obtained from I_n by multiplying the i-th row by a.
- For $i \leq i, j \leq n$, $i \neq j$, and $a \in \mathbb{F}$, we let $E_{i,j}(a)$ be the elementary matrix obtained from I_n by adding a times row i to row j.

The *type* of the elementary matrix is the type of the corresponding row operation performed on I_n.

Example 1.4.3. If $n = 4$, we have

$$P_{1,3} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_4(-2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}, \quad \text{and} \quad E_{2,4}(3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}.$$

Proposition 1.4.4. (a) *Every elementary matrix is invertible and the inverse is an elementary matrix of the same type.*

(b) *Performing an elementary row operation on a matrix A is equivalent to multiplying A on the left by the corresponding elementary matrix.*

Proof. (a) We leave it as an exercise (see Exercise 1.4.1) to check that

$$P_{i,j}P_{i,j} = I, \quad M_i(a)M_i(a^{-1}) = I, \quad E_{i,j}(a)E_{i,j}(-a) = I \tag{1.8}$$

(b) We give the proof for row operations of type III, and leave the proofs for types I and II as exercises. (See Exercise 1.4.1.) Fix $1 \leq i, j \leq n$, with $i \neq j$. Note that

- row k of $E_{i,j}(a)$ is $\mathbf{e}_k$ if $k \neq j$, and
- row j of $E_{i,j}(a)$ is $\mathbf{e}_j + a\mathbf{e}_i$.

Thus, if $k \neq j$, then row k of $E_{i,j}(a)A$ is

$$\mathbf{e}_k A = \text{row } k \text{ of } A,$$

and row j of $E_{i,j}A$ is

$$(\mathbf{e}_j + a\mathbf{e}_i)A = \mathbf{e}_j A + a\mathbf{e}_i A = (\text{row } j \text{ of } A) + a(\text{row } i \text{ of } A).$$

Therefore, $E_{i,j}(a)A$ is the result of adding a times row i to row j. □

Definition 1.4.5 (Row-echelon form)**.** A matrix is in *row-echelon form* (and will be called a *row-echelon matrix*) if:

(a) all nonzero rows are above all zero rows,

(b) the first nonzero entry from the left in each nonzero row is a 1, called the *leading* 1 for that row, and

(c) each leading 1 is strictly to the right of all leading 1s in rows above it.

A row-echelon matrix is in *reduced row-echelon form* (and will be called a *reduced row-echelon matrix*) if, in addition,

(d) each leading 1 is the only nonzero entry in its column.

Remark 1.4.6. Some references do not require the leading entry (i.e. the first nonzero entry from the left in a nonzero row) to be 1 in row-echelon form.

Proposition 1.4.7. *Every matrix $A \in M_{m,n}(\mathbb{F})$ can be transformed to a row-echelon form matrix R by performing elementary row operations. Equivalently, there exist finitely many elementary matrices $E_1, E_2, \ldots, E_k$ such that $R = E_1 E_2 \cdots E_k A$.*

Proof. You saw this in previous courses, and so we will omit the proof here. In fact, there is a precise algorithm, called the *gaussian algorithm*, for bringing a matrix to row-echelon form. See [Nic, Th. 1.2.1] for details. □

Proposition 1.4.8. *A square matrix is invertible if and only if it is a product of elementary matrices.*

Proof. Since the elementary matrices are invertible by Proposition 1.4.4(a), if A is a product of invertible matrices, then A is invertible. Conversely, suppose A is invertible. Then it can be row-reduced to the identity matrix I. Hence, by Proposition 1.4.7, there are elementary matrices $E_1, E_2, \ldots, E_k$ such that $I = E_1 E_2 \cdots E_k A$. Then

$$A = E_k^{-1} \cdots E_2^{-1} E_1^{-1}.$$

Since inverses of elementary matrices are elementary matrices by Proposition 1.4.4(a), we are done. □

Reducing a matrix all the way to reduced row-echelon form is sometimes called *Gauss–Jordan elimination*.

Recall that the *rank* of a matrix A, denoted $\operatorname{rank} A$, is the dimension of the column space of A. Equivalently, $\operatorname{rank} A$ is the number of nonzero rows (which is equal to the number of leading 1s) in any row-echelon matrix U that is row equivalent to A (i.e. that can be obtained from A by row operations). Thus we see that $\operatorname{rank} A$ is also the dimension of the row space of A, as noted earlier.

Recall that every linear system consisting of m linear equations in n variables can be written in matrix form

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

where A is an $m \times n$ matrix, called the *coefficient matrix*,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is the vector of *variables* (or *unknowns*), and $\mathbf{b}$ is the vector of constant terms. We say that

- the linear system is *overdetermined* if there are more equations than unknowns (i.e. if $m > n$),
- the linear system is *underdetermined* if there are more unknowns than equations (i.e. if $m < n$),
- the linear system is *square* if there are the same number of unknowns as equations (i.e. if $m = n$),
- an $m \times n$ matrix is *tall* if $m > n$, and
- an $m \times n$ matrix is *wide* if $m < n$.

It follows immediately that the linear system $A\mathbf{x} = \mathbf{b}$ is

- overdetermined if and only if A is tall,
- underdetermined if and only if A is wide, and
- square if and only if A is square.

Example 1.4.9. As a refresher, let's solve the following underdetermined system of linear equations:

$$\begin{array}{rcrcrcrcrcr} & & & & -4x_3 & + & x_4 & + & 2x_5 & = & 11 \\ 4x_1 & - & 2x_2 & + & 8x_3 & + & x_4 & - & 5x_5 & = & 5 \\ 2x_1 & - & x_2 & + & 2x_3 & + & x_4 & - & 3x_5 & = & 2 \end{array}$$

We write down the augmented matrix and row reduce:

$$\left[\begin{array}{ccccc|c} 0 & 0 & -4 & 1 & 2 & 11 \\ 4 & -2 & 8 & 1 & -5 & 5 \\ 2 & -1 & 2 & 1 & -3 & 2 \end{array}\right] \xrightarrow{R_1 \leftrightarrow R_3} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & -3 & 2 \\ 4 & -2 & 8 & 1 & -5 & 5 \\ 0 & 0 & -4 & 1 & 2 & 11 \end{array}\right]$$

$$\xrightarrow{R_2-2R_1} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & -3 & 2 \\ 0 & 0 & 4 & -1 & 1 & 1 \\ 0 & 0 & -4 & 1 & 2 & 11 \end{array}\right] \xrightarrow{R_3+R_2} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & -3 & 2 \\ 0 & 0 & 4 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 3 & 12 \end{array}\right]$$

One can now easily solve the linear system using a technique called *back substitution*, which we will discuss in Section 1.6.1. However, to further illustrate the process of row reduction, let's continue with gaussian elimination:

$$\left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & -3 & 2 \\ 0 & 0 & 4 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 3 & 12 \end{array}\right] \xrightarrow{\frac{1}{3}R_3} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & -3 & 2 \\ 0 & 0 & 4 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{array}\right]$$

$$\xrightarrow[R_2-R_3]{R_1+3R_3} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & 0 & 14 \\ 0 & 0 & 4 & -1 & 0 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{array}\right] \xrightarrow{\frac{1}{4}R_2} \left[\begin{array}{ccccc|c} 2 & -1 & 2 & 1 & 0 & 14 \\ 0 & 0 & 1 & -1/4 & 0 & -3/4 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{array}\right]$$

$$\xrightarrow{R_2-2R_2} \left[\begin{array}{ccccc|c} 2 & -1 & 0 & 3/2 & 0 & 31/2 \\ 0 & 0 & 1 & -1/4 & 0 & -3/4 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{array}\right] \xrightarrow{\frac{1}{2}R_1} \left[\begin{array}{ccccc|c} 1 & -1/2 & 0 & 3/4 & 0 & 31/4 \\ 0 & 0 & 1 & -1/4 & 0 & -3/4 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{array}\right]$$

The matrix is now in reduced row-echelon form. This reduced matrix corresponds to the equivalent linear system:

$$\begin{array}{rcrcrcl} x_1 & - & \frac{1}{2}x_2 & & + \frac{3}{4}x_4 & = & \frac{31}{4} \\ & & & x_3 & - \frac{1}{4}x_4 & = & \frac{-3}{4} \\ & & & & x_5 & = & 4 \end{array}$$

The *leading variables* are the variables corresponding to leading 1s in the reduced row-echelon matrix: x_1, x_3, and x_5. The *non-leading variables*, or *free variables*, are x_2 and x_4. We let the free variables be parameters:

$$x_2 = s, \quad x_4 = t, \quad s, t \in \mathbb{F}.$$

Then we solve for the leading variables in terms of these parameters giving the general solution in parametric form:

$$\begin{aligned} x_1 &= \frac{31}{4} + \frac{1}{2}s - \frac{3}{4}t \\ x_2 &= s \\ x_3 &= -\frac{3}{4} + \frac{1}{4}t \\ x_4 &= t \\ x_5 &= 4 \end{aligned}$$

Exercises.

1.4.1. Complete the proof of Proposition 1.4.4 by verifying the equalities in (1.8) and verifying part (b) for the case of elementary matrices of types I and II.

Additional recommended exercises: [Nic, §§1.1, 1.2, 2.5].

1.5 Matrix inverses

In earlier courses you learned about invertible matrices and how to find their inverses. We now revisit the topic of matrix inverses in more detail. In particular, we will discuss the more general notions of *one-sided* inverses. In this section we follow the presentation in [BV18, Ch. 11].

1.5.1 Left inverses

Definition 1.5.1 (Left inverse)**.** A matrix X satisfying

$$XA = I$$

is called a *left inverse* of A. If such a left inverse exists, we say that A is *left-invertible*. Note that if A has size $m \times n$, then any left inverse X will have size $n \times m$.

Examples 1.5.2. (a) If $A \in M_{1,1}(\mathbb{F})$, then we can think of A simply as a scalar. In this case a left inverse is equivalent to the inverse of the scalar. Thus, A is left-invertible if and only if it is nonzero, and in this case it has only one left inverse.

(b) Any nonzero vector $\mathbf{a} \in \mathbb{F}^n$ is left-invertible. Indeed, if $a_i \neq 0$ for some $1 \leq i \leq n$, then

$$\frac{1}{a_i}\mathbf{e}_i^T\mathbf{a} = \begin{bmatrix}1\end{bmatrix}.$$

Hence $\frac{1}{a_i}\mathbf{e}_i^T$ is a left inverse of $\mathbf{a}$. For example, if

$$\mathbf{a} = \begin{bmatrix}2\\0\\-1\\1\end{bmatrix},$$

then

$$\begin{bmatrix}1/2 & 0 & 0 & 0\end{bmatrix}, \quad \begin{bmatrix}0 & 0 & -1 & 0\end{bmatrix}, \quad \text{and} \quad \begin{bmatrix}0 & 0 & 0 & 1\end{bmatrix}$$

are all left inverses of $\mathbf{a}$. In fact, $\mathbf{a}$ has infinitely many left inverses. See Exercise 1.5.1.

(c) The matrix

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

has left inverses

$$B = \frac{1}{9}\begin{bmatrix} -7 & -8 & 11 \\ 11 & 10 & -16 \end{bmatrix} \quad \text{and} \quad C = \frac{1}{2}\begin{bmatrix} 0 & -1 & 4 \\ 0 & 1 & -6 \end{bmatrix}.$$

Indeed, one can check directly that $BA = CA = I$.

Proposition 1.5.3. *If A has a left inverse, then the columns of A are linearly independent.*

Proof. Suppose $A \in M_{m,n}(\mathbb{F})$ has a left inverse B, and let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be the columns of A. Suppose that

$$x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = 0$$

for some $x_1, \ldots, x_n \in \mathbb{F}$. Thus, taking $\mathbf{x} = (x_1, \ldots, x_n)$, we have

$$A\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = 0.$$

Thus

$$\mathbf{x} = I\mathbf{x} = BA\mathbf{x} = B0 = 0.$$

This implies that $x_1, \ldots, x_n = 0$. Hence the columns of A are linearly independent. □

We will prove the converse of Proposition 1.5.3 in Proposition 1.5.15.

Corollary 1.5.4. *If A has a left inverse, then A is square or tall.*

Proof. Suppose A is a wide matrix, i.e. A is $m \times n$ with $m < n$. Then it has n columns, each of which is a vector in $\mathbb{F}^m$. Since $m < n$, these columns cannot be linearly independent. Hence A cannot have a left inverse. □

So we see that *only square or tall matrices can be left-invertible*. Of course, not every square or tall matrix is left-invertible (e.g. consider the zero matrix).

Now suppose we want to solve a system of linear equations

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

in the case where A has a left inverse C. If this system has a solution $\mathbf{x}$, then

$$C\mathbf{b} = CA\mathbf{x} = I\mathbf{x} = \mathbf{x}.$$

So $\mathbf{x} = C\mathbf{b}$ is *the* solution of $A\mathbf{x} = \mathbf{b}$. However, we started with the assumption that the system has a solution. If, on the other hand, it has no solution, then $\mathbf{x} = C\mathbf{b}$ cannot satisfy $A\mathbf{x} = \mathbf{b}$.

This gives us a method to check if the linear system $A\mathbf{x} = \mathbf{b}$ has a solution, and to find the solution if it exists, provided A has a left inverse C. We simply compute $AC\mathbf{b}$.

- If $AC\mathbf{b} = \mathbf{b}$, then $\mathbf{x} = C\mathbf{b}$ is the *unique* solution of the linear system $A\mathbf{x} = \mathbf{b}$.
- If $AC\mathbf{b} \neq \mathbf{b}$, then the linear system $A\mathbf{x} = \mathbf{b}$ has no solution.

Keep in mind that this method only works when A has a left inverse. In particular, A must be square or tall. So this method only has a chance of working for square or overdetermined systems.

Example 1.5.5. Consider the matrices of Example 1.5.2(c): The matrix

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

has left inverses

$$B = \frac{1}{9}\begin{bmatrix} -7 & -8 & 11 \\ 11 & 10 & -16 \end{bmatrix} \quad \text{and} \quad C = \frac{1}{2}\begin{bmatrix} 0 & -1 & 4 \\ 0 & 1 & -6 \end{bmatrix}.$$

Suppose we want to solve the over-determined linear system

$$A\mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}. \tag{1.9}$$

We can use *either* left inverse and compute

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

Then we check to see if $(1, -1)$ is a solution:

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

So $(1, -1)$ is the *unique* solution to the linear system (1.9).

Example 1.5.6. Using the same matrix A from Example 1.5.5, consider the over-determined linear system

$$A\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

We compute

$$B\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/9 \\ 1/9 \end{bmatrix}.$$

Thus, *if* the system has a solution, it must be $(1/9, 1/9, 0)$. However, we check that

$$A \begin{bmatrix} 1/9 \\ 1/9 \\ 0 \end{bmatrix} = \begin{bmatrix} 7/9 \\ -10/9 \\ -2/9 \end{bmatrix} \neq \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

Thus, the system has no solution. Of course, we could also compute using the left inverse C as above, or see that

$$B \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/9 \\ 1/9 \end{bmatrix} \neq \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix} = C \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

If the system had a solution, both $(1/9, 1/9, 0)$ and $(1/2, -1/2, 0)$ would be the unique solution, which is clearly not possible.

1.5.2 Right inverses

Definition 1.5.7 (Right inverse)**.** A matrix X satisfying

$$AX = I$$

is called a *right inverse* of A. If such a right inverse exists, we say that A is *right-invertible*. Note that if A has size $m \times n$, then any right inverse X will have size $n \times m$.

Suppose A has a right inverse B. Then

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

and so B^T is a left inverse of A^T. Similarly, if A has a left inverse C, then

$$A^T C^T = (CA)^T = I,$$

and so C^T is a right inverse of A^T. This allows us to translate our results about left inverses to results about right inverses.

Proposition 1.5.8. (a) *The matrix A is left-invertible if and only if A^T is right invertible. Furthermore, if C is a left inverse of A, then C^T is a right inverse of A^T*

(b) *Similarly, A is right-invertible if and only if A^T is left-invertible. Furthermore, if B is a right inverse of A, then B^T is a left inverse of A^T.*

(c) *If a matrix is right-invertible then its rows are linearly independent.*

(d) *If A has a right inverse, then A is square or wide.*

Proof. (a) We proved this above.

(b) We proved this above. Alternatively, it follows from part (a) and the fact that $(A^T)^T = A$.

(c) This follows from Proposition 1.5.3 and part (b) since the rows of A are the columns of A^T.

(d) This follows from Corollary 1.5.4 and part (b) since A is square or wide if and only if A^T is square or tall. □

We can also transpose Examples 1.5.2.

Examples 1.5.9. (a) If $A \in M_{1,1}(\mathbb{F})$, then a right inverse is equivalent to the inverse of the scalar. Thus, A is right-invertible if and only if it is nonzero, and in this case it has only one right inverse.

(b) Any nonzero row matrix $\mathbf{a} = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in M_{1,n}(\mathbb{F})$ is right-invertible. Indeed, if $a_i \neq 0$ for some $1 \leq i \leq n$, then

$$\frac{1}{a_i}\mathbf{a}\mathbf{e}_i = \begin{bmatrix} 1 \end{bmatrix}.$$

Hence $\frac{1}{a_i}\mathbf{e}_i$ is a right inverse of $\mathbf{a}$.

(c) The matrix

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

has right inverses

$$B = \frac{1}{9}\begin{bmatrix} -7 & 11 \\ -8 & 10 \\ 11 & -16 \end{bmatrix} \quad \text{and} \quad C = \frac{1}{2}\begin{bmatrix} 0 & 0 \\ -1 & 1 \\ 4 & -6 \end{bmatrix}.$$

Now suppose we want to solve a linear system

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

in the case where A has a right inverse B. Note that

$$AB\mathbf{b} = I\mathbf{b} = \mathbf{b}.$$

Thus $\mathbf{x} = B\mathbf{b}$ is a solution to his system. Hence, the system has a solution for *any* $\mathbf{b}$. Of course, there can be other solutions; the solution $\mathbf{x} = B\mathbf{b}$ is just one of them.

This gives us a method to solve any linear system $A\mathbf{x} = \mathbf{b}$ in the case that A has a right inverse. Of course, this implies that A is square or wide. So this method only has a chance of working for square or underdetermined systems.

Example 1.5.10. Using the matrix A from Example 1.5.9(c) with right inverses B and C, the linear system

$$A\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

has solutions

$$B\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4/9 \\ 2/9 \\ -5/9 \end{bmatrix} \quad \text{and} \quad C\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

(Of course, there are more. As you learned in previous courses, any linear system with more than one solution has infinitely many solutions.) Indeed, we can find a solution of $A\mathbf{x} = \mathbf{b}$ for any $\mathbf{b}$.

1.5.3 Two-sided inverses

Definition 1.5.11 (Two-sided inverse)**.** A matrix X satisfying

$$AX = I = XA$$

is called a *two-sided inverse*, or simply an *inverse*, of A. If such an inverse exists, we say that A is *invertible* or *nonsingular*. A square matrix that is not invertible is called *singular*.

Proposition 1.5.12. *A matrix A is invertible if and only if it is both left-invertible and right-invertible. In this case, any left inverse of A is equal to any right inverse of A. Hence the inverse of A is unique.*

Proof. Suppose $AX = I$ and $YA = I$. Then

$$X = IX = YAX = YI = Y. \qquad \square$$

If A is invertible, we denote its inverse by A^{-1}.

Corollary 1.5.13. *Invertible matrices are square.*

Proof. If A is invertible, it is both left-invertible and right-invertible by Proposition 1.5.12. Then A must be square by Corollary 1.5.4 and Proposition 1.5.8(d). $\square$

You learned about inverses in previous courses. In particular, you learned:

- If A is an invertible matrix, then the linear system $A\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = A^{-1}\mathbf{b}$.
- If A is invertible, then the inverse can be computed by row-reducing an augmented matrix:
$$\left[\begin{array}{c|c} A & I \end{array}\right] \rightsquigarrow \left[\begin{array}{c|c} I & A^{-1} \end{array}\right].$$

Proposition 1.5.14. *If A is a square matrix, then the following are equivalent:*

(a) *A is invertible.*

(b) *The columns of A are linearly independent.*

(c) *The rows of A are linearly independent.*

(d) *A is left-invertible.*

(e) *A is right-invertible.*

Proof. Let $A \in M_{n,n}(\mathbb{F})$. First suppose that A is left-invertible. Then, by Proposition 1.5.3, the columns of A are linearly independent. Since there are n columns, this implies that they form a basis of $\mathbb{F}^n$. Thus any vector in $\mathbb{F}^n$ is a linear combination of the columns of A. In particular, each of the standard basis vector $\mathbf{e}_i$, $1 \leq i \leq n$, can be expressed as $\mathbf{e}_i = A\mathbf{b}_i$ for some $\mathbf{b}_i \in \mathbb{F}^n$. Then the matrix $B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{bmatrix}$ satisfies

$$AB = \begin{bmatrix} A\mathbf{b}_1 & A\mathbf{b}_2 & \cdots & A\mathbf{b}_n \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{bmatrix} = I.$$

Therefore B is a right inverse of A. So we have shown that

$$\text{(d)} \implies \text{(b)} \implies \text{(e)}.$$

Applying this result to the transpose of A, we see that

$$\text{(e)} \implies \text{(c)} \implies \text{(d)}.$$

Therefore (b) to (e) are all equivalent. Since, by Proposition 1.5.12, (a) is equivalent to ((d) and (e)), the proof is complete. □

1.5.4 The pseudoinverse

We learned in previous courses how to compute two-sided inverses of invertible matrices. But how do we compute left and right inverses in general? We will assume that $\mathbb{F} = \mathbb{R}$ here, but similar methods work over the complex numbers. We just have to replace the transpose by the conjugate transpose (see Definition 3.3.1).

If $A \in M_{m,n}(\mathbb{R})$, then the square matrix

$$A^T A \in M_{n,n}(\mathbb{R})$$

is called the *Gram matrix* of A.

Proposition 1.5.15. (a) *A matrix has linearly independent columns if and only if its Gram matrix $A^T A$ is invertible.*

(b) *A matrix is left-invertible if and only if its columns are linearly independent. Furthermore, if A is left-invertible, then $(A^T A)^{-1} A^T$ is a left inverse of A.*

Proof. (a) First suppose the columns of A are linearly independent. Assume that

$$A^T A\mathbf{x} = 0$$

for some $\mathbf{x} \in \mathbb{R}^n$. Multiplying on the left by $\mathbf{x}^T$ gives

$$0 = \mathbf{x}^T 0 = \mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T (A\mathbf{x}) = (A\mathbf{x}) \cdot (A\mathbf{x}),$$

which implies that $A\mathbf{x} = 0$. (Recall that, for any vector $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v} \cdot \mathbf{v} = 0$ if and only if $\mathbf{v} = 0$.) Because the columns of A are linearly independent, this implies that $\mathbf{x} = 0$. Since the only solution to $A^T A\mathbf{x} = 0$ is $\mathbf{x} = 0$, we conclude that $A^T A$ is invertible.

Now suppose the columns of A are linearly dependent. Thus, there exists a nonzero $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = 0$. Multiplying on the left by A^T gives

$$A^T A\mathbf{x} = 0, \quad \mathbf{x} \neq 0.$$

Thus the Gram matrix $A^T A$ is singular.

(b) We already saw in Proposition 1.5.3 that if A is left-invertible, then its columns are linearly independent. To prove the converse, suppose the columns of A are linearly independent. Then, by (a), the matrix $A^T A$ is invertible. Then we compute

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

Hence $(A^T A)^{-1} A^T$ is a left inverse of A. □

If the columns of A are linearly independent (in particular, A is square or tall), then the particular left inverse $(A^TA)^{-1}A^T$ described in Proposition 1.5.15 is called the *pseudoinverse* of A, the *generalized inverse* of A, or the *Moore–Penrose inverse* of A, and is denoted A^+. Recall that left inverses are *not* unique in general. So this is just *one* left inverse. However, when A is square, we have

$$A^+ = (A^TA)^{-1}A^T = A^{-1}(A^T)^{-1}A^T = A^{-1}I = A^{-1},$$

and so the pseudoinverse reduces to the ordinary inverse (which is unique). Note that this equation does not make sense when A is not square or, more generally, when A is not invertible.

We also have a right analogue of Proposition 1.5.15.

Proposition 1.5.16. (a) *The rows of A are linearly independent if and only if the matrix AA^T is invertible.*

(b) *A matrix is right-invertible if and only if its rows are linearly independent. Furthermore, if A is right-invertible, then $A^T(AA^T)^{-1}$ is a right inverse of A.*

Proof. Essentially, we take the transpose of the statements in Proposition 1.5.15.

(a) We have

$$\begin{aligned} \text{rows of } A \text{ are lin. ind.} &\iff \text{cols of } A^T \text{ are lin. ind.} \\ &\iff (A^T)^TA^T = AA^T \text{ is invertible} \quad \text{(by Proposition 1.5.15(a))}. \end{aligned}$$

(b) We have

$$\begin{aligned} A \text{ is right-invertible} &\iff A^T \text{ is left-invertible} \\ &\iff \text{columns of } A^T \text{ are lin. ind.} \quad \text{(by Proposition 1.5.15(b))} \\ &\iff \text{rows of } A \text{ are lin. ind.} \end{aligned}$$

Furthermore, if A is right-invertible, then, by part (a), AA^T is invertible. Then we have

$$AA^T(AA^T)^{-1} = I,$$

and so $A^T(AA^T)^{-1}$ is a right inverse of A. □

Propositions 1.5.15 and 1.5.16 give us a method to compute right and left inverses, if they exist. Precisely, these results reduce the problem to the computation of two-sided inverses, which you have done in previous courses. Later in the course we will develop other, more efficient, methods for computing left- and right-inverses. (See Sections 3.6 and 4.3.)

Exercises.

1.5.1. Suppose A is a matrix with left inverses B and C. Show that, for any scalars α and β satisfying $\alpha + \beta = 1$, the matrix $\alpha B + \beta C$ is also a left inverse of A. It follows that if a matrix has two different left inverses, then it has a infinite number of left inverses.

1.5.2. Let A be an $m \times n$ matrix. Suppose there exists a nonzero $n \times k$ matrix B such that $AB = 0$. Show that A has no left inverse. Formulate an analogous statement for right inverses.

1.5.3. Let $A \in M_{m,n}(\mathbb{R})$ and let $T_A \colon \mathbb{R}^n \to \mathbb{R}^m$ be the corresponding linear map (see (1.3)).

(a) Prove that A has a left inverse if and only if T_A is injective.

(b) Prove that A has a right inverse if and only if T_A is surjective.

(c) Prove that A has a two-sided inverse if and only if T_A is an isomorphism.

1.5.4. Consider the matrix

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

(a) Is A left-invertible? If so, find a left inverse.

(b) Compute AA^T.

(c) Is A right-invertible? If so, find a right inverse.

Additional recommended exercises from [BV18]: 11.2, 11.3, 11.5, 11.6, 11.7, 11.12, 11.13, 11.17, 11.18, 11.22.

1.6 LU factorization

In this section we discuss a certain factorization of matrices that is very useful in solving linear systems. We follow here the presentation in [Nic, §2.7].

1.6.1 Triangular matrices

If $A = [a_{ij}]$ is an $m \times n$ matrix, the elements $a_{11}, a_{22}, \ldots$ form the *main diagonal* of A, even if A is not square. We say A is *upper triangular* if every entry below and to the left of the main diagonal is zero. For example, the matrices

$$\begin{bmatrix} 5 & 2 & -5 \\ 0 & 0 & 3 \\ 0 & 0 & 8 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & 5 & -1 & 0 & 1 \\ 0 & -1 & 4 & 7 & -1 \\ 0 & 0 & 3 & 7 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 & 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 & -3 \\ 0 & 0 & 2 & 7 & -4 \end{bmatrix}$$

are all upper triangular. In addition

every row-echelon matrix is upper triangular.

Similarly, A is *lower triangular* if every entry above and to the right of the main diagonal is zero. Equivalently, A is lower triangular if and only if A^T is upper triangular. We say a matrix is *triangular* if it is upper or lower triangular.

If the coefficient matrix of a linear system is upper triangular, there is a particularly efficient way to solve the system, known as *back substitution*, where later variables are substituted into earlier equations.

Example 1.6.1. Let's solve the following system:

$$\begin{array}{rcrcrcrcl} 2x_1 & - & x_2 & + & 2x_3 & + & x_4 & - & 3x_5 & = & 2 \\ & & & & 4x_3 & - & x_4 & + & x_5 & = & 1 \\ & & & & & & & & 3x_5 & = & 12 \end{array}$$

Note that the coefficient matrix is

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

which is upper triangular. We obtained this matrix part-way through the row reduction in Example 1.4.9. As in gaussian elimination, we let the free variables (i.e. the non-leading variables) be parameters:

$$x_2 = s, \quad x_4 = t.$$

Then we solve for x_5, x_3, and x_1, in that order. The last equation gives

$$x_5 = \frac{12}{3} = 4.$$

Substitution into the second-to-last equation then gives

$$x_3 = \frac{1}{4}(1 + x_4 - x_5) = -\frac{3}{4} + \frac{1}{4}t.$$

Then, substitution into the first equation gives

$$x_1 = \frac{1}{2}(2 + x_2 - 2x_3 - x_4 + 3x_5) = \frac{1}{2}\left(2 + s + \frac{3}{2} - \frac{1}{2}t - t + 12\right) = \frac{31}{4} + \frac{1}{2}s - \frac{3}{4}t.$$

Note that this is the same solution we obtained in Example 1.4.9. But the method of back substitution involved less work!

Similarly, if the coefficient matrix of a system is lower triangular, it can be solved by *forward substitution*, where earlier variables are substituted into later equations. Back substitution is more numerically efficient than gaussian elimination. (With n equations, where n is large, gaussian elimination requires approximately $n^3/2$ multiplications and divisions, whereas back substitution requires about $n^3/3$.) Thus, if we want to solve a large number of linear systems with the same coefficient matrix, it would be useful to write the coefficient matrix in terms of lower and/or upper triangular matrices.

Suppose we have a linear system $A\mathbf{x} = \mathbf{b}$, where we can factor A as $A = LU$ for some lower triangular matrix L and upper triangular matrix U. Then we can efficiently solve the system $A\mathbf{x} = \mathbf{b}$ as follows:

(a) First solve $L\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$ using forward substitution.

(b) Then solve $U\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$ using back substitution.

Then we have

$$A\mathbf{x} = LU\mathbf{x} = L\mathbf{y} = \mathbf{b},$$

and so we have solved the system. Furthermore, *every* solution can be found using this procedure: if $\mathbf{x}$ is a solution, take $\mathbf{y} = U\mathbf{x}$. This is an efficient way to solve the system that can be implemented on a computer.

The following lemma will be useful in our exploration of this topic.

Lemma 1.6.2. *Suppose A and B are matrices.*

(a) *If A and B are both lower (upper) triangular, then so is AB (assuming it is defined).*

(b) *If A is a square lower (upper) triangular matrix, then A is invertible if and only if every main diagonal entry is nonzero. In this case A^{-1} is also lower (upper) triangular.*

Proof. The proof of this lemma is left as Exercise 1.6.1 □

1.6.2 LU factorization

Suppose A is an $m \times n$ matrix. Then we can use row reduction to transform A to a row-echelon matrix U, which is therefore upper triangular. As discussed in Section 1.4, this reduction can be performed by multiplying on the left by elementary matrices:

$$A \to E_1A \to E_2E_1A \to \cdots \to E_kE_{k-1}\cdots E_2E_1A = U.$$

It follows that

$$A = LU, \quad \text{where} \quad L = (E_kE_{k-1}\cdots E_2E_1)^{-1} = E_1^{-1}E_2^{-1}\cdots E_{k-1}^{-1}E_k^{-1}.$$

As long as we do not require that U be reduced then, except for row interchanges, none of the above row operations involve adding a row to a row *above* it. Therefore, if we can avoid row interchanges, all the E_i are *lower triangular*. In this case, L is lower triangular (and invertible) by Lemma 1.6.2. Thus, we have the following result. We say that a matrix can be *lower reduced* if it can be reduced to row-echelon form without using row interchanges.

Proposition 1.6.3. *If A can be lower reduced to a row-echelon (hence upper triangular) matrix U, then we have*

$$A = LU$$

for some lower triangular, invertible matrix L.

Definition 1.6.4 (LU factorization)**.** A factorization $A = LU$ as in Proposition 1.6.3 is called an *LU factorization* or *LU decomposition* of A.

It is possible that no LU factorization exists, when A cannot be reduced to row-echelon form without using row interchanges. We will discuss in Section 1.6.3 how to handle this situation. However, if an LU factorization exists, then the gaussian algorithm gives U and a procedure for finding L.

Example 1.6.5. Let's find an LU factorization of

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

We first lower reduce A to row-echelon form:

$$A = \begin{bmatrix} 0 & 1 & 2 & -1 & 3 \\ 0 & -2 & -4 & 4 & -2 \\ 0 & -1 & -2 & 4 & 3 \end{bmatrix} \xrightarrow[R_3+R_1]{R_2+2R_1} \begin{bmatrix} 0 & 1 & 2 & -1 & 3 \\ 0 & 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 3 & 6 \end{bmatrix} \xrightarrow[R_3-3R_2]{\frac{1}{2}R_2} \begin{bmatrix} 0 & 1 & 2 & -1 & 3 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = U.$$

We have highlighted the leading column at each step. In each leading column, we divide the top row by the top (pivot) entry to create a 1 in the pivot position. Then we use the leading one to create zeros below that entry. Then we have

$$A = LU, \quad \text{where} \quad L = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 2 & 0 \\ -1 & 3 & 1 \end{bmatrix}.$$

The matrix L is obtained from the identity matrix I_3 by replacing the bottom of the first two columns with the highlighted columns above. Note that $\operatorname{rank} A = 2$, which is the number of highlighted columns.

The method of Example 1.6.5 works in general, provided A can be lower reduced. Note that we did not need to calculate the elementary matrices used in our row operations.

Algorithm 1.6.6 (LU algorithm)**.** *Suppose A is an $m \times n$ matrix of rank r, and that A can be lower reduced to a row-echelon matrix U. Then $A = LU$ where L is a lower triangular, invertible matrix constructed as follows:*

(a) *If $A = 0$, take $L = I_m$ and $U = 0$.*

(b) *If $A \neq 0$, write $A_1 = A$ and let $\mathbf{c}_1$ be the leading column of A_1. Use $\mathbf{c}_1$ and lower reduction to create a leading one at the top of $\mathbf{c}_1$ and zeros below it. When this is completed, let A_2 denote the matrix consisting of rows 2 to m of the new matrix.*

(c) *If $A_2 \neq 0$, let $\mathbf{c}_2$ be the leading column of A_2 and repeat Step (b) on A_2 to create A_3.*

(d) *Continue in this way until U is reached, where all rows below the last leading 1 are zero rows. This will happen after r steps.*

(e) *Create L by replacing the bottoms of the first r columns of I_m with $\mathbf{c}_1, \ldots, \mathbf{c}_r$.*

Proof. If $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r$ are columns of lengths $m, m-1, \ldots, m-r+1$, respectively, let $L^{(m)}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r)$ denote the lower triangular $m \times m$ matrix obtained from I_m by placing

$\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r$ at the bottom of the first r columns:

$$L^{(m)}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r) = \begin{bmatrix} \mathbf{c}_1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ & \mathbf{c}_2 & \ddots & & & & & \vdots \\ & & \ddots & \ddots & & & & \vdots \\ & & & \mathbf{c}_r & \ddots & & & \vdots \\ & & & & 1 & \ddots & & \vdots \\ & & & & 0 & \ddots & \ddots & \vdots \\ & & & & \vdots & \ddots & \ddots & 0 \\ & & & & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

We prove the result by induction on n. The case where $A = 0$ or $n = 1$ is straightforward. Suppose $n > 1$ and that the result holds for $n-1$. Let $\mathbf{c}_1$ be the leading column of A. By the assumption that A can be lower reduced, there exist lower-triangular elementary matrices $E_1, \ldots, E_k$ such that, in block form,

$$(E_k \cdots E_2 E_1)A = \begin{bmatrix} 0 & \mathbf{e}_1 & X_1 \\ & & A_1 \end{bmatrix}, \quad \text{where} \quad (E_k \cdots E_2 E_1)\mathbf{c}_1 = \mathbf{e}_1.$$

(Recall that $\mathbf{e}_1 = (1, 0, \ldots, 0)$ is the standard basis element.) Define

$$G = (E_k \cdots E_2 E_2)^{-1} = E_1^{-1} E_2^{-1} \cdots E_k^{-1}.$$

Then we have $G\mathbf{e}_1 = \mathbf{c}_1$. By Lemma 1.6.2, G is lower triangular. In addition, each E_j, and hence each E_j^{-1}, is the result of either multiplying row 1 of I_m by a nonzero scalar or adding a multiple of row 1 to another row. Thus, in block form,

$$G = \begin{bmatrix} \mathbf{c}_1 & 0 \\ & I_{m-1} \end{bmatrix}$$

By our induction hypothesis, we have an LU factorization $A_1 = L_1 U_1$, where $L_1 = L^{(m-1)}(\mathbf{c}_2, \ldots, \mathbf{c}_r)$ and U_1 is row-echelon. Then block multiplication gives

$$G^{-1}A = \begin{bmatrix} 0 & \mathbf{e}_1 & X_1 \\ & & L_1 U_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & L_1 \end{bmatrix} \begin{bmatrix} 0 & 1 & X_1 \\ 0 & 0 & U_1 \end{bmatrix}.$$

Thus $A = LU$, where

$$U = \begin{bmatrix} 0 & 1 & X_1 \\ 0 & 0 & U_1 \end{bmatrix}$$

is row-echelon and

$$L = \begin{bmatrix} \mathbf{c}_1 & 0 \\ & I_{m-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & L_1 \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 & 0 \\ & L_1 \end{bmatrix} = L^{(m)}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r).$$

This completes the proof of the induction step. □

LU factorization is very important in practice. It often happens that one wants to solve a series of systems

$$A\mathbf{x} = \mathbf{b}_1, \quad A\mathbf{x} = \mathbf{b}_2, \cdots, \quad A\mathbf{x} = \mathbf{b}_k$$

with the same coefficient matrix. It is very efficient to first solve the first system by gaussian elimination, simultaneously creating an LU factorization of A. Then one can use this factorization to solve the remaining systems quickly by forward and back substitution.

Example 1.6.7. Let's find an LU factorization for

$$A = \begin{bmatrix} 3 & 6 & -3 & 0 & 3 \\ -2 & 4 & 6 & 8 & -2 \\ 1 & 0 & -2 & -5 & 0 \\ 1 & 2 & -1 & 6 & 3 \end{bmatrix}.$$

We reduce A to row-echelon form:

$$A = \begin{bmatrix} 3 & 6 & -3 & 0 & 3 \\ -2 & 4 & 6 & 8 & -2 \\ 1 & 0 & -2 & -5 & 0 \\ 1 & 2 & -1 & 6 & 3 \end{bmatrix} \xrightarrow[R_4-R_1]{\substack{\frac{1}{3}R_1 \\ R_2+2R_1 \\ R_3-R_1}} \begin{bmatrix} 1 & 2 & -1 & 0 & 1 \\ 0 & 8 & 4 & 8 & 0 \\ 0 & -2 & -1 & -5 & -1 \\ 0 & 0 & 0 & 6 & 2 \end{bmatrix}$$

$$\xrightarrow[R_3+2R_1]{\frac{1}{8}R_2} \begin{bmatrix} 1 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & -3 & -1 \\ 0 & 0 & 0 & 6 & 2 \end{bmatrix} \xrightarrow[R_4-6R_3]{-\frac{1}{3}R_3} \begin{bmatrix} 1 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = U.$$

Thus we have $A = LU$, with

$$L = \begin{bmatrix} 3 & 0 & 0 & 0 \\ -2 & 8 & 0 & 0 \\ 1 & -2 & -3 & 0 \\ 1 & 0 & 6 & 1 \end{bmatrix}.$$

Let's do one more example, this time where the matrix A is invertible.

Example 1.6.8. Let's find an LU factorization for

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

We reduce A to row-echelon form:

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 3 & 4 \\ 1 & -4 & -3 \end{bmatrix} \xrightarrow[R_3-R_1]{R_2+R_1} \begin{bmatrix} 1 & -1 & 2 \\ 0 & 2 & 6 \\ 0 & -3 & -5 \end{bmatrix} \xrightarrow[R_3+3R_2]{\frac{1}{2}R_2} \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 4 \end{bmatrix} \xrightarrow{\frac{1}{4}R_3} \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix} = U.$$

Then $A = LU$ with

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

1.6.3 PLU factorization

All of our examples in Section 1.6.2 worked because we started with matrices A that could be lower reduced. However, there are matrices that have no LU factorization. Consider, for example, the matrix

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

Suppose this matrix had an LU decomposition $A = LU$. Write

$$L = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.$$

Then we have

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \ell_{11}u_{11} & \ell_{11}u_{12} \\ \ell_{21}u_{11} & \ell_{21}u_{12} + \ell_{22}u_{22} \end{bmatrix}.$$

In particular, $\ell_{11}u_{11} = 0$, which implies that $\ell_{11} = 0$ or $u_{11} = 0$. But this would mean that L is singular or U is singular. In either case, we would have

$$\det(A) = \det(L)\det(U) = 0,$$

which contradicts the fact that $\det(A) = -1$. Therefore, *A has no LU decomposition.* By Algorithm 1.6.6, this means that A cannot be lower reduced. The problem is that we need to use a row interchange to reduce A to row-echelon form.

The following theorem tells us how to handle LU factorization in general.

Theorem 1.6.9. *Suppose an $m \times n$ matrix A is row reduced to a row-echelon matrix U. Let $P_1, P_2, \ldots, P_s$ be the elementary matrices corresponding (in order) to the row interchanges using in this reduction, and let $P = P_s \cdots P_2 P_1$. (If no interchanges are used, take $P = I_m$.) Then PA has an LU factorization.*

Proof. The only thing that can go wrong in the LU algorithm (Algorithm 1.6.6) is that, in step (b), the leading column (i.e. first nonzero column) may have a zero entry at the top. This can be remedied by an elementary row operation that swaps two rows. This corresponds to multiplication by a permutation matrix (see Proposition 1.4.4(b)). Thus, if U is a row echelon form of A, then we can write

$$L_r P_r L_{r-1} P_{r-1} \cdots L_2 P_2 L_1 P_1 A = U, \tag{1.10}$$

where U is a row-echelon matrix, each of the $P_1, \ldots, P_r$ is either an identity matrix or an elementary matrix corresponding to a row interchange, and, for each $1 \leq j \leq r$, $L_j = L^{(m)}(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{c}_j)^{-1}$ for some column $\mathbf{c}_j$ of length $m - j + 1$. (Refer to the proof of Algorithm 1.6.6 for notation.) It is not hard to check that, for each $1 \leq j \leq r$,

$$L_j = L^{(m)}(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{c}'_j)$$

for some column $\mathbf{c}'_j$ of length $m - j + 1$.

Now, each permutation matrix can be "moved past" each lower triangular matrix to the right of it, in the sense that, if $k > j$, then

$$P_k L_j = L'_j P_k,$$

where $L'_j = L^{(m)}(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}, \mathbf{c}''_j)$ for some column $\mathbf{c}''_j$ of length $m-j+1$. See Exercise 1.6.2. Thus, from (1.10) we obtain

$$(L_r L'_{r-1} \cdots L'_2 L'_1)(P_r P_{r-1} \cdots P_2 P_1) A = U,$$

for some lower triangular matrices $L'_1, L'_2, \ldots, L'_{r-1}$. Setting $P = P_r P_{r-1} \cdots P_2 P_1$, this implies that PA has an LU factorization, since $L_r L'_{r-1} \cdots L'_2 L'_1$ is lower triangular and invertible by Lemma 1.6.2. □

Note that Theorem 1.6.9 generalizes Proposition 1.6.3. If A can be lower reduced, then we can take $P = I_m$ in Theorem 1.6.9, which then states that A has an LU factorization.

A matrix that is the product of elementary matrices corresponding to row interchanges is called a *permutation matrix*. (We also consider the identity matrix to be a permutation matrix.) Every permutation matrix P is obtained from the identity matrix by permuting the rows. Then PA is the matrix obtained from A by performing the same permutation on the rows of A. The matrix P is a permutation matrix if and only if it has exactly one 1 in each row and column, and all other entries are zero. The elementary permutation matrices are those matrices obtained from the identity matrix by a single row exchange. Every permutation matrix is a product of elementary ones.

Example 1.6.10. Consider the matrix

$$A = \begin{bmatrix} 0 & 2 & 0 & 4 \\ 1 & 0 & 1 & 3 \\ 1 & 0 & 1 & 14 \\ -2 & -1 & -1 & -10 \end{bmatrix}.$$

Let's find a permutation matrix P such that PA has an LU factorization, and then find the factorization.

We first row reduce A:

$$A \xrightarrow{R_1 \leftrightarrow R_2} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 1 & 0 & 1 & 14 \\ -2 & -1 & -1 & -10 \end{bmatrix} \xrightarrow[R_4+2R_1]{R_3-R_1} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 0 & 11 \\ 0 & -1 & 1 & -4 \end{bmatrix} \xrightarrow{R_4+\frac{1}{2}R_2} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 0 & 11 \\ 0 & 0 & 1 & -2 \end{bmatrix}$$
$$\xrightarrow{R_3 \leftrightarrow R_4} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 11 \end{bmatrix}.$$

We used two row interchanges: first $R_1 \leftrightarrow R_2$ and then $R_3 \leftrightarrow R_4$. Thus, as in Theorem 1.6.9, we take

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

We now apply the LU algorithm to PA:

$$PA = \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ -2 & -1 & -1 & -10 \\ 1 & 0 & 1 & 14 \end{bmatrix} \xrightarrow[R_4-R_1]{R_3+2R_1} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 2 & 0 & 4 \\ 0 & -1 & 1 & -4 \\ 0 & 0 & 0 & 11 \end{bmatrix} \xrightarrow[R_3+R_2]{\frac{1}{2}R_2} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 11 \end{bmatrix}$$
$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 11 \end{bmatrix} \xrightarrow{\frac{1}{11}R_3} \begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = U.$$

Hence $PA = LU$, where

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ -2 & -1 & 1 & 0 \\ 1 & 0 & 0 & 11 \end{bmatrix}.$$

Theorem 1.6.9 is an important factorization theorem that applies to *any* matrix. If A is any matrix, this theorem asserts that there exists a permutation matrix P and an LU factorization $PA = LU$. Furthermore, it tells us how to find P, and we then know how to find L and U.

Note that $P_i = P_i^{-1}$ for each i (since any elementary permutation matrix is its own inverse). Thus, the matrix A can be factored as

$$A = P^{-1}LU,$$

where P^{-1} is a permutation matrix, L is lower triangular and invertible, and U is a row-echelon matrix. This is called a *PLU factorization* or a *$PA = LU$ factorization* of A.

1.6.4 Uniqueness

Theorem 1.6.9 is an *existence* theorem. It tells us that a PLU factorization always exists. However, it leaves open the question of *uniqueness*. In general, LU factorizations (and hence PLU factorizations) are *not* unique. For example,

$$\begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 4 & -5 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -4 & 5 \\ 4 & 16 & -20 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 4 & 8 \end{bmatrix} \begin{bmatrix} 1 & 4 & -5 \\ 0 & 0 & 0 \end{bmatrix}.$$

(In fact, one can put *any* value in the $(2,2)$-position of the 2×2 matrix and obtain the same result.) The key to this non-uniqueness is the zero row in the row-echelon matrix. Note that, if A is $m \times n$, then the matrix U has no zero row if and only if A has rank m.

Theorem 1.6.11. *Suppose A is an $m \times n$ matrix with LU factorization $A = LU$. If A has rank m (that is, U has no zero row), then L and U are* uniquely *determined by A.*

Proof. Suppose $A = MV$ is another LU factorization of A. Thus, M is lower triangular and invertible, and V is row-echelon. Thus we have

$$LU = MV, \tag{1.11}$$

and we wish to show that $L = M$ and $U = V$.

We have

$$V = M^{-1}LU = NU, \quad \text{where} \quad N = M^{-1}L.$$

Note that N is lower triangular and invertible by Lemma 1.6.2. It suffices to show that $N = I$. Suppose N is $m \times m$. We prove the result by induction on m.

First note that the first column of V is N times the first column of U. Since N is invertible, this implies that the first column of V is zero if and only if the first column of U is zero. Hence, by deleting zero columns if necessary, we can assume that the $(1,1)$-entry is 1 in both U and V.

If $m = 1$, then, since U is row-echelon, we have

$$LU = \begin{bmatrix}\ell_{11}\end{bmatrix}\begin{bmatrix}1 & u_{12} & \cdots & u_{1n}\end{bmatrix} = \begin{bmatrix}\ell_{11} & \ell_{11}u_{12} & \cdots & \ell_{11}u_{1n}\end{bmatrix} = A = \begin{bmatrix}a_{11} & \cdots & a_{1n}\end{bmatrix}.$$

Thus $\ell_{11} = a_{11}$. Similarly, $m_{11} = a_{11}$. So $L = \begin{bmatrix}a_{11}\end{bmatrix} = M$, as desired.

Now suppose $m > 1$ and that the result holds for N of size $(m-1) \times (m-1)$. As before, we can delete any zero columns. So we can write, in block form,

$$N = \begin{bmatrix} a & 0 \\ X & N_1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & Y \\ 0 & U_1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & Z \\ 0 & V_1 \end{bmatrix}.$$

Then

$$NU = V \implies \begin{bmatrix} a & aY \\ X & XY + N_1U_1 \end{bmatrix} = \begin{bmatrix} 1 & Z \\ 0 & V_1 \end{bmatrix}.$$

This implies that

$$a = 1, \quad Y = Z, \quad X = 0, \quad \text{and} \quad N_1U_1 = V_1.$$

By the induction hypothesis, the equality $N_1U_1 = V_1$ implies $N_1 = I$. Hence $N = I$, as desired. □

Recall that an $m \times m$ matrix is invertible if and only if it has rank m. Thus, we get the following special case of Theorem 1.6.11.

Corollary 1.6.12. *If an invertible matrix A has an LU factorization $A = LU$, then L and U are uniquely determined by A.*

Exercises.

1.6.1. Prove Lemma 1.6.2.

1.6.2 ([Nic, Ex. 2.7.11]). Recall the notation $L^{(m)}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_r)$ from the proof of Algorithm 1.6.6. Suppose $1 \leq i < j < k \leq m$, and let $\mathbf{c}_i$ be a column of length $m - i + 1$. Show that there is another column $\mathbf{c}_i'$ of length $m - i + 1$ such that

$$P_{j,k}L^{(m)}(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{c}_i) = L^{(m)}(\mathbf{e}_1, \ldots, \mathbf{e}_{i-1}, \mathbf{c}_i')P_{j,k}.$$

Here $P_{j,k}$ is the $m \times m$ elementary permutation matrix (see Definition 1.4.2). *Hint*: Recall that $P_{j,k}^{-1} = P_{j,k}$. Write

$$P_{j,k} = \begin{bmatrix} I_i & 0 \\ 0 & P_{j-i,k-i} \end{bmatrix}$$

in block form.

Additional recommended exercises: [Nic, §2.7].

Chapter 2

Matrix norms, sensitivity, and conditioning

In this chapter we will consider the issue of how sensitive a linear system is to small changes or errors in its coefficients. This is particularly important in applications, where these coefficients are often the results of measurement, and thus inherently subject to some level of error. Therefore, our goal is to develop some precise measure of how sensitive a linear system is to such changes.

2.1 Motivation

Consider the linear systems

$$A\mathbf{x} = \mathbf{b} \quad \text{and} \quad A\mathbf{x}' = \mathbf{b}'$$

where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.00001 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 2.00001 \end{bmatrix}, \quad \mathbf{b}' = \begin{bmatrix} 2 \\ 2.00002 \end{bmatrix}.$$

Since $\det A \neq 0$, the matrix A is invertible, and hence these linear systems have unique solutions. Indeed it is not hard to see that the solutions are

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{x}' = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

Note that even though the vector $\mathbf{b}'$ is very close to $\mathbf{b}$, the solutions to the two systems are quite different. So the solution is very sensitive to the entries of the vector of constants $\mathbf{b}$.

When the solution to the system

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

is highly sensitive to the entries of the coefficient matrix A or the vector $\mathbf{b}$ of constant terms, we say that the system is *ill-conditioned*. Ill-conditioned systems are especially problematic when the coefficients are obtained from experimental results (which always come associated with some error) or when computations are carried out by computer (which can involve round-off error).

So how do we know if a linear system is ill-conditioned? To do this, we need to discuss vector and matrix norms.

Exercises.

2.1.1. Consider the following linear systems:

$$\begin{bmatrix} 400 & -201 \\ -800 & 401 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ -200 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 401 & -201 \\ -800 & 401 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ -200 \end{bmatrix}.$$

Solve these two linear systems (feel free to use a computer) to see how the small change in the coefficient matrix results in a large change in the solution. So the solution is very sensitive to the entries of the coefficient matrix.

2.2 Normed vector spaces

Recall that every complex number $z \in \mathbb{C}$ can be written in the form

$$a + bi, \quad a, b \in \mathbb{R}.$$

The *complex conjugate* of z is

$$\bar{z} = a - bi.$$

Note that $\bar{z} = z$ if and only if $z \in \mathbb{R}$. The *absolute value*, or *magnitude*, of $z = a + bi \in \mathbb{C}$ is defined to be

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

Note that if $z \in \mathbb{R}$, then $|z|$ is the usual absolute value of z.

Definition 2.2.1 (Vector norm, normed vector space)**.** A *norm* (or *vector norm*) on an $\mathbb{F}$-vector space V (e.g. $V = \mathbb{F}^n$) is a function

$$\|\cdot\| \colon V \to \mathbb{R}$$

that satisfies the following axioms. For all $\mathbf{u}, \mathbf{v} \in V$ and $c \in \mathbb{F}$, we have

(N1) $\|\mathbf{v}\| \geq 0$,

(N2) if $\|\mathbf{v}\| = 0$ then $\mathbf{v} = 0$,

(N3) $\|c\mathbf{v}\| = |c| \, \|\mathbf{v}\|$, and

(N4) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ (the *triangule inequality*).

(Note that the codomain of the norm is $\mathbb{R}$, regardless of whether $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.) A vector space equipped with a norm is called a *normed vector space*. Thus, a normed vector space is a pair $(V, \|\cdot\|)$, where $\|\cdot\|$ is a norm on V. However, we will often just refer to V as a normed vector space, leaving it implied that we have a specific norm $\|\cdot\|$ in mind.

Note that (N3) implies that

$$\|0\| = |0|\,\|0\| = 0\|0\| = 0.$$

Thus, combined with (N2), we have

$$\|\mathbf{v}\| = 0 \iff \mathbf{v} = 0.$$

Example 2.2.2 (2-norm). Let $V = \mathbb{R}^3$. Then

$$\|(x, y, z)\|_2 := \sqrt{x^2 + y^2 + z^2}$$

is the usual norm, called the 2-*norm* or the *Euclidean norm*. It comes from the dot product, since $\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$. We can clearly generalize this to a norm on $V = \mathbb{R}^n$ by

$$\|(x_1, \ldots, x_n)\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

We can even define an analogous norm on the complex vector space $V = \mathbb{C}^n$ by

$$\|(x_1, \ldots, x_n)\|_2 := \sqrt{|x_1|^2 + \cdots + |x_n|^2}. \tag{2.1}$$

(Since the definition (2.1) works for $\mathbb{R}$ or $\mathbb{C}$, we will take it as the definition of the 2-norm from now on.) You have verified in previous courses that $\|\cdot\|$ satisfies axioms (N1)–(N4).

Example 2.2.3 (1-norm). Let $V = \mathbb{F}^n$ and define

$$\|(x_1, \ldots, x_n)\|_1 := |x_1| + \cdots + |x_n|.$$

Let's verify that this is a norm, called the 1-*norm*. Since $|c| \geq 0$ for all $c \in \mathbb{F}$, we have

$$\|(x_1, \ldots, x_n)\|_1 = |x_1| + \cdots + |x_n| \geq 0$$

and

$$\|(x_1, \ldots, x_n)\|_1 = 0 \implies |x_1| = \cdots = |x_n| = 0 \implies x_1 = \cdots = x_n = 0.$$

Thus, axioms (N1) and (N2) are satisfied. To verify axiom (N3), we see that

$$\begin{aligned}
\|c(x_1, \ldots, x_n)\| &= \|(cx_1, \ldots, cx_n)\| \\
&= |cx_1| + \cdots + |cx_n| \\
&= |c|\,|x_1| + \cdots + |c|\,|x_n| \\
&= |c|\big(|x_1| + \cdots + |x_n|\big) \\
&= |c|\,\|(x_1, \ldots, x_n)\|_1.
\end{aligned}$$

Also, we have

$$\begin{aligned}
\|(x_1, \ldots, x_n) + (y_1, \ldots, y_n)\|_1 &= \|(x_1 + y_1, \ldots, x_n + y_n)\|_1 \\
&= |x_1 + y_1| + \cdots + |x_n + y_n| \\
&\leq |x_1| + |y_1| + \cdots + |x_n| + |y_n| \qquad \text{(by the } \Delta \text{ inequality for } \mathbb{F}\text{)} \\
&= |x_1| + \cdots + |x_n| + |y_1| + \cdots + |y_n| \\
&= \|(x_1, \ldots, x_n)\| + \|(y_1, \ldots, y_n)\|.
\end{aligned}$$

So (N4) is satisfied.

In general, for any $p \in \mathbb{R}$, $p \geq 1$, one can define the *p-norm* by

$$\|(x_1, \ldots, x_n)\|_p := (|x_1|^p + \cdots + |x_2|^p)^{1/p}.$$

(It is a bit harder to show this is a norm in general. The proof uses Minkowski's inequality.) As p approaches ∞, this becomes the norm

$$\|\cdot\|_\infty \colon \mathbb{F}^n \to \mathbb{R}, \quad \|(x_1, \ldots, x_n)\|_\infty := \max\{|x_1|, \ldots, |x_n|\}, \tag{2.2}$$

which is called the *∞-norm* or *maximum norm*. See Exercise 2.2.1. In this course, we'll focus mainly on the cases $p = 1, 2, \infty$.

Remark 2.2.4. It is a theorem of analysis that all norms on $\mathbb{F}^n$ are equivalent in the sense that, if $\|\cdot\|$ and $\|\cdot\|'$ are two norms on $\mathbb{F}^n$, then there is a $c \in \mathbb{R}$, $c > 0$, such that

$$\frac{1}{c}\|\mathbf{v}\| \leq \|\mathbf{v}\|' \leq c\|\mathbf{v}\| \text{ for all } \mathbf{v} \in \mathbb{F}^n.$$

This implies that they induce the same *topology* on $\mathbb{F}^n$. That's beyond the scope of this course, but it means that, in practice, we can choose whichever norm best suits our particular application.

Exercises.

2.2.1. Show that (2.2) defines a norm on $\mathbb{F}^n$.

2.2.2. Suppose that $\|\cdot\|$ is a norm on $\mathbb{F}^n$.

(a) Show that $\|\mathbf{u} - \mathbf{v}\| \geq \big|\|\mathbf{u}\| - \|\mathbf{v}\|\big|$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$.

(b) Show that, if $\mathbf{v} \neq 0$, then $\|(1/c)\mathbf{v}\| = 1$ when $c = \|\mathbf{v}\|$.

2.2.3. Show that

$$\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2 \leq \sqrt{n}\|\mathbf{v}\|_\infty$$

for all $\mathbf{v} \in \mathbb{F}^n$.

2.2.4. Suppose that, for $p > 1$, we define

$$\|\mathbf{v}\| = |v_1|^p + |v_2|^p + \cdots + |v_n|^p \quad \text{for } \mathbf{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{F}^n.$$

Is this a norm on $\mathbb{F}^n$? If yes, prove it. If not, show that one of the axioms of a norm is violated.

2.3 Matrix norms

We would now like to define norms of *matrices*. In view of the motivation in Section 2.1, we would like the norm to somehow measure how much a matrix A can change a vector. If multiplication by A has a large effect, then small changes in $\mathbf{x}$ can result in large changes in $A\mathbf{x}$, which is a problem.

Definition 2.3.1 (Matrix norm)**.** Let $A \in M_{m,n}(\mathbb{F})$. If $\|\cdot\|_p$ and $\|\cdot\|_q$ are norms on $\mathbb{F}^n$ and $\mathbb{F}^m$ respectively, we define

$$\|A\|_{p,q} = \max\left\{\frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} : \mathbf{x} \in \mathbb{F}^n,\ \mathbf{x} \neq 0\right\}. \tag{2.3}$$

This is called the *matrix norm*, or *operator norm*, of A with respect to the norms $\|\cdot\|_p$ and $\|\cdot\|_q$. We also say that $\|A\|_{p,q}$ is the matrix norm *associated to* the norms $\|\cdot\|_p$ and $\|\cdot\|_q$.

The next lemma tells us that, in order to compute the matrix norm (2.3), it is not necessary to check the value of $\|A\mathbf{x}\|_q/\|\mathbf{x}\|_p$ for every $\mathbf{x} \neq 0$. Instead, it is enough to consider unit vectors.

Proposition 2.3.2. *In the setup Definition 2.3.1, we have*

$$\|A\|_{q,p} = \max\left\{\|A\mathbf{x}\|_q : \mathbf{x} \in \mathbb{F}^n,\ \|\mathbf{x}\|_p = 1\right\}.$$

Proof. Let

$$S = \left\{\frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} : \mathbf{x} \in \mathbb{F}^n,\ \mathbf{x} \neq 0\right\} \quad \text{and} \quad T = \left\{\|A\mathbf{x}\|_q : x \in \mathbb{F}^n,\ \|\mathbf{x}\|_p = 1\right\}.$$

These are both sets of nonnegative real numbers, and we wish to show that they have the same maximum. To do this, we will show that these sets are actually the same (which is a stronger assertion).

First note that, if $\|\mathbf{x}\|_p = 1$, then

$$\frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} = \|A\mathbf{x}\|_q.$$

Thus $T \subseteq S$.

Now we want to show the reverse inclusion: $S \subseteq T$. Suppose $s \in S$. Then there is some $\mathbf{x} \neq 0$ such that

$$s = \frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p}.$$

Define

$$c = \frac{1}{\|\mathbf{x}\|_p} \in \mathbb{R}.$$

Then, by (N3),

$$\|c\mathbf{x}\|_p = |c|\,\|\mathbf{x}\|_p = \frac{1}{\|\mathbf{x}\|_p}\|\mathbf{x}\|_p = 1.$$

Thus we have

$$s = \frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} = \frac{|c|\,\|A\mathbf{x}\|_q}{|c|\,\|\mathbf{x}\|_p} \overset{\text{(N3)}}{=} \frac{\|cA\mathbf{x}\|_q}{\|c\mathbf{x}\|_p} = \frac{\|A(c\mathbf{x})\|_q}{\|c\mathbf{x}\|_p} = \|A(c\mathbf{x})\|_q \in T,$$

since $\|c\mathbf{x}\|_p = 1$. Thus we have shown the reverse inclusion $S \subseteq T$.

Having shown both inclusions, it follows that $S = T$, and hence their maxima are equal. □

Remark 2.3.3. In general, a set of real numbers may not have a maximum. (Consider the set $\mathbb{R}$ itself.) Thus, it is not immediately clear that $\|A\|_{p,q}$ is indeed well defined. However, one can use Proposition 2.3.2 to show that it is well defined. The set $\{\mathbf{x} \in \mathbb{F}^n : \|x\|_p = 1\}$ is *compact*, and the function $\mathbf{x} \mapsto \|A\mathbf{x}\|_q$ is *continuous*. It follows from a theorem in analysis that this function attains a maximum.

When we use the same type of norm (e.g. the 1-norm, the 2-norm, or the ∞-norm) in both the domain and the codomain, we typically use a single subscript on the matrix norm. Thus, for instance,

$$\|A\|_1 := \|A\|_{1,1}, \quad \|A\|_2 := \|A\|_{2,2}, \quad \|A\|_\infty := \|A\|_{\infty,\infty}.$$

Note that, in principle, we could choose different types of norms for the domain and codomain. However, in practice, we usually choose the same one.

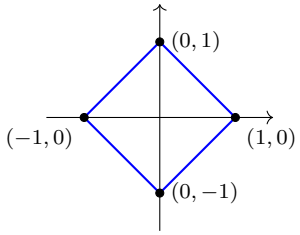
Example 2.3.4. Let's calculate $\|A\|_1$ for $\mathbb{F} = \mathbb{R}$ and

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

We will use Proposition 2.3.2, so we need to consider the set

$$\{(x, y) : \|(x, y)\|_1 = 1\} = \{(x, y) : |x| + |y| = 1\}.$$

This set is the union of the four blue line segments shown below:



By Proposition 2.3.2, we have

$$\begin{aligned} \|A\|_1 &= \max\{\|A\mathbf{x}\|_1 : \|x\|_1 = 1\} \\ &= \max\left\{\left\|A\begin{bmatrix} x \\ y \end{bmatrix}\right\|_1 : |x| + |y| = 1\right\} \\ &= \max\left\{\left\|\begin{bmatrix} 2x + 3y \\ x - 5y \end{bmatrix}\right\|_1 : |x| + |y| = 1\right\} \end{aligned}$$

$$\begin{aligned}
&= \max\{|2x+3y| + |x-5y| : |x|+|y| = 1\} \\
&\leq \max\{2|x| + 3|y| + |x| + 5|y| : |x|+|y| = 1\} \\
&= \max\{3|x| + 8|y| : |x|+|y| = 1\} \\
&= \max\{3|x| + 8(1-|x|) : 0 \leq |x| \leq 1\} \\
&= \max\{8 - 5x : 0 \leq x \leq 1\} \\
&= 8.
\end{aligned}$$

So we know that $\|A\|_1 = \max\{\|A\mathbf{x}\|_1 : \|x\|_1 = 1\}$ is at most 8. To show that it is *equal to* 8, it is enough to show that $8 \in \{\|A\mathbf{x}\|_1 : \|\mathbf{x}\|_1 = 1\}$. So we want to find an $\mathbf{x} \in \mathbb{R}^2$ such that

$$\|\mathbf{x}\|_1 = 1 \quad \text{and} \quad \|A\mathbf{x}\|_1 = 8.$$

Indeed, if we take $\mathbf{x} = (0,1)$, then $\|\mathbf{x}\|_1 = 1$, and

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

and so $\|A\mathbf{x}\|_1 = |3| + |-5| = 8$. Thus $\|A\|_1 = 8$.

Note that, in Example 2.3.4, the matrix norm $\|A\|_1$ was precisely the 1-norm of one of its columns (the second column, to be precise). The following result gives the general situation.

Theorem 2.3.5. *Suppose $A = [a_{ij}] \in M_{m,n}(\mathbb{F})$. Then*

(a) $\|A\|_1 = \max\{\sum_{i=1}^m |a_{ij}| : 1 \leq j \leq n\}$, *the maximum of the 1-norms of the columns of A;*

(b) $\|A\|_\infty = \max\left\{\sum_{j=1}^n |a_{ij}| : 1 \leq i \leq m\right\}$, *the maximum of the 1-norms of the rows of A;*

(c) $\|A\|_2 \leq \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} =: \|A\|_F$, *called the* Frobenius norm *of A, which is the 2-norm of A, viewing it as a vector in $\mathbb{F}^{nm}$.*

Proof. We will prove part (a) and leave parts (b) and (c) as Exercise 2.3.1.

Recall that $\mathbf{a}_j = A\mathbf{e}_j$ is the j-th column of A, for $1 \leq j \leq n$. We have $\|\mathbf{e}_j\|_1 = 1$ and

$$\|A\mathbf{e}_j\|_1 = \sum_{i=1}^m |a_{i,j}|$$

is the 1-norm of the j-th column of A. Thus, by definition,

$$\|A\|_1 \geq \max\left\{\sum_{i=1}^m |a_{ij}| : 1 \leq j \leq n\right\}.$$

It remains to prove the reverse inequality. If $\mathbf{x} = (x_1, \ldots, x_n)$, we have

$$A\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n.$$

Thus

$$\begin{aligned}\|A\mathbf{x}\|_1 &= \|x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n\|_1 \\ &\leq \|x_1\mathbf{a}_1\|_1 + \cdots + \|x_n\mathbf{a}_n\|_1 && \text{(by the triangle inequality (N4))} \\ &\leq |x_1|\,\|\mathbf{a}_1\|_1 + \cdots + |x_n|\,\|\mathbf{a}_n\|_1. && \text{(by (N3))}\end{aligned}$$

Now, suppose the column of A with the maximum 1-norm is the j-th column, and suppose $\|\mathbf{x}\|_1 = 1$. Then we have

$$\|A\mathbf{x}\|_1 \leq |x_1|\,\|\mathbf{a}_j\|_1 + \cdots + |x_n|\,\|\mathbf{a}_j\|_1 = (|x_1| + \cdots + |x_n|)\,\|\mathbf{a}_j\|_1 = \|\mathbf{a}_j\|_1.$$

This completes the proof. □

Note that part (c) of Theorem 2.3.5 involves an inequality. In practice, the norm $\|A\|_2$ is difficult to compute, while the Frobenius norm is much easier.

The following theorem summarizes the most important properties of matrix norms.

Theorem 2.3.6 (Properties of matrix norms)**.** *Suppose $\|\cdot\|$ is a family of norms on $\mathbb{F}^n$, $n \geq 1$. We also use the notation $\|A\|$ for the matrix norm with respect to these vector norms.*

(a) *For all $\mathbf{v} \in \mathbb{F}^n$ and $A \in M_{m,n}(\mathbb{F})$, we have $\|A\mathbf{v}\| \leq \|A\|\,\|\mathbf{v}\|$.*

(b) $\|I\| = 1$.

(c) *For all $A \in M_{m,n}(\mathbb{F})$, we have $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$.*

(d) *For all $c \in \mathbb{F}$ and $A \in M_{m,n}(\mathbb{F})$, we have $\|cA\| = |c|\,\|A\|$.*

(e) *For all $A, B \in M_{m,n}(\mathbb{F})$, we have $\|A + B\| \leq \|A\| + \|B\|$.*

(f) *For all $A \in M_{m,n}(\mathbb{F})$ and $B \in M_{n,k}(\mathbb{F})$, we have $\|AB\| \leq \|A\|\,\|B\|$.*

(g) *For all $A \in M_{n,n}(\mathbb{F})$, we have $\|A^k\| \leq \|A\|^k$ for all $k \geq 1$.*

(h) *If $A \in \mathrm{GL}(n, \mathbb{F})$, then $\|A^{-1}\| \geq \frac{1}{\|A\|}$.*

Proof. We prove parts (a) and (h) and leave the remaining parts as Exercise 2.3.3. Suppose $\mathbf{v} \in \mathbb{F}^n$ and $A \in M_{m,n}(\mathbb{F})$. If $\mathbf{v} = 0$, then

$$\|A\mathbf{v}\| = 0 = \|A\|\,\|\mathbf{v}\|.$$

Now suppose $\mathbf{v} \neq 0$. Then

$$\frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} \leq \max\left\{\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{F}^n,\ \mathbf{x} \neq 0\right\} = \|A\|.$$

Multiplying both sides by $\|\mathbf{v}\|$ then gives (a).

Now suppose $A \in \mathrm{GL}(n, \mathbb{F})$. By the definition of $\|A\|$, we can choose $\mathbf{x} \in \mathbb{F}^n$, $\mathbf{x} \neq 0$, such that

$$\|A\| = \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Then we have

$$\frac{1}{\|A\|} = \frac{\|\mathbf{x}\|}{\|A\mathbf{x}\|} = \frac{\|A^{-1}(A\mathbf{x})\|}{\|A\mathbf{x}\|} \leq \|A^{-1}\|$$

by part (a) for A^{-1}. □

Exercises.

2.3.1. Prove parts (b) and (c) of Theorem 2.3.5. For part (c), use the *Cauchy–Schwarz inequality*

$$|\mathbf{u}^T\mathbf{v}| \leq \|\mathbf{u}\|\,\|\mathbf{v}\|, \quad \mathbf{u},\mathbf{v} \in \mathbb{F}^n,$$

(here we view the 1×1 matrix $\mathbf{u}^T\mathbf{v}$ as an element of $\mathbb{F}$) and note that the entries of the product Ax are of the form $\mathbf{u}^T\mathbf{x}$, where $\mathbf{u}$ is a row of A.

2.3.2. Suppose $A \in M_{n,n}(\mathbb{F})$ and that λ is an eigenvalue of A. Show that, for *any* choice of vector norm on $\mathbb{F}^n$, we have $\|A\| \geq |\lambda|$, where $\|A\|$ is the associated matrix norm of A.

2.3.3. Complete the proof of Theorem 2.3.6.

2.3.4. What is the matrix norm of a zero matrix?

2.3.5. Suppose $A \in M_{m,n}(\mathbb{F})$ and that there is some fixed $k \in \mathbb{R}$ such that $\|A\mathbf{v}\| \leq k\|\mathbf{v}\|$ for all $\mathbf{v} \in \mathbb{F}^n$. (Here we have fixed some arbitrary norms on $\mathbb{F}^n$ and $\mathbb{F}^m$.) Show that $\|A\| \leq k$.

2.3.6. For each of the following matrices, find $\|A\|_1$ and $\|A\|_\infty$.

(a) $\begin{bmatrix} -4 \\ 1 \\ 5 \end{bmatrix}$

(b) $\begin{bmatrix} -4 & 1 & 5 \end{bmatrix}$

(c) $\begin{bmatrix} -9 & 0 & 2 & 9 \end{bmatrix}$

(d) $\begin{bmatrix} -5 & 8 \\ 6 & 2 \end{bmatrix}$

(e) $\begin{bmatrix} 2i & 5 \\ 4 & -i \\ 3 & 5 \end{bmatrix}$

2.4 Conditioning

Our goal is to develop some measure of how "good" a matrix is as a coefficient matrix of a linear system. That is, we want some measure that allows us to know whether or not a matrix can exhibit the bad behaviour we saw in Section 2.1.

Definition 2.4.1 (Condition number)**.** Suppose $A \in \mathrm{GL}(n,\mathbb{F})$ and let $\|\cdot\|$ denote a norm on $\mathbb{F}^n$ as well as the associated matrix norm. The value

$$\kappa(A) = \|A\|\,\|A^{-1}\| \geq 1$$

is called the *condition number* of the matrix A, relative to the choice of norm $\|\cdot\|$.

Note that the condition number depends on the choice of norm. The fact that $\kappa(A) \geq 1$ follows from Theorem 2.3.6(h).

Examples 2.4.2. (a) Consider the matrix from Section 2.1. We have

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.00001 \end{bmatrix} \quad \text{and} \quad A^{-1} = 10^5 \begin{bmatrix} 1.00001 & -1 \\ -1 & 1 \end{bmatrix}.$$

Therefore, with respect to the 1-norm,

$$\kappa(A) = (2.00001)(2.00001 \cdot 10^5) \geq 4 \cdot 10^5.$$

(b) If

$$B = \begin{bmatrix} 2 & 2 \\ 4 & 3 \end{bmatrix}, \quad \text{then} \quad B^{-1} = \frac{-1}{2} \begin{bmatrix} 3 & -2 \\ -4 & 2 \end{bmatrix}.$$

Thus, with respect to the 1-norm,

$$\kappa(B) = 6 \cdot \frac{7}{2} = 21.$$

(c) If $a \in \mathbb{R}$ and

$$C = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad C^{-1} = \begin{bmatrix} 1 & -a \\ 0 & 1 \end{bmatrix}$$

and so, with respect to the 1-norm,

$$\kappa(C) = (|a| + 1)^2.$$

Since a is arbitrary, this example shows that the condition number can be arbitrarily large.

Lemma 2.4.3. *If $c \in \mathbb{F}^\times$, then for every invertible matrix A, we have $\kappa(cA) = \kappa(A)$.*

Proof. Note that $(cA)^{-1} = c^{-1}A^{-1}$. Thus

$$\kappa(cA) = \|cA\| \, \|c^{-1}A^{-1}\| = |c| \, |c^{-1}| \, \|A\| \, \|A^{-1}\| = |cc^{-1}|\kappa(A) = \kappa(A). \qquad \square$$

Example 2.4.4. If

$$M = \begin{bmatrix} 10^{16} & 0 \\ 0 & 10^{16} \end{bmatrix},$$

then $M = 10^{16}I$. Hence $\kappa(M) = \kappa(I) = 1$ by Theorem 2.3.6(b).

Now that we've defined the condition number of a matrix, what does it have to do with the situation discussed in Section 2.1? Suppose we want to solve the system

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

where $\mathbf{b}$ is determined by some experiment (thus subject to measurement error) or computation (subject to rounding error). Let $\Delta\mathbf{b}$ be some small perturbation in $\mathbf{b}$, so that $\mathbf{b}' = \mathbf{b} + \Delta\mathbf{b}$ is close to $\mathbf{b}'$. Then there is some new solution

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

Let $\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x}$. We say that the error $\Delta\mathbf{b}$ induces (via A) the error $\Delta\mathbf{x}$ in $\mathbf{x}$.

The norms of the vectors $\Delta\mathbf{b}$ and $\Delta\mathbf{x}$ are the *absolute errors*; the quotients

$$\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \quad \text{and} \quad \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}$$

are the *relative errors*. In general, we are interested in the relative error.

Theorem 2.4.5. *With the above notation,*

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Proof. We have

$$A(\Delta\mathbf{x}) = A(\mathbf{x}' - \mathbf{x}) = A\mathbf{x}' - A\mathbf{x} = \mathbf{b}' - \mathbf{b} = \Delta\mathbf{b},$$

and so $\Delta\mathbf{x} = A^{-1}\Delta\mathbf{b}$. Thus, by Theorem 2.3.6(a), we have

$$\|\Delta\mathbf{x}\| \leq \|A^{-1}\|\,\|\Delta\mathbf{b}\| \quad \text{and} \quad \|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|.$$

Thus

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|A^{-1}\|\,\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|/\|A\|} = \kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

□

In fact, there always exists a choice of $\mathbf{b}$ and $\Delta\mathbf{b}$ such that we have equality in Theorem 2.4.5. See Exercise 2.4.1.

Theorem 2.4.5 says that, when solving the linear system,

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

the condition number is roughly the rate at which the solution $\mathbf{x}$ will change with respect to a change in $\mathbf{b}$. So if the condition number is large, then even a small error/change in $\mathbf{b}$ may cause a large error/change in $\mathbf{x}$. More precisely, the condition number is the maximum ratio of the relative error in $\mathbf{x}$ to the relative error in $\mathbf{b}$.

If $\kappa(A)$ is close to 1, we say that A is *well-conditioned*. On the other hand, if $\kappa(A)$ is large, we say that A is *ill-conditioned*. Note that these terms are a bit vague; we do not have a precise notion of how large $\kappa(A)$ needs to be before we say A is ill-conditioned. This depends somewhat on the particular situation.

Example 2.4.6. Consider the situation from Section 2.1 and Example 2.4.2(a):

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1.00001 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 2.00001 \end{bmatrix}, \quad \mathbf{b}' = \begin{bmatrix} 2 \\ 2.00002 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}' = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

Thus

$$\|\Delta\mathbf{b}\|_1 = 10^{-5}, \quad \|\mathbf{b}\|_1 = 4.00001, \quad \|\Delta\mathbf{x}\|_1 = 2, \quad \|\mathbf{x}\|_1 = 2.$$

So we have

$$\kappa(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} = (2.00001)^2 \cdot 10^5 \cdot \frac{10^{-5}}{4.00001} \geq 1 = \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|},$$

as predicted by Theorem 2.4.5. The fact that A is ill-conditioned explains the phenomenon we noticed in Section 2.1: that a small change in $\mathbf{b}$ can result in a large change in the solution $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{b}$.

What happens if there is some small change/error in the matrix A in addition to a change/error in $\mathbf{b}$? (See Exercise 2.1.1.) In general, one can show that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq c \cdot \kappa(A) \left(\frac{\|\Delta\mathbf{b}|}{\|\mathbf{b}\|} + \frac{\|\Delta A\|}{\|A\|} \right),$$

where $c = \|(\Delta A)A^{-1}\|$ or $c = \|A^{-1}(\Delta A)\|$ (and we assume that $c < 1$). For a proof see, for example, [ND77, Th. 6.29].

While an ill-conditioned coefficient matrix A tells us that the system $A\mathbf{x} = \mathbf{b}$ *can* be ill-conditioned (see Exercise 2.4.1), it does *not imply in general* that this system is ill-conditioned. See Exercise 2.4.2. Thus, $\kappa(A)$ measures the worse case scenario for a linear system with coefficient matrix A.

Exercises.

2.4.1. Show that there always exists a choice of $\mathbf{b}$ and $\Delta\mathbf{b}$ such that we have equality in Theorem 2.4.5.

2.4.2. If s is very large, we know from Example 2.4.2(c) that the matrix

$$C = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

is ill-conditioned. Show that, if $\mathbf{b} = (1, 1)$, then the system $C\mathbf{x} = \mathbf{b}$ satisfies

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq 3\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

and is therefore well-conditioned. (Here we use the 1-norm.) On the other hand, find a choice of $\mathbf{b}$ for which the system $C\mathbf{x} = \mathbf{b}$ is ill-conditioned.

2.4.3. Suppose $k \in \mathbb{F}^\times$ and find the condition number of the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{k} \end{bmatrix}$$

using either the 1-norm or the ∞-norm.

2.4.4 ([ND77, 6.5.16]). Consider the system of equations

$$\begin{array}{rcrcl} 0.89x_1 & + & 0.53x_2 & = & 0.36 \\ 0.47x_1 & + & 0.28x_2 & = & 0.19 \end{array}$$

with exact solution $x_1 = 1$, $x_2 = -1$.

(a) Find $\Delta\mathbf{b}$ so that if you replace the right-hand side $\mathbf{b}$ by $\mathbf{b} + \Delta\mathbf{b}$, the exact solution will be $x_1 = 0.47$, $x_2 = 0.11$.

(b) Is the system ill-conditioned or well-conditioned?

(c) Find the condition number for the coefficient matrix of the system using the ∞-norm.

Chapter 3

Orthogonality

The notion of orthogonality is fundamental in linear algebra. You've encountered this concept in previous courses. Here we will delve into this subject in further detail. We begin by briefly reviewing the Gram–Schmidt algorithm, orthogonal complements, orthogonal projection, and diagonalization. We then discuss hermitian and unitary matrices, which are complex analogues of symmetric and orthogonal matrices that you've seen before. Afterwards, we learn about Schur decomposition and prove the important spectral and Cayley–Hamilton theorems. We also define positive definite matrices and consider Cholesky and QR factorizations. We conclude with a discussion of computing/estimating eigenvalues, including the Gershgorin circle theorem.

3.1 Orthogonal complements and projections

In this section we briefly review the concepts of orthogonality, orthogonal complements, orthogonal projections, and the Gram–Schmidt algorithm. Since you have seen this material in previous courses, we will move quickly and omit proofs. We follow the presentation in [Nic, §8.1], and proofs can be found there.

Recall that $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. If $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}^n$, we define

$$\bar{\mathbf{v}} = (\overline{v_1}, \ldots, \overline{v_n}) \in \mathbb{F}^n.$$

Then we define the *inner product* on $\mathbb{F}^n$ as follows:

$$\langle \mathbf{u}, \mathbf{v} \rangle := \bar{\mathbf{u}}^T \mathbf{v} = \overline{u_1} v_1 + \cdots + \overline{u_n} v_n, \quad \mathbf{v} = (v_1, \ldots, v_n),\ \mathbf{u} = (u_1, \ldots, u_n). \tag{3.1}$$

When $\mathbb{F} = \mathbb{R}$, this is the usual *dot product*. (Note that, in [Nic, §8.7], the inner product is defined with the complex conjugation on the second vector.) The inner product has the following important properties: For all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{F}^n$ and $c, d \in \mathbb{F}$, we have

(IP1) $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$,

(IP2) $\langle c\mathbf{u} + d\mathbf{v}, \mathbf{w} \rangle = \bar{c}\langle \mathbf{u}, \mathbf{w} \rangle + \bar{d}\langle \mathbf{v}, \mathbf{w} \rangle$,

(IP3) $\langle \mathbf{u}, c\mathbf{v} + d\mathbf{w} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle + d\langle \mathbf{u}, \mathbf{w} \rangle$,

(IP4) $\langle \mathbf{u}, \mathbf{u} \rangle \in \mathbb{R}$ and $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$,

(IP5) $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = 0$.

More generally, if V is a vector space over $\mathbb{F}$, then any map

$$\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{R}$$

satisfying (IP1)–(IP5) is called an inner product on V. In light of (IP4), for any inner product, we may define

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle},$$

and one can check that this defines a norm on V. For the purposes of this course, we will stick to the particular inner product (3.1). In this case $\|\mathbf{v}\|$ is the usual 2-norm:

$$\|\mathbf{v}\| = \sqrt{|v_1|^2 + \cdots + |v_n|^2}.$$

Throughout this chapter, $\| \cdot \|$ will denote the 2-norm.

Definition 3.1.1 (Orthogonal, orthonormal)**.** We say that $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$ are *orthogonal*, and we write $\mathbf{u} \perp \mathbf{v}$, if

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

A set $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is orthogonal if $\mathbf{v}_i \perp \mathbf{v}_j$ for all $i \neq j$. If, in addition, we have $\|\mathbf{v}_i\| = 1$ for all i, then we say the set is *orthonormal*. An *orthogonal basis* is a basis that is also an orthogonal set. Similarly, an *orthonormal basis* is a basis that is also an orthonormal set.

Proposition 3.1.2. *Let U be a subspace of $\mathbb{F}^n$.*

(a) *Every orthogonal subset $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ in U is a subset of an orthogonal basis of U. (We say that we can* extend *any orthogonal subset to an orthogonal basis.)*

(b) *The subspace U has an orthogonal basis.*

Proof. You saw this in previous courses, so we will omit the proof here. It can be found in [Nic, Th. 8.1.1]. □

Theorem 3.1.3 (Gram–Schmidt orthogonalization algorithm)**.** *If $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is any basis of a subspace of $\mathbb{F}^n$, construct $\mathbf{u}_1, \ldots, \mathbf{u}_m \in U$ successively as follows:*

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1, \\
\mathbf{u}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1, \\
\mathbf{u}_3 &= \mathbf{v}_3 - \frac{\langle \mathbf{u}_1, \mathbf{v}_3 \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1 - \frac{\langle \mathbf{u}_2, \mathbf{v}_3 \rangle}{\|\mathbf{u}_2\|^2} \mathbf{u}_2, \\
&\vdots \\
\mathbf{u}_m &= \mathbf{v}_m - \frac{\langle \mathbf{u}_1, \mathbf{v}_m \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1 - \frac{\langle \mathbf{u}_2, \mathbf{v}_m \rangle}{\|\mathbf{u}_2\|^2} \mathbf{u}_2 - \cdots - \frac{\langle \mathbf{u}_{m-1}, \mathbf{v}_m \rangle}{\|\mathbf{u}_{m-1}\|^2} \mathbf{u}_{m-1}.
\end{aligned}$$

Then

(a) $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ *is an orthogonal basis of* U*, and*

(b) $\operatorname{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} = \operatorname{Span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ *for each* $k = 1, 2, \dots, m$.

Proof. You saw this in previous courses, and so we will not repeat the proof here. See [Nic, Th. 8.1.2]. Note that, in [Nic, Th. 8.1.2], it is assumed that $\mathbb{F} = \mathbb{R}$, in which case $\langle \mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. If we wish to allow $\mathbb{F} = \mathbb{C}$, we have (IP1) instead. Then it is important that we write $\langle \mathbf{u}_k, \mathbf{v} \rangle$ in the Gram–Schmidt algorithm instead of $\langle \mathbf{v}, \mathbf{u}_k \rangle$. □

Example 3.1.4. Let's find an orthogonal basis for the row space of

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

Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ denote the rows of A. One can check that these rows are linearly independent. (Reduce A to echelon form and note that it has rank 3.) So they give a basis of the row space. Let's use the Gram–Schmidt algorithm to find an orthogonal basis:

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 = (1, -1, 0, 1), \\
\mathbf{u}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{u}_1, \mathbf{v}_2 \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1 = (2, -1, -2, 3) - \frac{6}{3}(1, -1, 0, 1) = (0, 1, -2, 1), \\
\mathbf{u}_3 &= \mathbf{v}_3 - \frac{\langle \mathbf{u}_1, \mathbf{v}_3 \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1 - \frac{\langle \mathbf{u}_2, \mathbf{v}_3 \rangle}{\|\mathbf{u}_2\|^2} \mathbf{u}_2 \\
&= (0, 2, 0, 1) - \frac{-1}{3}(1, -1, 0, 1) - \frac{3}{6}(0, 1, -2, 1) = \left(\frac{1}{3}, \frac{7}{6}, 1, \frac{5}{6} \right)
\end{aligned}$$

It can be nice to eliminate the fractions (see Remark 3.1.5), so

$$\{(1, -1, 0, 1), (0, 1, -2, 1), (2, 7, 6, 5)\}$$

is an orthogonal basis for the row space of A. If we wanted an orthonormal basis, we would divide each of these basis vectors by its norm.

Remark 3.1.5. Note that, for $c \in \mathbb{F}^\times$ and $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$, we have

$$\frac{\langle c\mathbf{u}, \mathbf{v} \rangle}{\|c\mathbf{u}\|^2}(c\mathbf{u}) = \frac{\bar{c}\langle \mathbf{u}, \mathbf{v} \rangle}{|c|^2 \|\mathbf{u}\|^2}(c\mathbf{u}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|^2}\mathbf{u}.$$

Therefore, in the Gram–Schmidt algorithm, replacing some $\mathbf{u}_i$ by $c\mathbf{u}_i$, $c \neq 0$, does not affect any of the subsequent steps. This is useful in computations, since we can, for instance, clear denominators.

Orthogonal (especially orthonormal) bases are particularly nice since it is easy to write a vector as a linear combinations of the elements of such a basis.

Proposition 3.1.6. *Suppose* $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ *is an orthogonal basis of a subspace* U *of* $\mathbb{F}^n$*. Then, for any* $\mathbf{v} \in U$*, we have*

$$\mathbf{v} = \sum_{i=1}^{m} \frac{\langle \mathbf{u}_i, \mathbf{v} \rangle}{\|\mathbf{u}_i\|^2} \mathbf{u}_i. \tag{3.2}$$

In particular, if the basis is orthonormal, then

$$\mathbf{v} = \sum_{i=1}^{m} \langle \mathbf{u}_i, \mathbf{v} \rangle \mathbf{u}_i. \tag{3.3}$$

Proof. Since $\mathbf{v} \in U$, we can write

$$\mathbf{v} = \sum_{i=1}^{m} c_i \mathbf{u}_i \quad \text{for some } c_1, \ldots, c_m \in \mathbb{F}.$$

Then, for $j = 1, \ldots, m$, we have

$$\langle \mathbf{u}_j, \mathbf{v} \rangle = \left\langle \mathbf{u}_j, \sum_{i=1}^{m} c_i \mathbf{u}_i \right\rangle \overset{\text{(IP3)}}{=} \sum_{i=1}^{m} c_i \langle \mathbf{u}_j, \mathbf{u}_i \rangle = c_j \langle \mathbf{u}_j, \mathbf{u}_j \rangle = c_j \|\mathbf{u}_j\|^2.$$

Thus $c_j = \frac{\langle \mathbf{u}_j, \mathbf{v} \rangle}{\|\mathbf{u}_j\|^2}$, as desired. □

Remark 3.1.7. What happens if you apply the Gram–Schmidt algorithm to a set of vectors that is not linearly independent? Remember that a list of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is linearly dependent if and only if one of the vectors, say $\mathbf{v}_k$, is a linear combination of the previous ones. Then, using Proposition 3.1.6, one can see that the Gram–Schmidt algorithm will give $\mathbf{u}_k = 0$. Thus, you can still apply the Gram–Schmidt algorithm to linearly dependent sets, as long as you simply throw out any zero vectors that you obtain in the process.

Definition 3.1.8. If U is a subspace of $\mathbb{F}^n$, we define the *orthogonal complement* of U by

$$U^\perp := \{\mathbf{v} \in \mathbb{F}^n : \langle \mathbf{u}, \mathbf{v} \rangle = 0 \text{ for all } \mathbf{u} \in U\}.$$

We read $U^\perp$ as "U-perp".

Lemma 3.1.9. *Let U be a subspace of $\mathbb{F}^n$.*

(a) *$U^\perp$ is a subspace of $\mathbb{F}^n$.*

(b) *$\{0\}^\perp = \mathbb{F}^n$ and $(\mathbb{F}^n)^\perp = \{0\}$.*

(c) *If $U = \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, then $U^\perp = \{\mathbf{v} \in \mathbb{F}^n : \langle \mathbf{v}, \mathbf{u}_i \rangle = 0$ for all $i = 1, 2, \ldots, k\}$.*

Proof. You saw these properties of the orthogonal complement in previous courses, so we will not repeat the proofs there. See [Nic, Lem. 8.1.2]. □

Definition 3.1.10 (Orthogonal projection)**.** Let U be a subspace of $\mathbb{F}^n$ with orthogonal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$. If $\mathbf{v} \in \mathbb{F}^n$, then the vector

$$\operatorname{proj}_U \mathbf{v} = \frac{\langle \mathbf{u}_1, \mathbf{v} \rangle}{\|\mathbf{u}_1\|^2} \mathbf{u}_1 + \frac{\langle \mathbf{u}_2, \mathbf{v} \rangle}{\|\mathbf{u}_2\|^2} \mathbf{u}_2 + \cdots + \frac{\langle \mathbf{u}_m, \mathbf{v} \rangle}{\|\mathbf{u}_m\|^2} \mathbf{u}_m \tag{3.4}$$

is called the *orthogonal projection* of $\mathbf{v}$ onto U. For the zero subspace $U = \{0\}$, we define $\operatorname{proj}_{\{0\}} \mathbf{x} = 0$.

Noting the similarity between (3.4) and (3.2), we see that

$$\mathbf{v} = \operatorname{proj}_U \mathbf{v} \iff \mathbf{v} \in U.$$

(The forward implication follows from the fact that $\operatorname{proj}_U \mathbf{v} \in U$, while the reverse implication follows from Proposition 3.1.6.) Also note that the k-th step of Gram–Schmidt algorithm can be written as

$$\mathbf{u}_k = \mathbf{v}_k - \operatorname{proj}_{\operatorname{Span}\{\mathbf{u}_1,\dots,\mathbf{u}_{k-1}\}} \mathbf{v}_k.$$

Proposition 3.1.11. *Suppose U is a subspace of $\mathbb{F}^n$ and $\mathbf{v} \in \mathbb{F}^n$. Define $\mathbf{p} = \operatorname{proj}_U \mathbf{v}$.*

(a) *We have $\mathbf{p} \in U$ and $\mathbf{v} - \mathbf{p} \in U^\perp$.*

(b) *The vector $\mathbf{p}$ is the vector in U closest to $\mathbf{v}$ in the sense that*

$$\|\mathbf{v} - \mathbf{p}\| < \|\mathbf{v} - \mathbf{u}\| \text{ for all } \mathbf{u} \in U,\ \mathbf{u} \neq \mathbf{p}.$$

Proof. You saw this in previous courses, so we will not repeat the proofs here. See [Nic, Th. 8.1.3]. □

Proposition 3.1.12. *Suppose U is a subspace of $\mathbb{F}^n$. Define*

$$T\colon \mathbb{F}^n \to \mathbb{F}^n, \quad T(\mathbf{v}) = \operatorname{proj}_U \mathbf{v}, \quad \mathbf{v} \in \mathbb{F}^n.$$

(a) *T is a linear operator.*

(b) *$\operatorname{im} T = U$ and $\ker T = U^\perp$.*

(c) *$\dim U + \dim U^\perp = n$.*

Proof. See [Nic, Th. 8.1.4]. □

If U is a subspace of $\mathbb{F}^n$, then every $\mathbf{v} \in \mathbb{F}^n$ can be written *uniquely* as a sum of a vector in U and a vector in $U^\perp$. Precisely, we have

$$\mathbf{v} = \operatorname{proj}_U \mathbf{v} + \operatorname{proj}_{U^\perp} \mathbf{v}.$$

Exercises.

3.1.1. Prove that the inner product defined by (3.1) satisfies conditions (IP1)–(IP5).

3.1.2. Suppose U and V are subspaces of $\mathbb{F}^n$. We write $U \perp V$ if

$$u \perp v \quad \text{for all } u \in U,\ v \in V.$$

Show that if $U \perp V$, then $\operatorname{proj}_U \circ \operatorname{proj}_V = 0$.

Additional recommended exercises:

- [Nic, §8.1]: All exercises
- [Nic, §8.7]: 8.7.1–8.7.4

3.2 Diagonalization

In this section we quickly review the topics of eigenvectors, eigenvalues, and diagonalization that you saw in previous courses. For a more detailed review of this material, see [Nic, §3.3].

Throughout this section, we suppose that $A \in M_{n,n}(\mathbb{F})$. Recall that if

$$A\mathbf{v} = \lambda\mathbf{v} \quad \text{for some } \lambda \in \mathbb{F},\ \mathbf{v} \in \mathbb{F}^n,\ \mathbf{v} \neq 0, \tag{3.5}$$

then we say that $\mathbf{v}$ is an *eigenvector* of A with corresponding *eigenvalue* λ. So $\lambda \in \mathbb{F}$ is an eigenvector of A if (3.5) is satisfied for some *nonzero* vector $\mathbf{v} \in \mathbb{F}^n$.

You learned in previous courses how to find the eigenvalues of A. The eigenvalues of A are precisely the roots of the *characteristic polynomial*

$$c_A(x) := \det(xI - A).$$

If $\mathbb{F} = \mathbb{C}$, then the characteristic polynomial will factor completely. That is, we have

$$\det(xI - A) = (x - \lambda_1)^{m_{\lambda_1}}(x - \lambda_2)^{m_{\lambda_2}} \cdots (x - \lambda_k)^{m_{\lambda_k}},$$

where the $\lambda_1, \ldots, \lambda_k$ are distinct. Then m_{λ_i} is the called *algebraic multiplicity* of λ_i. It follows that $\sum_{i=1}^{k} m_{\lambda_i} = n$.

For an eigenvalue λ of A, the set of solutions to the equation

$$A\mathbf{x} = \lambda\mathbf{x} \qquad \text{or} \qquad (A - \lambda I)\mathbf{x} = 0$$

is the associated *eigenspace* E_λ. The eigenvectors corresponding to λ are precisely the nonzero vectors in E_λ. The dimension $\dim E_\lambda$ is called the *geometric multiplicity* of λ. We always have

$$1 \leq \dim E_\lambda \leq m_\lambda \tag{3.6}$$

for every eigenvalue λ.

Recall that the matrix A is *diagonalizable* if there exists an invertible matrix P such that

$$P^{-1}AP$$

is diagonal. If D is this diagonal matrix, then we have

$$A = PDP^{-1}.$$

We say that matrices $A, B \in M_{n,n}(\mathbb{C})$ are *similar* if there exists some invertible matrix P such that $A = PBP^{-1}$. So a matrix is diagonalizable if and only if it is similar to a triangular matrix.

Theorem 3.2.1. *Suppose $A \in M_{n,n}(\mathbb{C})$. The following statements are equivalent.*

(a) *$\mathbb{C}^n$ has a basis consisting of eigenvectors of A.*

(b) *$\dim E_\lambda = m_\lambda$ for every eigenvalue λ of A.*

(c) *A is diagonalizable.*

Corollary 3.2.2. *If $A \in M_{n,n}(\mathbb{C})$ has n distinct eigenvalues, then A is diagonalizable.*

Proof. Suppose A has n distinct eigenvalues. Then, $m_\lambda = 1$ for every eigenvalue λ. It then from (3.6) that $\dim E_\lambda = m_\lambda$ for every eigenvalue λ. Hence A is diagonalizable by Theorem 3.2.1. □

Suppose A is diagonalizable, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of eigenvectors. Then the matrix

$$P = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix}$$

is invertible. Furthermore, if we define

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix},$$

where λ_i is the eigenvalue corresponding to the eigenvector $\mathbf{v}_i$, then we have $A = PDP^{-1}$. Each eigenvalue appears on the diagonal of D a number of times equal to its (algebraic or geometric) multiplicity.

Example 3.2.3. Consider the matrix

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

The characteristic polynomial is

$$c_A(x) = \det(xI - A) = \begin{vmatrix} x-3 & 0 & 0 \\ -1 & x-3 & -1 \\ 4 & 0 & x+1 \end{vmatrix} = (x-3)\begin{vmatrix} x-3 & -1 \\ 0 & x+1 \end{vmatrix} = (x-3)^2(x+1).$$

Thus, the eigenvalues are -1 and 3, with algebraic multiplicities

$$m_{-1} = 1, \quad m_3 = 2.$$

For the eigenvalue 3, we compute the corresponding eigenspace E_3 by solving the system $(A - 3I)\mathbf{x} = 0$:

$$\left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ -4 & 0 & -4 & 0 \end{array}\right] \xrightarrow{\text{row reduce}} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Thus

$$E_3 = \text{Span}\{(1, 0, -1), (0, 1, 0)\}.$$

In particular, this eigenspace is 2-dimensional, with basis $\{(1, 0, -1), (0, 1, 0)\}$.

For the eigenvalue -1, we find E_{-1} by solving $(A+I)\mathbf{x}=0$:

$$\left[\begin{array}{ccc|c} -4 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 \\ -4 & 0 & 0 & 0 \end{array}\right] \xrightarrow{\text{row reduce}} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & -1/4 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Thus

$$E_{-1} = \operatorname{Span}\{(0,1,4)\}.$$

In particular, this eigenspace is 1-dimensional, with basis $\{(0,1,4)\}$.

Since we have $\dim E_\lambda = m_\lambda$ for each eigenvalue λ, the matrix A is diagonalizable. In particular, we have $A = PDP^{-1}$, where

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ -1 & 0 & 4 \end{bmatrix} \qquad \text{and} \qquad D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Example 3.2.4. Consider the matrix

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & -3 & 0 \end{bmatrix}.$$

Then

$$c_B(x) = \det(xI - B) = \begin{vmatrix} x & 0 & 0 \\ -2 & x & 0 \\ 0 & 3 & x \end{vmatrix} = x^3.$$

Thus B has only one eigenvalue $\lambda = 0$, with algebraic multiplicity 3. To find E_0 we solve

$$\begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & -3 & 0 \end{bmatrix} \mathbf{x} = 0$$

and find the solution set

$$E_0 = \operatorname{Span}\{(0,0,1)\}.$$

So $\dim E_0 = 1 < 3 = m_0$. Thus there is no basis for $\mathbb{C}^3$ consisting of eigenvectors. Hence B is not diagonalizable.

Exercises.

Recommended exercises: Exercises in [Nic, §3.3].

3.3 Hermitian and unitary matrices

In this section we introduce hermitian and unitary matrices, and study some of their important properties. These matrices are complex analogues of symmetric and orthogonal matrices, which you saw in previous courses.

Recall that real matrices (i.e. matrices with real entries) can have complex eigenvalues that are not real. For example, consider the matrix

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

Its characteristic polynomial is

$$c_A(x) = \det(xI - A) = x^2 + 1,$$

which has roots $\pm i$. Then one can find the associated eigenvectors in the usual way to see that

$$A \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} i \\ -1 \end{bmatrix} = i \begin{bmatrix} 1 \\ i \end{bmatrix} \quad \text{and} \quad A \begin{bmatrix} 1 \\ -i \end{bmatrix} = \begin{bmatrix} -i \\ -1 \end{bmatrix} = -i \begin{bmatrix} 1 \\ -i \end{bmatrix}.$$

Thus, when considering eigenvalues, eigenvectors, and diagonalization, it makes much more sense to work over the complex numbers.

Definition 3.3.1 (Conjugate transpose)**.** The *conjugate transpose* (or *hermitian conjugate*) A^H of a complex matrix is defined by

$$A^H = (\bar{A})^T = \overline{(A^T)}.$$

Other common notations for A^H are A^* and $A^\dagger$.

Note that $A^H = A^T$ when A is real. In many ways, the conjugate transpose is the "correct" complex analogue of the transpose for real matrices, in the sense that many theorems for real matrices involving the transpose remain true for complex matrices when you replace "transpose" by "conjugate transpose". We can also rewrite the inner product (3.1) as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^H \mathbf{v}. \tag{3.7}$$

Example 3.3.2. We have

$$\begin{bmatrix} 2-i & -3 & i \\ -5i & 7-3i & 0 \end{bmatrix}^H = \begin{bmatrix} 2+i & 5i \\ -3 & 7+3i \\ -i & 0 \end{bmatrix}.$$

Proposition 3.3.3. *Suppose $A, B \in M_{m,n}(\mathbb{C})$ and $c \in \mathbb{C}$.*

(a) $(A^H)^H = A$.

(b) $(A+B)^H = A^H + B^H$.

(c) $(cA)^H = \bar{c}A^H$.

If $A \in M_{m,n}(\mathbb{C})$ and $B \in M_{n,k}(\mathbb{C})$, then

(d) $(AB)^H = B^H A^H$.

Recall that a matrix is *symmetric* if $A^T = A$. The natural complex generalization of this concept is the following.

Definition 3.3.4 (Hermitian matrix)**.** A square complex matrix A is *hermitian* if $A^H = A$, equivalently, if $\bar{A} = A^T$.

Example 3.3.5. The matrix

$$\begin{bmatrix} -2 & 2-i & 5i \\ 2+i & 1 & 4 \\ -5i & 4 & 8 \end{bmatrix}$$

is hermitian, whereas

$$\begin{bmatrix} 3 & i \\ i & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} i & 2-i \\ 2+i & 3 \end{bmatrix}$$

are not. Note that entries on the main diagonal of a hermitian matrix must be real.

Proposition 3.3.6. *A matrix $A \in M_{n,n}(\mathbb{C})$ is hermitian if and only if*

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \langle A\mathbf{x}, \mathbf{y} \rangle \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{C}^n. \tag{3.8}$$

Proof. First suppose that $A = [a_{ij}]$ is hermitian. Then, for $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, we have

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \mathbf{x}^H A\mathbf{y} = \mathbf{x}^H A^H \mathbf{y} = (A\mathbf{x})^H \mathbf{y} = \langle A\mathbf{x}, \mathbf{y} \rangle.$$

Conversely, suppose that (3.8) holds. Then, for all $1 \le i, j \le n$, we have

$$a_{ij} = \mathbf{e}_i^T A \mathbf{e}_j = \langle \mathbf{e}_i, A\mathbf{e}_j \rangle = \langle A\mathbf{e}_i, \mathbf{e}_j \rangle = (A\mathbf{e}_i)^H \mathbf{e}_j = \mathbf{e}_i^H A^H \mathbf{e}_j = \mathbf{e}_i^T A^H \mathbf{e}_j = \overline{a_{ji}}.$$

Thus A is hermitian. □

Proposition 3.3.7. *If $A \in M_{n,n}(\mathbb{C})$ is hermitian, then every eigenvalue of A is real.*

Proof. Suppose we have

$$A\mathbf{x} = \lambda \mathbf{x}, \quad \mathbf{x} \in \mathbb{C}^n,\ \mathbf{x} \neq 0,\ \lambda \in \mathbb{C}.$$

Then

$$\begin{aligned} \lambda \langle \mathbf{x}, \mathbf{x} \rangle &= \langle \mathbf{x}, \lambda \mathbf{x} \rangle && \text{(by (IP3))} \\ &= \langle \mathbf{x}, A\mathbf{x} \rangle \\ &= \mathbf{x}^H A\mathbf{x} && \text{(by (3.7))} \\ &= \mathbf{x}^H A^H \mathbf{x} && \text{(since } A \text{ is hermitian)} \\ &= (A\mathbf{x})^H \mathbf{x} \\ &= \langle A\mathbf{x}, \mathbf{x} \rangle && \text{(by (3.7))} \end{aligned}$$

$$\begin{aligned} &= \langle \lambda\mathbf{x}, \mathbf{x}\rangle \\ &= \bar{\lambda}\langle \mathbf{x}, \mathbf{x}\rangle. && \text{(by (IP2))} \end{aligned}$$

Thus we have

$$(\lambda - \bar{\lambda})\langle \mathbf{x}, \mathbf{x}\rangle = 0.$$

Since $\mathbf{x} \neq 0$, (IP5) implies that $\lambda = \bar{\lambda}$. Hence $\lambda \in \mathbb{R}$. □

Proposition 3.3.8. *If $A \in M_{n,n}(\mathbb{C})$ is hermitian, then eigenvectors of A corresponding to distinct eigenvalues are orthogonal.*

Proof. Suppose

$$A\mathbf{x} = \lambda\mathbf{x},\ A\mathbf{y} = \mu\mathbf{y}, \quad \mathbf{x}, \mathbf{y} \neq 0,\ \lambda \neq \mu.$$

Since A is hermitian, we have $\lambda, \mu \in \mathbb{R}$ by Proposition 3.3.7. Then we have

$$\begin{aligned} \lambda\langle \mathbf{x}, \mathbf{y}\rangle &= \langle \lambda\mathbf{x}, \mathbf{y}\rangle && \text{(by (IP2) and } \lambda \in \mathbb{R}) \\ &= \langle A\mathbf{x}, \mathbf{y}\rangle \\ &= \langle \mathbf{x}, A\mathbf{y}\rangle && \text{(by Proposition 3.3.6)} \\ &= \langle \mathbf{x}, \mu\mathbf{y}\rangle \\ &= \mu\langle \mathbf{x}, \mathbf{y}\rangle. && \text{(by (IP3))} \end{aligned}$$

Thus we have

$$(\lambda - \mu)\langle \mathbf{x}, \mathbf{y}\rangle = 0.$$

Since $\lambda \neq \mu$, it follows that $\langle \mathbf{x}, \mathbf{y}\rangle = 0$. □

Proposition 3.3.9. *The following conditions are equivalent for a matrix $U \in M_{n,n}(\mathbb{C})$.*

(a) *U is invertible and $U^{-1} = U^H$.*

(b) *The rows of U are orthonormal.*

(c) *The columns of U are orthonormal.*

Proof. The proof of this result is almost identical to the characterization of orthogonal matrices that you saw in previous courses. For details, see [Nic, Th. 8.2.1] and [Nic, Th. 8.7.6]. □

Definition 3.3.10 (Unitary matrix)**.** A square complex matrix U is *unitary* if $U^{-1} = U^H$.

Recall that a matrix P is *orthogonal* if $P^{-1} = P^T$. Note that a *real* matrix is unitary if and only if it is orthogonal. You should think of unitary matrices as the correct complex analogue of orthogonal matrices.

Example 3.3.11. The matrix

$$\begin{bmatrix} i & 1-i \\ 1 & 1+i \end{bmatrix}$$

has orthogonal columns. However, the columns are not orthonormal (and the rows are not orthogonal). So the matrix is not unitary. However, if we normalize the columns, we obtain the unitary matrix

$$\frac{1}{2}\begin{bmatrix} \sqrt{2}i & 1-i \\ \sqrt{2} & 1+i \end{bmatrix}.$$

You saw in previous courses that symmetric real matrices are always diagonalizable. We will see in the next section that the same is true for complex hermitian matrices. Before discussing the general theory, let's do a simple example that illustrates some of the ideas we've seen in this section.

Example 3.3.12. Consider the hermitian matrix

$$A = \begin{bmatrix} 2 & i \\ -i & 2 \end{bmatrix}.$$

Its characteristic polynomial is

$$c_A(x) = \det(xI - A) = \begin{vmatrix} x-2 & -i \\ i & x-2 \end{vmatrix} = (x-2)^2 - 1 = x^2 - 4x + 3 = (x-1)(x-3).$$

Thus, the eigenvalues are 1 and 3, which are real, as predicted by Proposition 3.3.7. The corresponding eigenvectors are

$$\begin{bmatrix} 1 \\ i \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} i \\ 1 \end{bmatrix}.$$

These are orthogonal, as predicted by Proposition 3.3.8. Each has length $\sqrt{2}$, so an orthonormal basis of eigenvectors is

$$\left\{ \tfrac{1}{\sqrt{2}}(1, i), \tfrac{1}{\sqrt{2}}(i, 1) \right\}.$$

Thus, the matrix

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

is unitary, and we have that

$$U^H A U = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

is diagonal.

Exercises.

3.3.1. Recall that, for $\theta \in \mathbb{R}$, we define the complex exponential

$$e^{i\theta} = \cos\theta + i \sin\theta.$$

Find necessary and sufficient conditions on $\alpha, \beta, \gamma, \theta \in \mathbb{R}$ for the matrix

$$\begin{bmatrix} e^{i\alpha} & e^{i\beta} \\ e^{i\gamma} & e^{i\theta} \end{bmatrix}$$

to be hermitian. Your final answer should not involve any complex exponentials or trigonometric functions.

3.3.2. Show that if A is a square matrix, then $\det(A^H) = \overline{\det A}$.

Additional recommended exercises: Exercises 8.7.12–8.7.17 in [Nic, §8.7].

3.4 The spectral theorem

In previous courses you studied the diagonalization of symmetric matrices. In particular, you learned that every real symmetric matrix was orthogonally diagonalizable. In this section we will study the complex analogues of those results.

Definition 3.4.1 (Unitarily diagonalizable matrix)**.** An matrix $A \in M_{n,n}(\mathbb{C})$ is said to be *unitarily diagonalizable* if there exists a unitary matrix U such that $U^{-1}AU = U^HAU$ is diagonal.

One of our goals in this section is to show that every hermitian matrix is unitarily diagonalizable. We first prove an important theorem which has this result as an easy consequence.

Theorem 3.4.2 (Schur's theorem)**.** *If $A \in M_{n,n}(\mathbb{C})$, then there exists a unitary matrix U such that*

$$U^HAU = T$$

is upper triangular. Moreover, the entries on the main diagonal of T are the eigenvalues of A (including multiplicities).

Proof. We prove the result by induction on n. If $n = 1$, then A is already upper triangular, and we are done (just take $U = I$). Now suppose $n > 1$, and that the theorem holds for $(n-1) \times (n-1)$ complex matrices.

Let λ_1 be an eigenvalue of A, and let $\mathbf{y}_1$ be a corresponding eigenvector with $\|\mathbf{y}_1\| = 1$. By Proposition 3.1.2, we can extend this to an orthonormal basis

$$\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$$

of $\mathbb{C}^n$. Then

$$U_1 = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{bmatrix}$$

is a unitary matrix. In block form, we have

$$U_1^HAU_1 = \begin{bmatrix} \mathbf{y}_1^H \\ \mathbf{y}_2^H \\ \vdots \\ \mathbf{y}_n^H \end{bmatrix} \begin{bmatrix} \lambda_1\mathbf{y}_1 & A\mathbf{y}_2 & \cdots & A\mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & X_1 \\ 0 & A_1 \end{bmatrix}.$$

Now, by the induction hypothesis applied to the $(n-1) \times (n-1)$ matrix A_1, there exists a unitary $(n-1) \times (n-1)$ matrix W_1 such that

$$W_1^HA_1W_1 = T_1$$

is upper triangular. Then

$$U_2 = \begin{bmatrix} 1 & 0 \\ 0 & W_1 \end{bmatrix}$$

is a unitary $n \times n$ matrix. If we let $U = U_1 U_2$, then

$$U^H = (U_1 U_2)^H = U_2^H U_1^H = U_2^{-1} U_1^{-1} = (U_1 U_2)^{-1} = U^{-1},$$

and so U is unitary. Furthermore,

$$U^H A U = U_2^H (U_1^H A U_1) U_2 = \begin{bmatrix} 1 & 0 \\ 0 & W_1^H \end{bmatrix} \begin{bmatrix} \lambda_1 & X_1 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & W_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & X_1 W_1 \\ 0 & T_1 \end{bmatrix} = T$$

is upper triangular.

Finally, since A and T are similar matrices, they have the same eigenvalues, and these eigenvalues are the diagonal entries of T since T is upper triangular. □

By Schur's theorem (Theorem 3.4.2), every matrix $A \in M_{n,n}(\mathbb{C})$ can be written in the form

$$A = UTU^H = UTU^{-1} \tag{3.9}$$

where U is unitary, T is upper triangular, and the diagonal entries of T are the eigenvalues of T. The expression (3.9) is called a *Schur decomposition* of A.

Recall that the *trace* of a square matrix $A = [a_{ij}]$ is

$$\operatorname{tr} A = a_{11} + a_{22} + \ldots + a_{nn}.$$

In other words, $\operatorname{tr} A$ is the sum of the entries of A on the main diagonal. Similar matrices have the same trace and determinant (see Exercises 3.4.1 and 3.4.2).

Corollary 3.4.3. *Suppose $A \in M_{n,n}(\mathbb{C})$, and let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of A, including multiplicities. Then*

$$\det A = \lambda_1 \lambda_2 \cdots \lambda_n \quad \text{and} \quad \operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Proof. Since the statements are clear true for triangular matrices, the corollary follows from the fact mentioned above that similar matrices have the same determinant and trace. □

Schur's theorem states that every complex square matrix can be "unitarily triangularized". However, not every complex square matrix can be unitarily diagonalized. For example, the matrix

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

cannot be unitarily diagonalized. You can see this by find the eigenvectors of A and seeing that there is no basis of eigenvectors (there is only one eigenvalue, but its corresponding eigenspace only has dimension one).

Theorem 3.4.4 (Spectral theorem)**.** *Every hermitian matrix is unitarily diagonalizable. In other words, if A is a hermitian matrix, then there exists a unitary matrix U such that $U^H A U$ is diagonal.*

Proof. Suppose A is a hermitian matrix. By Schur's Theorem (Theorem 3.4.2), there exists a unitary matrix U such that $U^H AU = T$ is upper triangular. Then we have

$$T^H = (U^H AU)^H = U^H A^H (U^H)^H = U^H AU = T.$$

Thus T is both upper and lower triangular. Hence T is diagonal. □

The terminology "spectral theorem" comes from the fact that the set of distinct eigenvalues is called the *spectrum* of a matrix. In previous courses, you learned the following real analogue of the spectral theorem.

Theorem 3.4.5 (Real spectral theorem, principal axes theorem)**.** *The following conditions are equivalent for $A \in M_{n,n}(\mathbb{R})$.*

(a) *A has an orthonormal set of n eigenvectors in $\mathbb{R}^n$.*

(b) *A is* orthogonally diagonalizable*. That is, there exists a real orthogonal matrix P such that $P^{-1}AP = P^T AP$ is diagonal.*

(c) *A is symmetric.*

A set of orthonormal eigenvectors of a symmetric matrix A is called a set of *principle axes* for A (hence the name of Theorem 3.4.5.).

Note that the principle axes theorem states that a real matrix is orthogonally diagonalizable if *and only if* the matrix is symmetric. However, the converse of the spectral theorem (Theorem 3.4.4) is false, as the following example shows.

Example 3.4.6. Consider the non-hermitian matrix

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

The characteristic polynomial is

$$c_A(x) = \det(xI - A) = x^2 + 4.$$

Thus the eigenvalues are $2i$ and $-2i$. The corresponding eigenvectors are

$$\begin{bmatrix} -1 \\ i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} i \\ -1 \end{bmatrix}.$$

These vectors are orthogonal and both have length $\sqrt{2}$. Therefore

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & i \\ i & -1 \end{bmatrix}$$

is a unitary matrix such that

$$U^H AU = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$$

is diagonal.

Why does the converse of Theorem 3.4.4 fail? Why doesn't the proof that an orthogonally diagonalizable real matrix is symmetric carry over to the complex setting? Let's recall the proof in the real case. Suppose A is orthogonally diagonalizable. Then there exists a real orthogonal matrix P (so $P^{-1} = P^T$) and a real diagonal matrix D such that $A = PDP^T$. Then we have

$$A^T = \left(PDP^T\right)^T = PD^TP^T = PDP^T = A,$$

where we used the fact that $D = D^T$ for a diagonal matrix. Hence A is symmetric. However, suppose we assume that A is a unitarily diagonalizable. Then there exists a unitary matrix U and a complex diagonal matrix D such that $A = UDU^T$. We have

$$A^H = \left(UDU^H\right)^H = UD^HU^H,$$

and here we're stuck. We won't have $D^H = D$ unless the entries of the diagonal matrix D are all real. So the argument fails. It turns out that we need to introduce a stronger condition on the matrix A.

Definition 3.4.7 (Normal matrix)**.** A matrix $N \in M_{n,n}(\mathbb{C})$ is *normal* if $NN^H = N^HN$.

Clearly every hermitian matrix is normal. Note that the matrix A in Example 3.4.6 is also normal since

$$AA^H = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix}\begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix} = A^HA.$$

Theorem 3.4.8. *A complex square matrix is unitarily diagonalizable if and only if it is normal.*

Proof. First suppose that $A \in M_{n,n}(\mathbb{C})$ is unitarily diagonalizable. So we have

$$U^HAU = D$$

for some unitary matrix U and diagonal matrix D. Since diagonal matrices commute with each other, we have $DD^H = D^HD$. Now

$$DD^H = (U^HAU)(U^HA^HU) = U^HAA^HU$$

and

$$D^HD = (U^HA^HU)(U^HAU) = U^HA^HAU.$$

Hence

$$U^H(AA^H)U = U^H(A^HA)U.$$

Multiplying on the left by U and on the right by U^H gives $AA^H = A^HA$, as desired.

Now suppose that $A \in M_{n,n}(\mathbb{C})$ is normal, so that $AA^H = A^HA$. By Schur's theorem (Theorem 3.4.2), we can write

$$U^HAU = T$$

for some unitary matrix U and upper triangular matrix T. Then T is also normal since

$$TT^H = (U^HAU)(U^HA^HU) = U^H(AA^H)U = U^H(A^HA)U = (U^HA^HU)(U^HAU) = T^HT.$$

So it is enough to show that a normal $n \times n$ upper triangular matrix is diagonal. We prove this by induction on n. The case $n = 1$ is clear, since all 1×1 matrices are diagonal. Suppose $n > 1$ and that all normal $(n-1) \times (n-1)$ upper triangular matrices are diagonal. Let $T = [t_{ij}]$ be a normal $n \times n$ upper triangular matrix. Equating the $(1,1)$-entries of TT^H and $T^H T$ gives

$$|t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2 = |t_{11}|^2.$$

It follows that

$$t_{12} = t_{13} = \cdots = t_{1n} = 0.$$

Thus, in block form, we have

$$T = \begin{bmatrix} t_{11} & 0 \\ 0 & T_{11} \end{bmatrix}$$

Then

$$T^H = \begin{bmatrix} \overline{t_{11}} & 0 \\ 0 & T_{11}^H \end{bmatrix}$$

and so we have

$$\begin{bmatrix} |t_{11}|^2 & 0 \\ 0 & T_1 T_1^H \end{bmatrix} = TT^H = T^H T = \begin{bmatrix} |t_{11}^2| & 0 \\ 0 & T_1^H T_1. \end{bmatrix}$$

Thus $T_1 T_1^H = T_1^H T_1$. By our induction hypothesis, this implies that T_1 is diagonal. Hence T is diagonal, completing the proof of the induction step. □

We conclude this section with a famous theorem about matrices. Recall that if A is a square matrix, then we can form powers A^k, $k \geq 0$, of A. (We define $A^0 = I$). Thus, we can substitute A into polynomials. For example, if

$$p(x) = 2x^3 - 3x^2 + 4x + 5 = 2x^3 - 3x^2 + 4x + 5x^0,$$

then

$$p(A) = 2A^3 - 3A^2 + 4A + 5I.$$

Theorem 3.4.9 (Cayley–Hamilton Theorem)**.** *If $A \in M_{n,n}(\mathbb{C})$, then $c_A(A) = 0$. In other words, every square matrix is a "root" of its characteristic polynomial.*

Proof. Note that, for any $k \geq 0$ and invertible matrix P, we have

$$(P^{-1}AP)^k = P^{-1}A^k P.$$

It follows that if $p(x)$ is any polynomial, then

$$p(P^{-1}AP) = P^{-1}p(A)P.$$

Thus, $p(A) = 0$ if and only if $p(P^{-1}AP) = 0$. Therefore, by Schur's theorem (Theorem 3.4.2), we may assume that A is upper triangular. Then the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of A appear on the main diagonal and we have

$$c_A(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n).$$

Therefore

$$c_A(A) = (A - \lambda_1 I)(A - \lambda_2 I) \cdots (A - \lambda_n I).$$

Each matrix $A - \lambda_i I$ is upper triangular. Observe that:

(a) $A - \lambda_1 I$ has zero first column, since the first column of A is $(\lambda_1, 0, \ldots, 0)$.

(b) Then $(A - \lambda_1 I)(A - \lambda_2 I)$ has the first two columns zero because the second column of $(A - \lambda_2 I)$ is of the form $(b, 0, \ldots, 0)$ for some $b \in \mathbb{C}$.

(c) Next $(A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I)$ has the first three columns zero because the third column of $(A - \lambda_3 I)$ is of the form $(c, d, 0, \ldots, 0)$ for some $c, d \in \mathbb{C}$.

Continuing in this manner, we see that $(A - \lambda_1 I)(A - \lambda_2 I) \cdots (A - \lambda_n I)$ has all n columns zero, and hence is the zero matrix. □

Exercises.

3.4.1. Suppose that A and B are similar square matrices. Show that $\det A = \det B$.

3.4.2. Suppose $A, B \in M_{n,n}(\mathbb{C})$.

(a) Show that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

(b) Show that if A and B are similar, then $\operatorname{tr} A = \operatorname{tr} B$.

3.4.3. (a) Suppose that $N \in M_{n,n}(\mathbb{C})$ is upper triangular with diagonal entries equal to zero. Show that, for all $j = 1, 2, \ldots, n$, we have

$$N\mathbf{e}_j \in \operatorname{Span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{j-1}\},$$

where $\mathbf{e}_i$ is the i-th standard basis vector. (When $j = 1$, we interpret the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{j-1}\}$ as the empty set. Recall that $\operatorname{Span} \varnothing = \{0\}$.)

(b) Again, suppose that $N \in M_{n,n}(\mathbb{C})$ is upper triangular with diagonal entries equal to zero. Show that $N^n = 0$.

(c) Suppose $A \in N_{n,n}(\mathbb{C})$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, with multiplicity. Show that $A = P + N$ for some $P, N \in M_{n,n}(\mathbb{C})$ satisfying $N^n = 0$ and $P = UDU^T$, where U is unitary and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. *Hint*: Use Schur's Theorem.

Additional exercises from [Nic, §8.7]: 8.7.5–8.7.9, 8.7.18–8.7.25.

3.5 Positive definite matrices

In this section we look at symmetric matrices whose eigenvalues are all positive. These matrices are important in applications including optimization, statistics, and geometry. We follow the presentation in [Nic, §8.3], except that we will consider the complex case (whereas [Nic, §8.3] works over the real numbers).

Let

$$\mathbb{R}_{>0} = \{a \in \mathbb{R} : a > 0\}$$

denote the set of positive real numbers.

Definition 3.5.1 (Positive definite)**.** A hermitian matrix $A \in M_{n,n}(\mathbb{C})$ is *positive definite* if

$$\langle \mathbf{x}, A\mathbf{x} \rangle \in \mathbb{R}_{>0} \text{ for all } \mathbf{x} \in \mathbb{C}^n, \ \mathbf{x} \neq 0.$$

By Proposition 3.3.7, we know that the eigenvalues of a hermitian matrix are real.

Proposition 3.5.2. *A hermitian matrix is positive definite if and only if all its eigenvalues λ are positive, that is, $\lambda > 0$.*

Proof. Suppose A is a hermitian matrix. By the spectral theorem (Theorem 3.4.4), there exists a unitary matrix U such that $U^H AU = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. For $\mathbf{x} \in \mathbb{C}^n$, define

$$\mathbf{y} = U^H \mathbf{x} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Then

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^H A\mathbf{x} &= \mathbf{x}^H (UDU^H)\mathbf{x} \\ &= (U^H\mathbf{x})^H DU^H\mathbf{x} = \mathbf{y}^H D\mathbf{y} = \lambda_1|y_1|^2 + \lambda_2|y_2|^2 + \cdots \lambda_n|y_n|^2. \quad (3.10) \end{aligned}$$

If every $\lambda_i > 0$, then (3.10) implies that $\langle \mathbf{x}, A\mathbf{x} \rangle > 0$ since some $y_j > 0$ (because $\mathbf{x} \neq 0$ and U is invertible). So A is positive definite.

Conversely, suppose A is positive definite. For $j \in \{1, 2, \ldots, n\}$, let $\mathbf{x} = U\mathbf{e}_j \neq 0$. Then $\mathbf{y} = \mathbf{e}_j$, and so (3.10) gives

$$\lambda_j = \langle \mathbf{x}, A\mathbf{x} \rangle > 0.$$

Hence all the eigenvalues of A are positive. □

Remark 3.5.3. A hermitian matrix is *positive semi-definite* if $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{C}^n$. Then one can show that a hermitian matrix is positive semi-definite if and only if all its eigenvalues λ are nonnegative, that is $\lambda \geq 0$. (See Exercise 3.5.1.) One can also consider *negative definite* and *negative semi-definite* matrices and they have analogous properties. However, we will focus here on positive definite matrices.

Example 3.5.4. Consider the matrix

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

Note that A is not hermitian (which is the same as symmetric, since A is real). For any $\mathbf{x} = (x_1, x_2)$, we have

$$\mathbf{x}^T A\mathbf{x} = x_1^2 + x_2^2.$$

Thus, if $\mathbf{x} \in \mathbb{R}^2$, then $\mathbf{x}^T A\mathbf{x}$ is always positive when $\mathbf{x} \neq 0$. However, if $\mathbf{x} = (1, i)$, we have

$$\mathbf{x}^H A\mathbf{x} = 2 + 2i,$$

which is not real. However, if $A \in M_{n,n}(\mathbb{R})$ is *symmetric*, then A is positive definite if and only if $\langle \mathbf{x}, A\mathbf{x} \rangle > 0$ for all $\mathbf{x} \in \mathbb{R}^n$. That is, for real symmetric matrices, it is enough to check this condition for real vectors.

Corollary 3.5.5. *If A is positive definite, then it is invertible and* $\det A \in \mathbb{R}_{>0}$.

Proof. Suppose A is positive definite. Then, by Proposition 3.5.2, the eigenvalues of A are positive real numbers. Hence, by Corollary 3.4.3, $\det A \in \mathbb{R}_{>0}$. In particular, since its determinant is nonzero, A is invertible. □

Example 3.5.6. Let's show that, for any invertible matrix $U \in M_{n,n}(\mathbb{C})$, the matrix $A = U^H U$ is positive definite. Indeed, we have

$$A^H = (U^H U)^H = U^H U = A,$$

and so A is hermitian. Also, for $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq 0$, we have

$$\mathbf{x}^H A \mathbf{x} = \mathbf{x}^H (U^H U)\mathbf{x} = (U\mathbf{x})^H (U\mathbf{x}) = \|U\mathbf{x}\|^2 > 0,$$

where the last equality follows from (IP5) and the fact that $U\mathbf{x} \neq 0$, since $\mathbf{x} \neq 0$ and U is invertible.

In fact, we will see that the converse to Example 3.5.6 is also true. Before verifying this, we discuss another important concept.

Definition 3.5.7 (Principal submatrices)**.** If $A \in M_{n,n}(\mathbb{C})$, let ${}^{(r)}A$ denote the $r \times r$ submatrix in the upper-left corner of A; that is, ${}^{(r)}A$ is the matrix obtained from A by deleting the last $n - r$ rows and columns. The matrices ${}^{(1)}A$, ${}^{(2)}A$, …, ${}^{(n)}A = A$ are called the *principal submatrices* of A.

Example 3.5.8. If

$$A = \begin{bmatrix} 5 & 7 & 2-i \\ 6i & 0 & -3i \\ 2+9i & -3 & 1 \end{bmatrix},$$

then

$${}^{(1)}A = \begin{bmatrix} 5 \end{bmatrix}, \quad {}^{(2)}A = \begin{bmatrix} 5 & 7 \\ 6i & 0 \end{bmatrix}, \quad {}^{(3)}A = A.$$

Lemma 3.5.9. *If $A \in M_{n,n}(\mathbb{C})$ is positive definite, then so is each principal matrix ${}^{(r)}A$ for $r = 1, 2, \ldots, n$.*

Proof. Write

$$A = \begin{bmatrix} {}^{(r)}A & P \\ Q & R \end{bmatrix}$$

in block form. First note that

$$\begin{bmatrix} {}^{(r)}A & P \\ Q & R \end{bmatrix} = A = A^H = \begin{bmatrix} \left({}^{(r)}A\right)^H & Q^H \\ P^H & R^H \end{bmatrix}.$$

Hence ${}^{(r)}A = \left({}^{(r)}A\right)^H$, and so A is hermitian.

Now let $\mathbf{y} \in \mathbb{C}^r$, $\mathbf{y} \neq 0$. Define

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \in \mathbb{C}^n.$$

Then $\mathbf{x} \neq 0$ and so, since A is positive definite, we have

$$\mathbb{R}_{>0} \ni \mathbf{x}^H A \mathbf{x} = \begin{bmatrix} \mathbf{y}^H & 0 \end{bmatrix} \begin{bmatrix} {}^{(r)}A & P \\ Q & R \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = \mathbf{y}^H \left({}^{(r)}A \right) \mathbf{y}.$$

Thus ${}^{(r)}A$ is positive definite. □

We can now prove a theorem that includes the converse to Example 3.5.6.

Theorem 3.5.10. *The following conditions are equivalent for any hermitian $A \in M_{n,n}(\mathbb{C})$.*

(a) *A is positive definite.*

(b) $\det \left({}^{(r)}A \right) \in \mathbb{R}_{>0}$ *for each $r = 1, 2, \ldots, n$.*

(c) *$A = U^H U$ for some upper triangular matrix U with positive real entries on the diagonal.*

Furthermore, the factorization in (c) is unique, called the Cholesky factorization *of A.*

Proof. We already saw that (c) $\implies$ (a) in Example 3.5.6. Also, (a) $\implies$ (b) by Lemma 3.5.9 and Corollary 3.5.5. So it remains to show (b) $\implies$ (c).

Assume that (b) is true. We prove that (c) holds by induction on n. If $n = 1$, then $A = \begin{bmatrix} a \end{bmatrix}$, where $a \in \mathbb{R}_{>0}$ by (b). So we can take $U = \begin{bmatrix} \sqrt{a} \end{bmatrix}$.

Now suppose $n > 1$ and that the result holds for matrices of size $(n-1) \times (n-1)$. Define

$$B = {}^{(n-1)}A.$$

Then B is hermitian and satisfies (b). Hence, by Lemma 3.5.9 and our induction hypothesis, we have

$$B = U^H U$$

for some upper triangular matrix $U \in M_{n-1,n-1}(\mathbb{C})$ with positive real entries on the main diagonal. Since A is hermitian, it has block form

$$A = \begin{bmatrix} B & \mathbf{p} \\ \mathbf{p}^H & b \end{bmatrix}, \quad \mathbf{p} \in \mathbb{C}^{n-1}, \ b \in \mathbb{R}.$$

If we define

$$\mathbf{x} = (U^H)^{-1} \mathbf{p} \quad \text{and} \quad c = b - \mathbf{x}^H \mathbf{x},$$

then block multiplication gives

$$A = \begin{bmatrix} U^H U & \mathbf{p} \\ \mathbf{p}^H & b \end{bmatrix} = \begin{bmatrix} U^H & 0 \\ \mathbf{x}^H & 1 \end{bmatrix} \begin{bmatrix} U & \mathbf{x} \\ 0 & c \end{bmatrix}. \tag{3.11}$$

Taking determinants and using (1.2) gives

$$\det A = (\det U^H)(\det U)c = c \left| \det U \right|^2.$$

(Here we use Exercise 3.3.2.) Since $\det A > 0$ by (b), it follows that $c > 0$. Thus, the factorization (3.11) can be modified to give

$$A = \begin{bmatrix} U^H & 0 \\ \mathbf{x}^H & \sqrt{c} \end{bmatrix} \begin{bmatrix} U & \mathbf{x} \\ 0 & \sqrt{c} \end{bmatrix}.$$

Since U is upper triangular with positive real entries on the main diagonal, this proves the induction step.

It remains to prove the uniqueness assertion in the statement of the theorem. Suppose that

$$A = U^H U = U_1^H U_1$$

are two Cholesky factorizations. Then, by Lemma 1.6.2,

$$D = U U_1^{-1} = (U^H)^{-1} U_1^H \tag{3.12}$$

is both upper triangular (since U and U_1 are) and lower triangular (since U^H and U_1^H are). Thus D is a diagonal matrix. It follows from (3.12) that

$$U = DU_1 \quad \text{and} \quad U_1 = D^H U,$$

and so

$$U = DU_1 = DD^H U.$$

Since U is invertible, this implies that $DD^H = I$. Because the diagonal entries of D are positive real numbers (since this is true for U and U_1), it follows that $D = I$. Thus $U = U_1$, as desired. □

Remark 3.5.11. (a) If the *real* matrix $A \in M_{n,n}(\mathbb{R})$ is symmetric (hence also hermitian), then the matrix U appearing in the Cholesky factorization $A = U^H U$ also has real entries, and so $A = U^T U$. See [Nic, Th. 8.3.3].

(b) Positive semi-definite matrices also have Cholesky factorizations, as long as we allow the diagonal matrices of U to be zero. However, the factorization is no longer unique in general.

Theorem 3.5.10 tells us that every positive definite matrix has a Cholesky factorization. But how do we *find* the Cholesky factorization?

Algorithm 3.5.12 (Algorithm for the Cholesky factorization)**.** *If A is a positive definite matrix, then the Cholesky factorization $A = U^H U$ can be found as follows:*

(a) *Transform A to an upper triangular matrix U_1 with positive real diagonal entries using row operations, each of which adds a multiple of a row to a lower row.*

(b) *Obtain U from U_1 by dividing each row of U_1 by the square root of the diagonal entry in that row.*

The key is that step (a) is *possible* for any positive definite matrix A. Let's do an example before proving Algorithm 3.5.12.

Example 3.5.13. Consider the hermitian matrix

$$A = \begin{bmatrix} 2 & i & -3 \\ -i & 5 & 2i \\ -3 & -2i & 10 \end{bmatrix}.$$

We can compute

$$\det {}^{(1)}A = 2 > 0, \quad \det {}^{(2)}A = 11 > 0, \quad \det {}^{(3)}A = \det A = 49 > 0.$$

Thus, by Theorem 3.5.10, A is positive definite and has a unique Cholesky factorization. We carry out step (a) of Algorithm 3.5.12 as follows:

$$A = \begin{bmatrix} 2 & i & -3 \\ -i & 5 & 2i \\ -3 & -2i & 10 \end{bmatrix} \xrightarrow[R_3+\frac{3}{2}R_1]{R_2+\frac{i}{2}R_1} \begin{bmatrix} 2 & i & -3 \\ 0 & 9/2 & i/2 \\ 0 & -i/2 & 11/2 \end{bmatrix} \xrightarrow{R_3+\frac{i}{9}R_2} \begin{bmatrix} 2 & i & -3 \\ 0 & 9/2 & i/2 \\ 0 & 0 & 49/9 \end{bmatrix} = U_1.$$

Now we carry out step (b) to obtain

$$U = \begin{bmatrix} \sqrt{2} & \frac{i}{\sqrt{2}} & \frac{-3}{\sqrt{2}} \\ 0 & \frac{3}{\sqrt{2}} & \frac{i\sqrt{2}}{6} \\ 0 & 0 & \frac{\sqrt{49}}{3} \end{bmatrix}.$$

You can then check that $A = U^H U$.

Proof of Algorithm 3.5.12. Suppose A is positive definite, and let $A = U^H U$ be the Cholesky factorization. Let $D = \operatorname{diag}(d_1, \ldots, d_n)$ be the common diagonal of U and U^H. (So the d_i are positive real numbers.) Then $U^H D^{-1}$ is *lower unitriangular* (lower triangular with ones on the diagonal). Thus $L = (U^H D^{-1})^{-1}$ is also lower unitriangular. Therefore we can write

$$L = E_r \cdots E_2 E_1 I_n,$$

where each E_i is an elementary matrix corresponding to a row operation that adds a multiple of one row to a lower row (we modify columns right to left). Then we have

$$E_r \cdots E_2 E_1 A = LA = \left(D(U^H)^{-1}\right)\left(U^H U\right) = DU$$

is upper triangular with positive real entries on the diagonal. This proves that step (a) of the algorithm is possible.

Now consider step (b). We have already shown that we can find a lower unitriangular matrix L_1 and an invertible upper triangular matrix U_1, with positive real entries on the diagonal, such that $L_1 A = U_1$. (In the notation above, $L_1 = E_r \cdots E_1$ and $U_1 = DU$.) Since A is hermitian, we have

$$L_1 U_1^H = L_1 (L_1 A)^H = L_1 A^H L_1^H = L_1 A L_1^H = U_1 L_1^H. \tag{3.13}$$

Let $D_1 = \operatorname{diag}(d_1, \ldots, d_n)$ denote diagonal matrix with the same diagonal entries as U_1. Then (3.13) implies that

$$L_1 \left(U_1^H D_1^{-1}\right) = U_1 L_1^H D_1^{-1}.$$

This is both upper triangular (since $U_1 L_1^H D_1^{-1}$ is) and lower unitriangular (since $L_1 \left(U_1^H D_1^{-1}\right)$ is), and so must equal I_n. Thus

$$U_1^H D_1^{-1} = L_1^{-1}.$$

Now let

$$D_2 = \operatorname{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right),$$

so that $D_2^2 = D_1$. If we define $U = D_2^{-1}U_1$, then

$$U^H U = \left(U_1^H D_2^{-1}\right)\left(D_2^{-1}U_1\right) = U_1^H \left(D_2^2\right)^{-1} U_1 = \left(U_1^H D_1^{-1}\right) U_1 = L_1^{-1}U_1 = A.$$

Since $U = D_2^{-1}U_1$ is the matrix obtained from U_1 by dividing each row by the square root of its diagonal entry, this completes the proof of step (b). □

Suppose we have a linear system

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

where A is a hermitian (e.g. real symmetric) matrix. Then we can find the Cholesky decomposition $A = U^H U$ and consider the linear system

$$U^H U\mathbf{x} = \mathbf{b}.$$

As with the LU decomposition, we can first solve $U^H\mathbf{y} = \mathbf{b}$ by forward substitution, and then solve $U\mathbf{x} = \mathbf{y}$ by back substitution. For linear systems that can be put in symmetric form, using the Cholesky decomposition is roughly twice as efficient as using the LU decomposition.

Exercises.

3.5.1. Show that a hermitian matrix is positive semi-definite if and only if all its eigenvalues λ are nonnegative, that is, $\lambda \geq 0$.

Additional recommended exercises: [Nic, §8.3].

3.6 QR factorization

Unitary matrices are very easy to invert, since the conjugate transpose is the inverse. Thus, it is useful to factor an arbitrary matrix as a product of a unitary matrix and a triangular matrix (which we've seen are also nice in many ways). We tackle this problem in this section. A good reference for this material is [Nic, §8.4]. (However, see Remark 3.6.2.)

Definition 3.6.1 (QR factorization)**.** A *QR factorization* of $A \in M_{m,n}(\mathbb{C})$, $m \geq n$, is a factorization $A = QR$, where Q is an $m \times m$ unitary matrix and R is an $m \times n$ upper triangular matrix whose entries on the main diagonal are nonnegative real numbers.

Note that the bottom $m - n$ rows of an $m \times n$ upper triangular matrix (with $m \geq n$) are zero rows. Thus, we can write a QR factorization $A = QR$ in block form

$$A = QR = Q\begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$$

where R_1 is an $n \times n$ upper triangular matrix whose entries on the main diagonal are nonnegative real numbers, 0 is the $(m-n) \times n$ zero matrix, Q_1 is $m \times n$, Q_2 is $m \times (m-n)$, and Q_1, Q_2 both have orthonormal columns. The factorization $A = Q_1 R_1$ is called a *thin QR factorization* or *reduced QR factorization* of A.

Remark 3.6.2. You may find different definitions of QR factorization in other references. In particular, some references, including [Nic, §8.6], refer to the reduced QR factorization as a QR factorization (without using the word "reduced"). In addition, [Nic, §8.6] imposes the condition that the columns of A are linearly independent. This is not needed for the existence result we will prove below (Theorem 3.6.3), but it will be needed for uniqueness (Theorem 3.6.6). When consulting other references, be sure to look at which definition they are using to avoid confusion.

Note that, given a reduced QR factorization, one can easily obtain a QR factorization by extending the columns of Q_1 to an orthonormal basis, and defining Q_2 to be the matrix whose columns are the additional vectors in this basis. It follows that, when $m > n$, the QR factorization is not unique. (Once can extend to an orthonormal basis in more than one way.) However the *reduced* QR factorization has some chance of being unique. (See Theorem 3.6.6.)

The power of the QR factorization comes from the fact that there are computer algorithms that can compute it with good control over round-off error. Finding the QR factorization involves the Gram–Schmidt algorithm.

Recall that a matrix $A \in M_{m,n}(\mathbb{C})$ has linearly independent columns if and only if its rank is n, which can only occur if A is tall or square (i.e. $m \geq n$).

Theorem 3.6.3. *Every tall or square matrix has a QR factorization.*

Proof. We will prove the theorem under the additional assumption that A has full rank (i.e. $\operatorname{rank} A = n$), and then make some remarks about how one can modify the proof to work without this assumption. We show that A has a reduced QR factorization, from which it follows (as discussed above) that A has a QR factorization.

Suppose

$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix} \in M_{m,n}(\mathbb{C})$$

with linearly independent columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. We can use the Gram-Schmidt algorithm to obtain an orthogonal set $\mathbf{u}_1, \ldots, \mathbf{u}_n$ spanning the column space of A. Namely, we set $\mathbf{u}_1 = \mathbf{a}_1$ and

$$\mathbf{u}_k = \mathbf{a}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{u}_j, \mathbf{a}_k \rangle}{\|\mathbf{u}_j^2\|} \mathbf{u}_j \quad \text{for } k = 2, 3, \ldots, n. \tag{3.14}$$

If we define

$$\mathbf{q}_k = \frac{1}{\|\mathbf{u}_k\|} \mathbf{u}_k \quad \text{for each } k = 1, 2, \ldots, n,$$

then the $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are orthonormal and (3.14) becomes

$$\|\mathbf{u}_k\| \mathbf{q}_k = \mathbf{a}_k - \sum_{j=1}^{k-1} \langle \mathbf{q}_j, \mathbf{a}_k \rangle \mathbf{q}_j. \tag{3.15}$$

Using (3.15), we can express each $\mathbf{a}_k$ as a linear combination of the $\mathbf{q}_j$:

$$\begin{aligned}
\mathbf{a}_1 &= \|\mathbf{u}_1\|\mathbf{q}_1,\\
\mathbf{a}_2 &= \langle \mathbf{q}_1, \mathbf{a}_2\rangle \mathbf{q}_1 + \|\mathbf{u}_2\|\mathbf{q}_2,\\
\mathbf{a}_3 &= \langle \mathbf{q}_1, \mathbf{a}_3\rangle \mathbf{q}_1 + \langle \mathbf{q}_2, \mathbf{a}_3\rangle \mathbf{q}_2 + \|\mathbf{u}_3\|\mathbf{q}_3\\
&\vdots\\
\mathbf{a}_n &= \langle \mathbf{q}_1, \mathbf{a}_n\rangle \mathbf{q}_1 + \langle \mathbf{q}_2, \mathbf{a}_n\rangle \mathbf{q}_2 + \ldots + \langle \mathbf{q}_{n-1}, \mathbf{a}_n\rangle \mathbf{q}_{n-1} + \|\mathbf{u}_n\|\mathbf{q}_n.
\end{aligned}$$

Writing these equations in matrix form gives us the factorization we're looking for:

$$\begin{aligned}
A &= \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \cdots & \mathbf{a}_n\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_n\end{bmatrix}
\begin{bmatrix}
\|\mathbf{u}_1\| & \langle \mathbf{q}_1, \mathbf{a}_2\rangle & \langle \mathbf{q}_1, \mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_1, \mathbf{a}_n\rangle\\
0 & \|\mathbf{u}_2\| & \langle \mathbf{q}_2, \mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_2, \mathbf{a}_n\rangle\\
0 & 0 & \|\mathbf{u}_3\| & \cdots & \langle \mathbf{q}_3, \mathbf{u}_n\rangle\\
0 & 0 & 0 & \cdots & \|\mathbf{u}_n\|
\end{bmatrix}.
\end{aligned}$$

What do we do if $\operatorname{rank} A < n$? In this case, the columns of A are linearly dependent, so some of the columns are in the span of the previous ones. Suppose that $\mathbf{a}_k$ is the first such column. Then, as noted in Remark 3.1.7, we get $\mathbf{u}_k = 0$. We can fix the proof as follows: Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be an orthonormal basis of $(\operatorname{col} A)^\perp$. Then, if we obtain $\mathbf{u}_k = 0$ in the Gram–Schmidt algorithm above, we let $\mathbf{q}_k$ be one of the $\mathbf{v}_i$, using each $\mathbf{v}_i$ exactly once. We then continue the proof as above, and the matrix R will have a row of zeros in the k-th row for each k that gave $\mathbf{u}_k = 0$ in the Gram–Schmidt algorithm. □

Remark 3.6.4. Note that for a tall or square *real* matrix, the proof of Theorem 3.6.3 shows us that we can find a QR factorization with Q and R *real* matrices.

Example 3.6.5. Let's find a QR factorization of

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

We first find a reduced QR factorization. We have

$$\mathbf{a}_1 = (1, 0, 1) \quad \text{and} \quad \mathbf{a}_2 = (2, 1, 0).$$

Thus

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{a}_1,\\
\mathbf{u}_2 &= \mathbf{a}_2 - \frac{\langle \mathbf{u}_1, \mathbf{a}_2\rangle}{\|\mathbf{u}_1\|^2}\mathbf{u}_1 = (2, 1, 0) - \frac{2}{2}(1, 0, 1) = (1, 1, -1).
\end{aligned}$$

So we have

$$\mathbf{q}_1 = \frac{1}{\|\mathbf{u}_1\|}\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 0, 1),$$

$$\mathbf{q}_2 = \frac{1}{\|\mathbf{u}_2\|}\mathbf{u}_2 = \frac{1}{\sqrt{3}}(1,1,-1).$$

We define

$$Q_1 = \begin{bmatrix}\mathbf{q}_1 & \mathbf{q}_2\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}}\end{bmatrix} \quad \text{and} \quad R_1 = \begin{bmatrix}\|\mathbf{u}_1\| & \langle \mathbf{q}_1, \mathbf{a}_2\rangle \\ 0 & \|\mathbf{u}_2\|\end{bmatrix} = \begin{bmatrix}\sqrt{2} & \sqrt{2} \\ 0 & \sqrt{3}\end{bmatrix}.$$

We can then verify that $A = Q_1R_1$. To get a QR factorization, we need to complete $\{\mathbf{q}_1, \mathbf{q}_2\}$ to an orthonormal basis. So we need to find a vector orthogonal to both $\mathbf{q}_1$ and $\mathbf{q}_2$ (equivalently, to their multiples $\mathbf{u}_1$ and $\mathbf{u}_2$). A vector $\mathbf{u}_3 = (x, y, z)$ is orthogonal to both when

$$0 = \langle \mathbf{u}_1, (x,y,z)\rangle = x + z \quad \text{and} \quad 0 = \langle \mathbf{u}_2, (x,y,z)\rangle = x + y - z.$$

Solving this system, we see that $\mathbf{u}_3 = (1, -2, -1)$ is orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$. So we define

$$\mathbf{q}_3 = \frac{1}{\|\mathbf{u}_3\|}\mathbf{u}_3 = \frac{1}{\sqrt{6}}(1,-2,-1).$$

Then, setting

$$Q = \begin{bmatrix}\mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}}\end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix}\sqrt{2} & \sqrt{2} \\ 0 & \sqrt{3} \\ 0 & 0\end{bmatrix},$$

we have a QR factorization $A = QR$.

Now that we know that QR factorizations exist, what about uniqueness? As noted earlier, here we should focus on *reduced* QR factorizations.

Theorem 3.6.6. *Every tall or square matrix A with linearly independent columns has a* unique *reduced QR factorization $A = QR$. Furthermore, the matrix R is invertible.*

Proof. Suppose $A \in M_{m,n}(\mathbb{C})$, $m \geq n$, has linearly independent columns. We know from Theorem 3.6.3 that A has a QR factorization $A = QR$. Furthermore, since Q is invertible, we have

$$\operatorname{rank} R = \operatorname{rank}(QR) = \operatorname{rank} A = n.$$

Since R is $m \times n$ and $m \geq n$, the columns of R are also linearly independent. It follows that the entries on the main diagonal of R must be nonzero, and hence positive (since they are nonnegative real numbers). This also holds for the upper triangular matrix appearing in a *reduced* QR factorization.

Now suppose

$$A = QR \quad \text{and} \quad A = Q_1R_1$$

are two reduced QR factorizations of A. By the above, the entries on the main diagonal of R and R_1 are positive. Since R and R_1 are square and upper triangular, this implies that

they are invertible. We wish to show that $Q = Q_1$ and $R = R_1$. Label the columns of Q and Q_1:

$$Q = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix} \quad \text{and} \quad Q_1 = \begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_n \end{bmatrix}.$$

Since Q and Q_1 have orthonormal columns, we have

$$Q^H Q = I_n = Q_1^H Q_1.$$

(Note that Q and Q_1 are *not* unitary matrices unless they are square. If they are not square, then QQ^H and $Q_1 Q_1^H$ are not defined. Recall our discussion of one-sided inverses in Section 1.5.) Therefore, the equality $QR = Q_1 R_1$ implies $Q_1^H Q = R_1 R^{-1}$. Let

$$[t_{ij}] = Q_1^H Q = R_1 R^{-1}. \tag{3.16}$$

Since R and R_1 are upper triangular with positive real diagonal elements, we have

$$t_{ii} \in \mathbb{R}_{>0} \quad \text{and} \quad t_{ij} = 0 \text{ for } i > j.$$

On the other hand, the (i, j)-entry of $Q_1^H Q$ is $\mathbf{d}_i^H \mathbf{c}_j$. So we have

$$\langle \mathbf{d}_i, \mathbf{c}_j \rangle = \mathbf{d}_i^H \mathbf{c}_j = t_{ij} \quad \text{for all } i, j.$$

Since $Q = Q_1(R_1 R^{-1})$, each $\mathbf{c}_j$ is in $\text{Span}\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$. Thus, by (3.3), we have

$$\mathbf{c}_j = \sum_{i=1}^{n} \langle \mathbf{d}_i, \mathbf{c}_j \rangle \mathbf{d}_i = \sum_{i=1}^{j} t_{ij} \mathbf{d}_i,$$

since $\langle \mathbf{d}_i, \mathbf{c}_j \rangle = t_{ij} = 0$ for $i > j$. Writing out these equations explicitly, we have:

$$\begin{aligned} \mathbf{c}_1 &= t_{11}\mathbf{d}_1, \\ \mathbf{c}_2 &= t_{12}\mathbf{d}_1 + t_{22}\mathbf{d}_2, \\ \mathbf{c}_3 &= t_{13}\mathbf{d}_1 + t_{23}\mathbf{d}_2 + t_{33}\mathbf{d}_3, \\ \mathbf{c}_4 &= t_{14}\mathbf{d}_1 + t_{24}\mathbf{d}_2 + t_{34}\mathbf{d}_3 + t_{44}\mathbf{d}_4, \\ &\vdots \end{aligned} \tag{3.17}$$

The first equation gives

$$1 = \|\mathbf{c}_1\| = \|t_{11}\mathbf{d}_1\| = |t_{11}| \, \|\mathbf{d}_1\| = t_{11}, \tag{3.18}$$

since $t_{11} \in \mathbb{R}_{>0}$. Thus $\mathbf{c}_1 = \mathbf{d}_1$. Then we have

$$t_{12} = \langle \mathbf{d}_1, \mathbf{c}_2 \rangle = \langle \mathbf{c}_1, \mathbf{c}_2 \rangle = 0,$$

and so the second equation in (3.17) gives

$$\mathbf{c}_2 = t_{22}\mathbf{d}_2.$$

As in (3.18), this implies that $\mathbf{c}_2 = \mathbf{d}_2$. Then $t_{13} = 0$ and $t_{23} = 0$ follows in the same way. Continuing in this way, we conclude that $\mathbf{c}_i = \mathbf{d}_i$ for all i. Thus, $Q = Q_1$ and, by (3.16), $R = R_1$, as desired. □

So far our results are all about tall or square matrices. What about wide matrices?

Corollary 3.6.7. *Every wide or square matrix A factors as $A = LP$, where P is unitary matrix, and L is a lower triangular matrix whose entries on the diagonal are nonnegative real numbers. It also factors as $A = L_1P_1$, where L_1 is a square lower triangular matrix whose entries on the main diagonal are nonnegative real numbers and P_1 has orthonormal rows. If A has linearly independent rows, then the second factorization is unique and the diagonal entries of L_1 are positive.*

Proof. We apply Theorems 3.6.3 and 3.6.6 to A^T. □

Corollary 3.6.8. *Every invertible (hence square) matrix A has unique factorizations $A = QR$ and $Q = LP$ where Q and P are unitary, R is upper triangular with positive real diagonal entries, and L is lower triangular with positive real diagonal entries.*

QR factorizations have a number of applications. We learned in Proposition 1.5.15 that a matrix has linearly independent columns if and only if it has a left inverse if and only if A^HA is invertible. Furthermore, if A is left-invertible, then $(A^HA)^{-1}A^H$ is a left inverse of A. (We worked over $\mathbb{R}$ in Section 1.5.4, but the results are true over $\mathbb{C}$ as long as we replace the transpose by the conjugate transpose.) So, in this situation, it is useful to compute $(A^HA)^{-1}$. Here Theorem 3.6.6 guarantees us that we have a reduced QR factorization $A = QR$ with R invertible. (And we have an algorithm for finding this factorization!) Since Q has orthonormal columns, we have $Q^HQ = I$. Thus

$$A^HA = R^HQ^HQR = R^HR,$$

and so

$$(A^HA)^{-1} = R^{-1}\left(R^{-1}\right)^H.$$

Since R is upper triangular, it is easy to find its inverse. So this gives us an efficient method of finding left inverses.

Students who took MAT 2342 learned about finding best approximations to (possibly inconsistent) linear systems. (See [Nic, §5.6].) In particular, if $A\mathbf{x} = \mathbf{b}$ is a linear system, then any solution $\mathbf{z}$ to the *normal equations*

$$(A^HA)\mathbf{z} = A^H\mathbf{b} \tag{3.19}$$

is a best approximation to a solution to $A\mathbf{x} = \mathbf{b}$ in the sense that $\|\mathbf{b} - A\mathbf{z}\|$ is the minimum value of $\|\mathbf{b} - A\mathbf{x}\|$ for $\mathbf{x} \in \mathbb{C}^n$. (You probably worked over $\mathbb{R}$ in MAT 2342.) As noted above, A has linearly independent columns if and only if A^HA is invertible. In this case, there is a unique solution $\mathbf{z}$ to (3.19), and it is given by

$$\mathbf{z} = (A^HA)^{-1}A^H\mathbf{b}.$$

As noted above, $(A^HA)^{-1}A^H$ is a left inverse to A. We saw in Section 1.5.1 that *if* $A\mathbf{x} = \mathbf{b}$ has a solution, then it must be $(A^HA)^{-1}A^H\mathbf{b}$. What we're saying here is that, even there is no solution, $(A^HA)^{-1}A^H\mathbf{b}$ is the best *approximation* to a solution.

Exercises.

Recommended exercises: Exercises in [Nic, §8.4]. Keep in mind the different definition of QR factorization used in [Nic] (see [Nic, Def. 8.6]).

3.7 Computing eigenvalues

Until now, you've found eigenvalues by finding the roots of the characteristic polynomial. In practice, this is almost never done. For large matrices, finding these roots is difficult. Instead, iterative methods for estimating eigenvalues are much better. In this section, we will explore such methods. A reference for some of this material is [Nic, §8.5].

3.7.1 The power method

Throughout this subsection, we suppose that $A \in M_{n,n}(\mathbb{C})$. We will use $\|\cdot\|$ to denote the 2-norm on $\mathbb{C}^n$.

An eigenvalue λ for A is called a *dominant eigenvalue* if λ has algebraic multiplicity one, and

$$|\lambda| > |\mu| \text{ for all eigenvalues } \mu \neq \lambda.$$

Any eigenvector corresponding to a dominant eigenvalue is called a *dominant eigenvector* of A.

Suppose A is diagonalizable, and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, with multiplicity, and suppose that

$$|\lambda_1| \leq \cdots \leq |\lambda_{n-1}| < \lambda_n.$$

(We are implicitly saying that λ_n is real here.) In particular, λ_n is a dominant eigenvalue. Fix a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of unit eigenvectors of A, so that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for each $i = 1, \ldots, n$.

Let start with some unit vector $\mathbf{x}_0 \in \mathbb{C}^n$ and recursively define a sequence of vectors $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ and positive real numbers $\|\mathbf{y}_1\|, \|\mathbf{y}_2\|, \|\mathbf{y}_3\|, \ldots$ by

$$\mathbf{y}_k = A\mathbf{x}_{k-1}, \quad \mathbf{x}_k = \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|}, \qquad k \geq 1. \tag{3.20}$$

The *power method* uses the fact that, under some mild assumptions, the sequence $\|\mathbf{y}_1\|, \|\mathbf{y}_2\|, \ldots$ converges to λ_n, and the sequence $\mathbf{x}_1, \mathbf{x}_2, \cdots$ converges to a corresponding eigenvector.

We should be precise about what we mean by convergence here. You learned about convergence of sequences of real numbers in calculus. What about convergence of vectors? We say that a sequence of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$ converges to a vector $\mathbf{v}$ if

$$\lim_{k \to \infty} \|\mathbf{x}_k - \mathbf{v}\| = 0.$$

This is equivalent to the components of the vectors $\mathbf{x}_k$ converging (in the sense of sequences of scalars) to the components of $\mathbf{v}$.

Theorem 3.7.1. *With the notation and assumptions from above, suppose that the initial vector* $\mathbf{x}_0$ *is of the form*

$$\mathbf{x}_0 = \sum_{i=1}^{n} a_i \mathbf{v}_i \quad \text{with } a_n \neq 0.$$

Then the power method converges, that is

$$\lim_{k\to\infty} \|\mathbf{y}_k\| = \lambda_n \quad \text{and} \quad \lim_{k\to\infty} \mathbf{x}_k = \frac{a_n}{|a_n|}\mathbf{v}_n.$$

In particular, the $\mathbf{x}_k$ *converge to an eigenvector of eigenvalue* λ_n.

Proof. Note that, by definition, $\mathbf{x}_k$ is a unit vector that is a positive real multiple of

$$A^k\mathbf{x}_0 = A^k\left(\sum_{i=1}^{n} a_i\mathbf{v}_i\right) = \sum_{i=1}^{n} a_i A^k\mathbf{v}_i = \sum_{i=1}^{n} a_i\lambda_i^k\mathbf{v}_i.$$

Thus

$$\mathbf{x}_k = \frac{\sum_{i=1}^{n} a_i\lambda_i^k\mathbf{v}_i}{\|\sum_{i=1}^{n} a_i\lambda_i^k\mathbf{v}_i\|} = \frac{a_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^k\mathbf{v}_i}{\left\|a_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^k\mathbf{v}_i\right\|}.$$

(Note that we could divide by λ_n since $\lambda_n \neq 0$. Also, since $a_n \neq 0$, the norm in the denominator is nonzero.) Since $|\lambda_i| < \lambda_n$ for $i \neq n$, we have

$$\left(\frac{\lambda_i}{\lambda_n}\right)^k \to 0 \quad \text{as} \quad k \to \infty.$$

Hence, as $k \to \infty$, we have

$$\mathbf{x}_k \to \frac{a_n\mathbf{v}_n}{\|a_n\mathbf{v}_n\|} = \frac{a_n}{\|a_n\|}\mathbf{v}_n.$$

Similarly,

$$\begin{aligned}\|\mathbf{y}_{k+1}\| = \|A\mathbf{x}_k\| &= \frac{\left\|a_n\lambda_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^k\lambda_i\mathbf{v}_i\right\|}{\left\|a_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^k\mathbf{v}_i\right\|} \\ &= \lambda_n\frac{\left\|a_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^{k+1}\mathbf{v}_i\right\|}{\left\|a_n\mathbf{v}_n + \sum_{i=1}^{n-1} a_i\left(\frac{\lambda_i}{\lambda_n}\right)^k\mathbf{v}_i\right\|} \to \lambda_n. \quad \square\end{aligned}$$

Remarks 3.7.2. (a) It is crucial that the largest eigenvalue be real. On the other hand, the assumption that it is positive can be avoided since, if it is negative, we can apply Theorem 3.7.1 to $-A$.

(b) If there are several eigenvalues with maximum norm, then the sequences $\|\mathbf{y}_k\|$ and $\mathbf{x}_k$ will not converge in general. On the other hand, if λ_n has multiplicity greater than one, but is the unique eigenvalue with maximum norm, then the sequence $\|\mathbf{y}_k\|$ will always converge to λ_n, but the sequence $\mathbf{x}_k$ may not converge.

(c) If you choose an initial vector $\mathbf{x}_0$ at random, it is *very* unlikely that you will choose one with $a_n = 0$. (This would be the same as choosing a random real number and ending up with the real number 0.) Thus, the condition in Theorem 3.7.1 that $a_n \neq 0$ is not a serious obstacle in practice.

(d) It is possible to compute the smallest eigenvalue (in norm) by applying the power method to A^{-1}. This is called the *inverse power method*. It is computationally more involved, since one must solve a linear system at each iteration.

Example 3.7.3. Consider the matrix

$$A = \begin{bmatrix} 3 & 3 \\ 5 & 1 \end{bmatrix}.$$

We leave it as an exercise to verify that the eigenvalues of A are -2 and 6. So 6 is a dominant eigenvalue. Let

$$\mathbf{x}_0 = (1, 0)$$

be our initial vector. Then we compute

$$\mathbf{y}_1 = A\mathbf{x}_0 = \begin{bmatrix} 3 & 3 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \qquad \|\mathbf{y}_1\| \approx 5.830951, \quad \mathbf{x}_1 = \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|} \approx \begin{bmatrix} 0.514496 \\ 0.857493 \end{bmatrix},$$

$$\mathbf{y}_2 = A\mathbf{x}_1 \approx \begin{bmatrix} 4.115967 \\ 3.429973 \end{bmatrix}, \qquad \|\mathbf{y}_2\| \approx 5.357789, \quad \mathbf{x}_2 = \frac{\mathbf{y}_2}{\|\mathbf{y}_2\|} \approx \begin{bmatrix} 0.768221 \\ 0.640184 \end{bmatrix},$$

$$\mathbf{y}_3 = A\mathbf{x}_2 \approx \begin{bmatrix} 4.225215 \\ 4.481289 \end{bmatrix}, \qquad \|\mathbf{y}_3\| \approx 6.159090, \quad \mathbf{x}_3 = \frac{\mathbf{y}_3}{\|\mathbf{y}_3\|} \approx \begin{bmatrix} 0.686012 \\ 0.727589 \end{bmatrix},$$

$$\mathbf{y}_4 = A\mathbf{x}_3 \approx \begin{bmatrix} 4.240803 \\ 4.157649 \end{bmatrix}, \qquad \|\mathbf{y}_4\| \approx 5.938893, \quad \mathbf{x}_4 = \frac{\mathbf{y}_4}{\|\mathbf{y}_4\|} \approx \begin{bmatrix} 0.714073 \\ 0.700071 \end{bmatrix},$$

$$\mathbf{y}_5 = A\mathbf{x}_4 \approx \begin{bmatrix} 4.242432 \\ 4.270436 \end{bmatrix}, \qquad \|\mathbf{y}_5\| \approx 6.019539, \quad \mathbf{x}_5 = \frac{\mathbf{y}_5}{\|\mathbf{y}_5\|} \approx \begin{bmatrix} 0.704777 \\ 0.709429 \end{bmatrix}.$$

The $\|\mathbf{y}_k\|$ are converging to 6, while the $\mathbf{x}_k$ are converging to $\frac{1}{\sqrt{2}}(1, 1) \approx (0.707107, 0.707107)$, which is an eigenvector of eigenvalue 6.

Students who took MAT 2342 learned about *Markov chains*. Markov chains are a particular case of the power method. The matrix A is *stochastic* if each entry is a nonnegative real number and the sum of the entries in each column is equal to one. A stochastic matrix is *regular* if there exists a positive integer k such that all entries of A^k are strictly positive. One can prove that every regular stochastic matrix has 1 as a dominant eigenvalue and a unique *steady state vector*, which is an eigenvector of eigenvalue 1, whose entries are all nonnegative and sum to 1. One can then use the Theorem 3.7.1 to find this steady state vector. See [Nic, §2.9] for further details.

Remark 3.7.4. Note that the power method only allows us to compute the *dominant* eigenvalue (or the smallest eigenvalue in norm if we use the inverse power method). What if we want to find other eigenvalues? In this case, the power method has serious limitations. If A is hermitian, one can first find the dominant eigenvalue λ_n with eigenvector $\mathbf{v}_n$, and then repeat the power method with an initial vector orthogonal to $\mathbf{v}_n$. At each iteration, we subtract the projection onto $\mathbf{v}_n$ to ensure that we remain in the subspace orthogonal to $\mathbf{v}_n$. However, this is quite computationally intensive, and so is not practical for computing all eigenvalues.

3.7.2 The QR method

The QR method is the most-used algorithm to compute all the eigenvalues of a matrix. Here we will restrict our attention to the case where $A \in M_{n,n}(\mathbb{R})$ is a real matrix whose eigenvalues have distinct norms:

$$0 < |\lambda_1| < |\lambda_2| < \cdots < |\lambda_{n-1}| < |\lambda_n|. \tag{3.21}$$

(The general case is beyond the scope of this course.) These conditions on A ensure that it is invertible and diagonalizable, with distinct real eigenvalues. We do *not* assume that A is symmetric.

The *QR method* consists in computing a sequence of matrices $A_1, A_2, \ldots$ with $A_1 = A$ and

$$A_{k+1} = R_k Q_k,$$

where $A_k = Q_k R_k$ is the QR factorization of A_k, for $k \geq 1$. Note that, since A is invertible, it has a unique QR factorization by Corollary 3.6.8. Recall that the matrices Q_k are orthogonal (since they are real and unitary) and the matrices R_k are upper triangular.

Since

$$A_{k+1} = R_k Q_k = Q_k^{-1} A_k Q_k = Q_k^T A_k Q_k, \tag{3.22}$$

the matrices $A_1, A_2, \ldots$ are all similar, and hence have the same eigenvalues.

Theorem 3.7.5. *Suppose $A \in M_{n,n}(\mathbb{R})$ satisfies (3.21). In addition, assume that P^{-1} admits an LU factorization, where P is the matrix of eigenvectors of A, that is, $A = P \operatorname{diag}(\lambda_1, \ldots, \lambda_n) P^{-1}$. Then the sequence $A_1, A_2, \ldots$ produced by the QR method converges to an upper triangular matrix whose diagonal entries are the eigenvalues of A.*

Proof. We will omit the proof of this theorem. The interested student can find a proof in [AK08, Th. 10.6.1]. □

Remark 3.7.6. If the matrix A is symmetric, then so are the A_k, by (3.22). Thus, the limit of the A_k is a diagonal matrix.

Example 3.7.7. Consider the symmetric matrix

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

Using the QR method gives

$$A_{20} = \begin{bmatrix} 7.07467 & 0.0 & 0.0 \\ 0.0 & -3.18788 & 0.0 \\ 0.0 & 0.0 & -0.88679 \end{bmatrix}.$$

The diagonal entries here are the eigenvalues of A to within 5×10^{-6}.

In practice, the convergence of the QR method can be slow when the eigenvalues are close together. The speed can be improved in certain ways.

- *Shifting*: If, at stage k of the algorithm, a number s_k is chosen and $A_k - s_k I$ is factored in the form $Q_k R_k$ rather than A_k itself, then

$$Q_k^{-1} A_k Q_k = Q_k^{-1}(Q_k R_k + s_k I)Q_k = R_k Q_k + s_k I,$$

  and so we take $A_{k+1} = R_k Q_k + s_k I$. If the shifts s_k are carefully chosen, one can greatly improve convergence.
- One can first bring the matrix A to *upper Hessenberg form*, which is a matrix that is nearly upper triangular (one allows nonzero entries just below the diagonal), using a technique based on *Householder reduction*. Then the convergence in the QR algorithm is faster.

See [AK08, §10.6] for further details.

3.7.3 Gershgorin circle theorem

We conclude this section with a result that can be used to bound the spectrum (i.e. set of eigenvalues) of a square matrix. We suppose in this subsection that $A = [a_{ij}] \in M_{n,n}(\mathbb{C})$.

For $i = 1, 2, \ldots, n$, let

$$R_i = \sum_{j : j \neq i} |a_{ij}|$$

be the sum of the absolute values of the non-diagonal entries in the i-th row. For $a \in \mathbb{C}$ and $r \in \mathbb{R}_{\geq 0}$, let

$$D(a, r) = \{z \in \mathbb{C} : |z - a| \leq r\}$$

be the closed disc with centre a and radius r. The discs $D(a_{ii}, R_i)$ are called the *Gershgorin discs* of A.

Theorem 3.7.8 (Gershgorin circle theorem)**.** *Every eigenvalue of A lies in at least one of the Gershgorin discs $D(a_{ii}, R_i)$.*

Proof. Suppose λ is an eigenvalue of A. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a corresponding eigenvector. Dividing $\mathbf{x}$ by its component with largest absolute value, we can assume that, for some $i \in \{1, 2, \ldots, n\}$, we have

$$x_i = 1 \quad \text{and} \quad |x_j| \leq 1 \text{ for all } j.$$

Equating the i-th entries of $A\mathbf{x} = \lambda\mathbf{x}$, we have

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i = \lambda.$$

Since $x_i = 1$, this implies that

$$\sum_{j:j\neq i} a_{ij}x_j + a_{ii} = \lambda.$$

Then we have

$$\begin{aligned} |\lambda - a_{ii}| &= \left|\sum_{j\neq i} a_{ij}x_j\right| \\ &\leq \sum_{j\neq i} |a_{ij}|\,|x_j| && \text{(by the triangle inequality)} \\ &\leq \sum_{j\neq i} |a_{ij}| && \text{(since } |x_j| \leq 1 \text{ for all } j\text{)} \\ &= R_i. && \square \end{aligned}$$

Corollary 3.7.9. *The eigenvalues of A lie in the Gershgorin discs corresponding to the columns of A. More precisely, each eigenvalue lies in at least one of the discs $D(a_{jj}, \sum_{i:i\neq j} |a_{ij}|)$.*

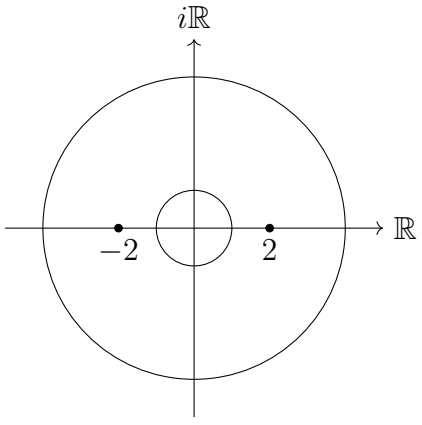
Proof. Since A and A^T have the same eigenvalues, we can apply Theorem 3.7.8 to A^T. $\square$

One can interpret the Gershgorin circle theorem as saying that if the off-diagonal entries of a square matrix have small norms, then the eigenvalues of the matrix are close to the diagonal entries of the matrix.

Example 3.7.10. Consider the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 4 & 0 \end{bmatrix}.$$

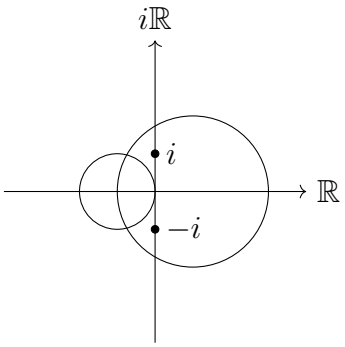
It has characteristic polynomial $x^2 - 4 = (x-2)(x+2)$, and so its eigenvalues are ± 2. The Gershgorin circle theorem tells us that the eigenvalues lie in the discs $D(0,1)$ and $D(0,4)$.



Example 3.7.11. Consider the matrix

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

It has characteristic polynomial $x^2 + 1$, and so its eigenvalues are $\pm i$. The Gershgorin circle theorem tells us that the eigenvalues lie in the discs $D(1, 2)$ and $D(-1, 1)$.



Note that, in Examples 3.7.10 and 3.7.11, it was *not* the case that each Gershgorin disc contained one eigenvalue. There was one disc that contained no eigenvalues, and one disc that contained two eigenvalues. In general, one has the following strengthened version of the Gershgorin circle theorem.

Theorem 3.7.12. *If the union of k Gershgorin discs is disjoint from the union of the other $n - k$ discs, then the former union contains exactly k eigenvalues of A, and the latter union contains exactly $n - k$ eigenvalues of A.*

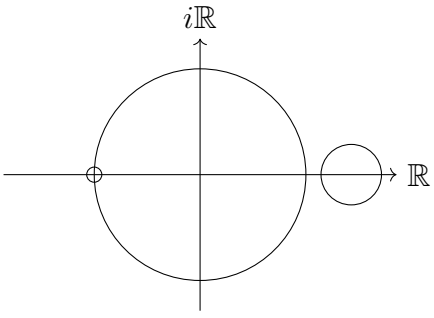
Proof. The proof of this theorem uses a continuity argument, where one starts with Gershgorin discs that are points, and gradually enlarges them. For details, see [Mey00, p. 498]. □

Example 3.7.13. Let's use the Gershgorin circle theorem to estimate the eigenvalues of

$$A = \begin{bmatrix} -7 & 0.3 & 0.2 \\ 5 & 0 & 2 \\ 1 & -1 & 10 \end{bmatrix}.$$

The Gershgorin discs are

$$D(-7, 0.5), \quad D(0, 7), \quad D(10, 2).$$



By Theorem 3.7.12, we know that two eigenvalues lie in the union of discs $D(-7, 0.5) \cup D(0, 7)$ and one lies in the disc $D(10, 2)$.

Exercises.

3.7.1. Suppose $A \in M_{n,n}(\mathbb{C})$ has eigenvalues $\lambda_1, \ldots, \lambda_r$. (We only list each eigenvalue once here, even if it has multiplicity greater than one.) Prove that the Gershgorin discs for A are precisely the sets $\{\lambda_1\}, \ldots, \{\lambda_r\}$ if and only if A is diagonal.

3.7.2. Let

$$A = \begin{bmatrix} 1.3 & 0.5 & 0.1 & 0.2 & 0.1 \\ -0.2 & 0.7 & 0 & 0.2 & 0.1 \\ 1 & -2 & 4 & 0.1 & -0.1 \\ 0 & 0.2 & -0.1 & 2 & 1 \\ 0.05 & 0 & 0.1 & 0.5 & 1 \end{bmatrix}.$$

Use the Gershgorin circle theorem to prove that A is invertible.

3.7.3. Suppose that P is a permutation matrix. (Recall that this means that each row and column of P have one entry equal to 1 and all other entries equal to 0.)

(a) Show that there are two possibilities for the Gershgorin discs of P.

(b) Show, using different methods, that the eigenvalues of a permutation all have absolute value 1. Compare this with your results from (a).

3.7.4. Fix $z \in \mathbb{C}^\times$, and define

$$A = \begin{bmatrix} 0 & 0.1 & 0.1 \\ 10 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} z & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(a) Calculate the matrix $B = CAC^{-1}$. Recall that A and B have the same eigenvalues.

(b) Give the Gershgorin discs for B, and find the values of z that give the strongest conclusion for the eigenvalues.

(c) What can we conclude about the eigenvalues of A?

Additional recommended exercises: Exercises in [Nic, §8.5].

Chapter 4

Generalized diagonalization

In this and previous courses, you've seen the concept of diagonalization. Diagonalizing a matrix makes it very easy to work with in many ways: you know the eigenvalues and eigenvectors, you can easily compute powers of the matrix, etc. However, you know that *not* all matrices are diagonalizable. So it is natural to ask if there is some slightly more general result concerning a nice form in which *all* matrices can be written. In this chapter we will consider two such forms: singular value decomposition and Jordan canonical form. Students who took MAT 2342 also saw singular value decomposition a bit in that course.

4.1 Singular value decomposition

One of the most useful tools in applied linear algebra is a factorization called *singular value decomposition* (SVD). A good reference for the material in this section is [Nic, §8.6.1]. Note however, that [Nic, §8.6] works over $\mathbb{R}$, whereas we will work over $\mathbb{C}$.

Definition 4.1.1 (Singular value decomposition)**.** A *singular value decomposition (SVD)* of $A \in M_{m,n}(\mathbb{C})$ is a factorization

$$A = P\Sigma Q^H,$$

where P and Q are unitary and, in block form,

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m\times n}, \qquad D = \operatorname{diag}(d_1, d_2, \ldots, d_r), \quad d_1, d_2, \ldots, d_r \in \mathbb{R}_{>0}.$$

Note that if $P = Q$ in Definition 4.1.1, then A is unitary diagonalizable. So SVD is a kind of generalization of unitary diagonalization. Our goal in this section is to prove that every matrix has a SVD, and to describe an algorithm for finding this decomposition. Later, we'll discuss some applications of SVDs.

Recall that hermitian matrices are nice in many ways. For instance, their eigenvalues are real and they are unitarily diagonalizable by the spectral theorem (Theorem 3.4.4). Note that for any complex matrix A (not necessarily square), both A^HA and AA^H are hermitian. It turns out that, to find a SVD of A, we should study these matrices.

Lemma 4.1.2. *Suppose $A \in M_{m,n}(\mathbb{C})$.*

(a) *The eigenvalues of A^HA and AA^H are real and nonnegative.*

(b) *The matrices A^HA and AA^H have the same set of positive eigenvalues.*

Proof. (a) Suppose λ is an eigenvalue of A^HA with eigenvector $\mathbf{v} \in \mathbb{C}^n$, $\mathbf{v} \neq 0$. Then we have

$$\|A\mathbf{v}\|^2 = (A\mathbf{v})^HA\mathbf{v} = \mathbf{v}^HA^HA\mathbf{v} = \mathbf{v}^H(\lambda\mathbf{v}) = \lambda\mathbf{v}^H\mathbf{v} = \lambda\|\mathbf{v}\|^2.$$

Thus $\lambda = \|A\mathbf{v}\|^2/\|\mathbf{v}\|^2 \in \mathbb{R}_{\geq 0}$. The proof for AA^H is analogous, replacing A by A^H.

(b) Suppose λ is a positive eigenvalues of A^HA with eigenvector $\mathbf{v} \in \mathbb{C}^n$, $\mathbf{v} \neq 0$. Then $A\mathbf{v} \in \mathbb{C}^m$ and

$$AA^H(A\mathbf{v}) = A\left((A^HA)\mathbf{v}\right) = A(\lambda\mathbf{v}) = \lambda(A\mathbf{v}).$$

Also, we have $A\mathbf{v} \neq 0$ since $A^HA\mathbf{v} = \lambda\mathbf{v} \neq 0$ (because $\lambda \neq 0$ and $\mathbf{v} \neq 0$). Thus λ is an eigenvalue of AA^H. This proves that every positive eigenvalue of A^HA is an eigenvalue of AA^H. For the reverse inclusion, we replace A everywhere by A^H. □

We now analyze the symmetric matrix A^HA, called the *Gram matrix* of A (see Section 1.5.4), in more detail.

Step 1: Unitarily diagonalize A^HA

Since the matrix A^HA is hermitian, the spectral theorem (Theorem 3.4.4) states that we can choose an orthonormal basis

$$\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\} \subseteq \mathbb{C}^n$$

consisting of eigenvectors of A^HA with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. By Lemma 4.1.2(a), we have $\lambda_i \in \mathbb{R}_{\geq 0}$ for each i. We may choose the order of the $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \quad \text{and} \quad \lambda_i = 0 \text{ for } i > r. \tag{4.1}$$

(We allow the possibility that $r = 0$, so that $\lambda_i = 0$ for all i, and also the possibility that $r = n$.) By Proposition 3.3.9,

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} \text{ is unitary} \quad \text{and} \quad \text{unitarily diagonalizes } A^HA. \tag{4.2}$$

Step 2: Show that $\operatorname{rank} A = r$

Recall from Section 1.3 that

$$\operatorname{rank} A = \dim(\operatorname{col} A) = \dim(\operatorname{im} T_A).$$

We wish to show that $\operatorname{rank} A = r$, where r is defined in (4.1). We do this by showing that

$$\{A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r\} \text{ is an orthogonal basis of } \operatorname{im} T_A. \tag{4.3}$$

First note that, for all i, j, we have

$$\langle A\mathbf{q}_i, A\mathbf{q}_j\rangle = (A\mathbf{q}_i)^HA\mathbf{q}_j = \mathbf{q}_i^H(A^HA)\mathbf{q}_j = \mathbf{q}_i^H(\lambda_j\mathbf{q}_j) = \lambda_j\mathbf{q}_i^H\mathbf{q}_j = \lambda_j\langle\mathbf{q}_i, \mathbf{q}_j\rangle.$$

Since the set $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n\}$ is orthonormal, this implies that

$$\langle A\mathbf{q}_i, A\mathbf{q}_j \rangle = 0 \text{ if } i \neq j \quad \text{and} \quad \|A\mathbf{q}_i\|^2 = \lambda_i \|\mathbf{q}_i\|^2 = \lambda_i \text{ for each } i. \tag{4.4}$$

Thus, using (4.1), we see that

$$\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\} \subseteq \operatorname{im} T_A \text{ is an orthogonal set} \quad \text{and} \quad A\mathbf{q}_i = 0 \text{ if } i > r. \tag{4.5}$$

Since $\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\}$ is orthogonal, it is linearly independent. It remains to show that it spans $\operatorname{im} T_A$. So we need to show that

$$U := \operatorname{Span}\{A\mathbf{q}_1, A\mathbf{q}_2, \dots, A\mathbf{q}_r\}$$

is equal to $\operatorname{im} T_A$. Since we already know $U \subseteq \operatorname{im} T_A$, we just need to show that $\operatorname{im} T_A \subseteq U$. So we need to show

$$A\mathbf{x} \in U \quad \text{for all } \mathbf{x} \in \mathbb{C}^n.$$

Suppose $\mathbf{x} \in \mathbb{C}^n$. Since $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n\}$ is a basis for $\mathbb{C}^n$, we can write

$$\mathbf{x} = t_1\mathbf{q}_1 + t_2\mathbf{q}_2 + \cdots + t_n\mathbf{q}_n \quad \text{for some } t_1, t_2, \dots, t_n \in \mathbb{C}.$$

Then, by (4.5), we have

$$A\mathbf{x} = t_1 A\mathbf{q}_1 + t_2 A\mathbf{q}_2 + \cdots + t_n A\mathbf{q}_n = t_1 A\mathbf{q}_1 + \cdots + t_r A\mathbf{q}_r \in U,$$

as desired.

Step 3: Some definitions

Definition 4.1.3 (Singular values of A)**.** The real numbers

$$\sigma_i = \sqrt{\lambda_i} \overset{(4.4)}{=} \|A\mathbf{q}_i\|, \quad i = 1, 2, \dots, n \tag{4.6}$$

are called the *singular values* of the matrix A.

Note that, by (4.1), we have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_i = 0 \text{ if } i > r.$$

So the number of *positive* singular values is equal to $r = \operatorname{rank} A$.

Definition 4.1.4 (Singular matrix of A)**.** Define

$$D_A = \operatorname{diag}(\sigma_1, \dots, \sigma_r).$$

Then the $m \times n$ matrix

$$\Sigma_A := \begin{bmatrix} D_A & 0 \\ 0 & 0 \end{bmatrix}$$

(in block form) is called the *singular matrix* of A.

Remark 4.1.5. Don't confuse the "singular matrix of A" with the concept of a "singular matrix" (a square matrix that is not invertible).

Step 4: Find an orthonormal basis of $\mathbb{C}^m$ compatible with $\operatorname{col} A$

Normalize the vectors $A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r$ by defining

$$\mathbf{p}_i = \frac{1}{\|A\mathbf{q}_i\|} A\mathbf{q}_i \overset{(4.6)}{=} \frac{1}{\sigma_i} A\mathbf{q}_i \in \mathbb{C}^m \quad \text{for } i = 1, 2, \ldots, r. \tag{4.7}$$

It follows from (4.3) that

$$\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_r\} \text{ is an } \textit{orthonormal} \text{ basis of } \operatorname{im} T_A = \operatorname{col} A \subseteq \mathbb{C}^m. \tag{4.8}$$

By Proposition 3.1.2, we can find $\mathbf{p}_{r+1}, \ldots, \mathbf{p}_m$ so that

$$\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\} \text{ is an orthonormal basis of } \mathbb{C}^m. \tag{4.9}$$

Step 5: Find the decomposition

By Section 4.1 and (4.2) we have two unitary matrices

$$P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_m \end{bmatrix} \in M_{m,m}(\mathbb{C}) \quad \text{and} \quad Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} \in M_{n,n}(\mathbb{C}).$$

We also have

$$\sigma_i \mathbf{p}_i \overset{(4.6)}{=} \|A\mathbf{q}_i\| \mathbf{p}_i \overset{(4.7)}{=} A\mathbf{q}_i \quad \text{for } i = 1, 2, \ldots, r.$$

Using this and (4.5), we have

$$AQ = \begin{bmatrix} A\mathbf{q}_1 & A\mathbf{q}_2 & \cdots & A\mathbf{q}_n \end{bmatrix} = \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & 0 & \cdots & 0 \end{bmatrix}.$$

Then we compute

$$\begin{aligned} P\Sigma_A &= \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \mathbf{p}_{r+1} & \cdots & \mathbf{p}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots 0 & \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 \cdots & 0 & \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & 0 & \cdots & 0 \end{bmatrix} \\ &= AQ. \end{aligned}$$

Since $Q^{-1} = Q^H$, it follows that $A = P\Sigma_A Q^H$. Thus we have proved the following theorem.

Theorem 4.1.6. *Let $A \in M_{m,n}(\mathbb{C})$, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ be the positive singular values of A. Then $r = \operatorname{rank} A$ and we have a factorization*

$$A = P\Sigma_A Q^H \quad \textit{where } P \textit{ and } Q \textit{ are unitary matrices.}$$

In particular, every complex matrix has a SVD.

Note that the SVD is *not* unique. For example, if $r < m$, then there are infinitely many ways to extend $\{\mathbf{p}_1, \ldots, \mathbf{p}_r\}$ to an orthonormal basis $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ of $\mathbb{C}^m$. Each such extension leads to a different matrix P in the SVD. For another example illustrating non-uniqueness, consider $A = I_n$. Then $\Sigma_A = I_n$, and $A = P\Sigma_A P^H$ is a SVD of A for *any* unitary $n \times n$ matrix P.

Remark 4.1.7 (Real SVD). If $A \in M_{m,n}(\mathbb{R})$, then we can find a SVD where P and Q are *real* (hence orthogonal) matrices. To see this, observe that our proof is valid if we replace $\mathbb{C}$ by $\mathbb{R}$ everywhere.

Our proof of Theorem 4.1.6 gives an algorithm for finding SVDs.

Example 4.1.8. Let's find a SVD of the matrix

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

Since this matrix is real, we can write A^T instead of A^H in our SVD algorithm. We have

$$A^T A = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 8 & 0 \\ -2 & 0 & 2 \end{bmatrix} \quad \text{and} \quad AA^T = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix}.$$

As expected, both of these matrices are symmetric. It's easier to find the eigenvalues of AA^T. This has characteristic polynomial

$$(x-6)^2 - 4 = x^2 - 12x + 32 = (x-4)(x-8).$$

Thus, the eigenvalues of AA^T are $\lambda_1 = 8$ and $\lambda_2 = 4$, both with multiplicity one. It follows from Lemma 4.1.2(b) that the eigenvalues of $A^T A$ are $\lambda_1 = 8$, $\lambda_2 = 4$, and $\lambda_3 = 0$, all with multiplicity one. So the positive singular values of A are

$$\sigma_1 = \sqrt{\lambda_1} = 2\sqrt{2} \quad \text{and} \quad \sigma_2 = \sqrt{\lambda_2} = 2.$$

We now find an orthonormal basis of each eigenspace of $A^T A$. We find the eigenvectors

$$\mathbf{q}_1 = (0, 1, 0), \quad \mathbf{q}_2 = \left(\frac{-1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right), \quad \mathbf{q}_3 = \left(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right).$$

We now compute

$$\mathbf{p}_1 = \frac{1}{\sigma_1} A\mathbf{q}_1 = \frac{1}{2\sqrt{2}}(-2, 2) = \left(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),$$
$$\mathbf{p}_2 = \frac{1}{\sigma_2} A\mathbf{q}_2 = \frac{1}{2}(-\sqrt{2}, -\sqrt{2}) = \left(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right).$$

Then define

$$P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix},$$

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = \begin{bmatrix} 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix},$$
$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} 2\sqrt{2} & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}.$$

We can then check that $A = P\Sigma Q^T$.

In practice, SVDs are not computed using the above method. There are sophisticated numerical algorithms for calculating the singular values, P, Q, and the rank of an $m \times n$ matrix to a high degree of accuracy. Such algorithms are beyond the scope of this course.

Our algorithm gives us a way of finding *one* SVD. However, since SVDs are not unique, it is natural to ask how they are related.

Lemma 4.1.9. *If $A = P\Sigma Q^H$ is any SVD for A as in Definition 4.1.1, then*

(a) $r = \operatorname{rank} A$, *and*

(b) *the $d_1, \ldots, d_r$ are the singular values of A in some order.*

Proof. We have

$$A^H A = \left(P\Sigma Q^H\right)^H \left(P\Sigma Q^H\right) = Q\Sigma^H P^H P\Sigma Q^H = Q\Sigma^H \Sigma Q^H. \tag{4.10}$$

Thus $\Sigma^H\Sigma$ and $A^H A$ are similar matrices, and so $\operatorname{rank}(A^H A) = \operatorname{rank}(\Sigma^H\Sigma) = r$. As we saw in Step 2 above, $\operatorname{rank} A = \operatorname{rank} A^H A$. Hence $\operatorname{rank} A = r$.

It also follows from the fact that $\Sigma^H\Sigma$ and $A^H A$ are similar that they have the same eigenvalues. Thus

$$\{d_1^2, d_2^2, \ldots, d_r^2\} = \{\lambda_1, \lambda_2, \ldots, \lambda_r\},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the positive eigenvalues of $A^H A$. So there is some permutation τ of $\{1, 2, \ldots, r\}$ such that $d_i^2 = \lambda_{\tau(i)}$ for each $i = 1, 2, \ldots, r$. Therefore $d_i = \sqrt{\lambda_{\tau(i)}} = \sigma_{\tau(i)}$ for each i by (4.6). □

Exercises.

4.1.1. Show that $A \in M_{m,n}(\mathbb{C})$ is the zero matrix if and only if all of its singular values are zero.

Recommended exercises: Exercises in [Nic, §8.6]: 8.6.2–8.6.12, 8.6.16.

4.2 Fundamental subspaces and principal components

A singular value decomposition contains a lot of useful information about a matrix. We explore this phenomenon in this section. A good reference for most of the material here is [Nic, §8.6.2].

Definition 4.2.1 (Fundamental subspaces)**.** The *fundamental subspaces* of $A \in M_{m,n}(\mathbb{F})$ are:

- $\operatorname{row} A = \operatorname{Span}\{\mathbf{x} : \mathbf{x} \text{ is a row of } A\}$,
- $\operatorname{col} A = \operatorname{Span}\{\mathbf{x} : \mathbf{x} \text{ is a column of } A\}$,
- $\operatorname{null} A = \{\mathbf{x} : A\mathbf{x} = 0\}$,
- $\operatorname{null} A^H = \{\mathbf{x} : A^H\mathbf{x} = 0\}$.

Recall Definition 3.1.8 of the orthogonal complement $U^\perp$ of a subspace U of $\mathbb{F}^n$. We will need a few facts about the orthogonal complement.

Lemma 4.2.2. (a) *If $A \in M_{m,n}(\mathbb{F})$, then*

$$(\operatorname{row} \bar{A})^\perp = \operatorname{null} A \quad \textit{and} \quad (\operatorname{col} A)^\perp = \operatorname{null} A^H.$$

(b) *If U is any subspace of $\mathbb{F}^n$, then $(U^\perp)^\perp = U$.*

(c) *Suppose $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is an orthonormal basis of $\mathbb{F}^m$. If $U = \operatorname{Span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ for some $1 \le k \le m$, then*

$$U^\perp = \operatorname{Span}\{\mathbf{v}_{k+1}, \dots, \mathbf{v}_m\}.$$

Proof. (a) Let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be the rows of A. Then

$$\begin{aligned}
\mathbf{x} \in \operatorname{null} A &\iff A\mathbf{x} = 0 \\
&\iff \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix} \mathbf{x} = 0 \\
&\iff \begin{bmatrix} \mathbf{a}_1\mathbf{x} \\ \vdots \\ \mathbf{a}_m\mathbf{x} \end{bmatrix} = 0 && \text{(block multiplication)} \\
&\iff \mathbf{a}_i\mathbf{x} = 0 \text{ for all } i \\
&\iff \langle \overline{\mathbf{a}_i}, \mathbf{x} \rangle = 0 \text{ for all } i \\
&\iff \mathbf{x} \in (\operatorname{Span}\{\overline{\mathbf{a}_1}, \dots, \overline{\mathbf{a}_m}\})^\perp = (\operatorname{row} \bar{A})^\perp. && \text{(by Lemma 3.1.9(c))}
\end{aligned}$$

Thus $\operatorname{null} A = (\operatorname{row} \bar{A})^\perp$. Replacing A by A^H, we get

$$\operatorname{null} A^H = (\operatorname{row} A^T)^\perp = (\operatorname{col} A)^\perp.$$

(b) You saw this in previous courses, so we will not repeat the proof here. See [Nic, Lem. 8.6.4].

(c) We leave this as an exercise. The proof can be found in [Nic, Lem. 8.6.4]. □

Now we can see that any SVD for a matrix A immediately gives orthonormal bases for the fundamental subspaces of A.

Theorem 4.2.3. *Suppose $A \in M_{m,n}(\mathbb{F})$. Let $A = P\Sigma Q^H$ be a SVD for A, where*

$$P = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \in M_{m,m}(\mathbb{F}) \quad \text{and} \quad Q = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} \in M_{n,n}(\mathbb{F})$$

are unitary (hence orthogonal if $\mathbb{F} = \mathbb{R}$) and

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}, \quad \text{where} \quad D = \operatorname{diag}(d_1, d_2, \ldots, d_r), \text{ with each } d_i \in \mathbb{R}_{>0}.$$

Then

(a) *$r = \operatorname{rank} A$, and the positive singular values of A are $d_1, d_2, \ldots, d_r$;*

(b) *the fundamental spaces are as follows:*

(i) *$\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal basis of $\operatorname{col} A$,*

(ii) *$\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ is an orthonormal basis of $\operatorname{null} A^H$,*

(iii) *$\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\operatorname{null} A$,*

(iv) *$\{\overline{\mathbf{v}_1}, \ldots, \overline{\mathbf{v}_r}\}$ is an orthonormal basis of $\operatorname{row} A$.*

Proof. (a) This is Lemma 4.1.9.

(b) (i) Since Q is invertible, we have $\operatorname{col} A = \operatorname{col}(AQ) = \operatorname{col}(P\Sigma)$. Also

$$P\Sigma = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \operatorname{diag}(d_1, \ldots, d_r) & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} d_1\mathbf{u}_1 & \cdots & d_r\mathbf{u}_r & 0 & \cdots & 0 \end{bmatrix}.$$

Thus

$$\operatorname{col} A = \operatorname{Span}\{d_1\mathbf{u}_1, \ldots, d_r\mathbf{u}_r\} = \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}.$$

Since the $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal, they are linearly independent. So $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for $\operatorname{col} A$.

(ii) We have

$$\begin{aligned} \operatorname{null} A^H &= (\operatorname{col} A)^\perp && \text{(by Lemma 4.2.2(a))} \\ &= (\operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\})^\perp \\ &= \operatorname{Span}\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}. && \text{(by Lemma 4.2.2(c))} \end{aligned}$$

(iii) We first show that the proposed basis has the correct size. By the Rank-Nullity Theorem (see (1.7)), we have

$$\dim(\operatorname{null} A) = n - r = \dim(\operatorname{Span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}).$$

So, if we can show that

$$\operatorname{Span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\} \subseteq \operatorname{null} A, \tag{4.11}$$

it will follow that $\operatorname{Span}\{\mathbf{v}_{r+1},\dots,\mathbf{v}_n\} = \operatorname{null} A$ (since the two spaces have the same dimension), and hence that $\{\mathbf{v}_{r+1},\dots,\mathbf{v}_n\}$ is a basis for $\operatorname{null} A$ (since it is linearly independent because it is orthonormal).

To show the inclusion (4.11), it is enough to show that $\mathbf{v}_j \in \operatorname{null} A$ (i.e. $A\mathbf{v}_j = 0$) for $j > r$. Define

$$d_{r+1} = \cdots = d_n = 0, \quad \text{so that} \quad \Sigma^H\Sigma = \operatorname{diag}(d_1^2,\dots,d_n^2).$$

Then, for $1 \le j \le n$, we have

$$\begin{aligned}(A^HA)\mathbf{v}_j &= (Q\Sigma^H\Sigma Q^H)\mathbf{v}_j && \text{(by (4.10))}\\ &= (Q\Sigma^H\Sigma Q^H)Q\mathbf{e}_j\\ &= Q(\Sigma^H\Sigma)\mathbf{e}_j\\ &= Q(d_j^2\mathbf{e}_j)\\ &= d_j^2Q\mathbf{e}_j\\ &= d_j^2\mathbf{v}_j.\end{aligned}$$

Thus, for $1 \le i \le n$,

$$\|A\mathbf{v}_j\|^2 = (A\mathbf{v}_j)^HA\mathbf{v}_j = \mathbf{v}_j^HA^HA\mathbf{v}_j = \mathbf{v}_j^H(d_j^2\mathbf{v}_j) = d_j^2\|\mathbf{v}_j\|^2 = d_j^2.$$

In particular, $A\mathbf{v}_j = 0$ for $j > r$, as desired.

(iv) First note that

$$\begin{aligned}(\operatorname{row}\bar{A})^\perp &= \operatorname{null} A && \text{(by Lemma 4.2.2(a))}\\ &= \operatorname{Span}\{\mathbf{v}_{r+1},\dots,\mathbf{v}_n\}. && \text{(by (iii))}\end{aligned}$$

Then

$$\begin{aligned}\operatorname{row}\bar{A} &= \big((\operatorname{row}\bar{A})^\perp\big)^\perp && \text{(by Lemma 4.2.2(b))}\\ &= (\operatorname{Span}\{\mathbf{v}_{r+1},\dots,\mathbf{v}_n\})^\perp\\ &= \operatorname{Span}\{\mathbf{v}_1,\dots,\mathbf{v}_r\}. && \text{(by Lemma 4.2.2(c))}\end{aligned}$$

Taking complex conjugates, we have $\operatorname{row} A = \operatorname{Span}\{\overline{\mathbf{v}_1},\dots,\overline{\mathbf{v}_r}\}$. □

Example 4.2.4. Suppose we want to solve the homogeneous linear system

$$A\mathbf{x} = 0 \quad \text{of } m \text{ equations in } n \text{ variables.}$$

The set of solutions is precisely $\operatorname{null} A$. If we compute a SVD $A = U\Sigma V^T$ for A then, in the notation of Theorem 4.2.3, the set $\{\mathbf{v}_{r+1},\dots,\mathbf{v}_n\}$ is an orthonormal basis of the solution set.

SVDs are also closely related to *principal component analysis*. If $A = P\Sigma Q^T$ is a SVD with, in the notation of Theorem 4.2.3,

$$\sigma_1 = d_1 > \sigma_2 = d_2 > \cdots > \sigma_n = d_n > 0,$$

then we have

$$\begin{aligned} A &= P\Sigma Q^T \\ &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}^H \\ &= \begin{bmatrix} \sigma_1\mathbf{u}_1 & \cdots & \sigma_r\mathbf{u}_r & 0 & \cdots 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^H \\ \vdots \\ \mathbf{v}_n^H \end{bmatrix} \\ &= \sigma_1\mathbf{u}_1\mathbf{v}_1^H + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^H. \qquad \text{(block multiplication)} \end{aligned}$$

Thus, we have written A as a sum of r rank one matrices (see Exercise 4.2.1), called the *principle components* of A. For $1 \leq t \leq r$, the matrix

$$A_t := \sigma_1\mathbf{u}_1\mathbf{v}_1^H + \ldots + \sigma_t\mathbf{u}_t\mathbf{v}_t^H$$

is a *truncated matrix* that is the closest approximation to A by a rank t matrix, in the sense that the difference between A and A_t has the smallest possible Frobenius norm. This result is known as the *Eckart–Young–Mirsky Theorem*. Such approximations are particularly useful in data analysis and machine learning. There exist algorithms for finding A_t without computing a full SVD.

Exercises.

4.2.1. Suppose $\mathbf{u} \in \mathbb{C}^n$ and $\mathbf{v} \in \mathbb{C}^m$. Show that if $\mathbf{u}$ and $\mathbf{v}$ are both nonzero, then the rank of the matrix $\mathbf{u}\mathbf{v}^H$ is 1.

4.3 Pseudoinverses

In Section 1.5 we discussed left and right inverses. In particular, we discussed the *pseudoinverse* in Section 1.5.4. This was a particular left/right inverse; in general left/right inverses are not unique. We'll now discuss this concept in a bit more detail, seeing what property uniquely characterizes the pseudoinverse and how we can compute it using a SVD. A good reference for this material is [Nic, §8.6.4].

Definition 4.3.1 (Middle inverse)**.** A *middle inverse* of $A \in M_{m,n}(\mathbb{F})$ is a matrix $B \in M_{n,m}(\mathbb{F})$ such that

$$ABA = A \quad \text{and} \quad BAB = B.$$

Examples 4.3.2. (a) Suppose A is left-invertible, with left inverse B. Then

$$ABA = AI = A \quad \text{and} \quad BAB = IB = B,$$

so B is a middle inverse of A. Conversely if C is any other middle inverse of A, then

$$ACA = A \implies BACA = BA \implies CA = I,$$

and so C is a left inverse of A. Thus, middle inverses of A are the same as left inverses.

(b) If A right invertible, then middle inverses of A are the same as right inverses. The proof is analogous to the one above.

(c) It follows that if A is invertible, then middle inverse are the same as inverses.

In general, middle inverses are not unique, even for square matrices.

Example 4.3.3. If

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

then

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

is a middle inverse for any b.

While the middle inverse is not unique in general, it turns out that it is unique if we require that AB and BA be hermitian.

Theorem 4.3.4 (Penrose' Theorem)**.** *For any $A \in M_{m,n}(\mathbb{C})$, there exists a unique $B \in M_{n,m}(\mathbb{C})$ such that*

(P1) *$ABA = A$ and $BAB = B$.*

(P2) *Both AB and BA are hermitian.*

Proof. A proof can be found in [Pen55]. □

Definition 4.3.5 (Pseudoinverse)**.** The *pseudoinverse* (or *Moore–Penrose inverse*) of $A \in M_{m,n}(\mathbb{C})$ is the unique $A^+ \in M_{n,m}(\mathbb{C})$ such that A and A^+ satisfy (P1) and (P2), that is:

$$AA^+A = A, \quad A^+AA^+ = A^+, \quad \text{and both } AA^+ \text{ and } A^+A \text{ are hermitian.}$$

If A is invertible, then $A^+ = A^{-1}$, as follows from Example 4.3.2. Also, the symmetry in the conditions (P1) and (P2) imply that $A^{++} = A$.

The following proposition shows that the terminology *pseudoinverse*, as used above, coincides with our use of this terminology in Section 1.5.4.

Proposition 4.3.6. *Suppose $A \in M_{m,n}(\mathbb{C})$.*

(a) *If $\operatorname{rank} A = m$, then AA^H is invertible and $A^+ = A^H(AA^H)^{-1}$.*

(b) *If $\operatorname{rank} A = n$, then A^HA is invertible and $A^+ = (A^HA)^{-1}A^H$.*

Proof. We proof the first statement; the proof of the second is similar. If $\operatorname{rank} A = m$, then the rows of A are linearly independent and so, by Proposition 1.5.16(a) (we worked over $\mathbb{R}$ there, but the same result holds over $\mathbb{C}$ if we replace the transpose by the conjugate transpose), AA^H is invertible. Then

$$A\left(A^H(AA^H)^{-1}\right)A = (AA^H)(AA^H)^{-1}A = IA = A$$

and

$$\left(A^H(AA^H)^{-1}\right)A\left(A^H(AA^H)^{-1}\right) = A^H(AA^H)^{-1}(AA^H)(AA^H)^{-1} = A^H(AA^H)^{-1}.$$

Furthermore, $A\left(A^H(AA^H)^{-1}\right) = I$ is hermitian, and

$$\left(\left(A^H(AA^H)^{-1}\right)A\right)^H = \left(A^H(AA^H)^{-1}\right)A,$$

so $\left(A^H(AA^H)^{-1}\right)A$ is also hermitian. □

In turns out that if we have a SVD for A, then it is particularly easy to compute the pseudoinverse.

Proposition 4.3.7. *Suppose $A \in M_{m,n}(\mathbb{C})$ and $A = P\Sigma Q^H$ is a SVD for A as in Definition 4.1.1, with*

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m\times n}, \quad D = \operatorname{diag}(d_1, d_2, \ldots, d_r), \quad d_1, d_2, \ldots, d_r \in \mathbb{R}_{>0}.$$

Then $A^+ = Q\Sigma' P^H$, where

$$\Sigma^{-1} = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}_{n\times m}.$$

Proof. It is straightforward to verify that

$$\Sigma\Sigma'\Sigma = \Sigma, \quad \Sigma'\Sigma\Sigma' = \Sigma', \quad \Sigma\Sigma' = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{m\times m}, \quad \Sigma'\Sigma = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{n\times n}.$$

In particular, Σ' is the pseudoinverse of Σ. Now let $B = Q\Sigma' P^H$. Then

$$ABA = (P\Sigma Q^H)(Q\Sigma' P^H)(P\Sigma Q^H) = P\Sigma\Sigma'\Sigma Q^H = P\Sigma Q^H = A.$$

Similarly, $BAB = B$. Furthermore,

$$AB = U(\Sigma\Sigma')U^H \quad \text{and} \quad BA = Q(\Sigma'\Sigma)Q^H$$

are both hermitian. Thus $B = A^+$. □

Example 4.3.8. Consider the matrix

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

We saw in Example 4.3.3 that

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

is a middle inverse for B for any choice of b. We have

$$AB = \begin{bmatrix} 1 & b \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad BA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

So BA is always symmetric, but AB is symmetric exactly when $b = 0$. Hence $B = A^+$ if and only if $b = 0$.

Let's compute A^+ using a SVD of A. The matrix

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

has eigenvalues $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0$ with orthonormal eigenvectors

$$\mathbf{q}_1 = (1, 0, 0), \quad \mathbf{q}_2 = (0, 1, 0), \quad \mathbf{q}_3 = (0, 0, 1).$$

Thus we take $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = I_3$. In addition, A has rank 1 with singular values

$$\sigma_1 = 1, \quad \sigma_2 = 0, \quad \sigma_3 = 0.$$

Thus

$$\Sigma_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = A \quad \text{and} \quad \Sigma' = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = A^T.$$

We then compute

$$\mathbf{p}_1 = \frac{1}{\sigma_1} A\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We extend this to an orthonormal basis of $\mathbb{C}^2$ by choosing

$$\mathbf{p}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Hence

$$P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 \end{bmatrix} = I_2.$$

Thus a SVD for A is $A = P\Sigma_A Q^T = \Sigma_A = A$. Therefore the pseudoinverse of A is

$$A^+ = Q\Sigma'_A P^T = \Sigma'_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = A^T.$$

We conclude this section with a list of properties of the pseudoinverse, many of which parallel properties of the inverse.

Proposition 4.3.9. *Suppose $A \in M_{m,n}(\mathbb{C})$.*

(a) $A^{++} = A$.

(b) *If A is invertible, then $A^+ = A^{-1}$.*

(c) *The pseudoinverse of a zero matrix is its transpose.*

(d) $(A^T)^+ = (A^+)^T$.

(e) $(\bar{A})^+ = \overline{A^+}$.

(f) $(A^H)^+ = (A^+)^H$.

(g) $(zA)^+ = z^{-1}A^+$ *for* $z \in \mathbb{C}^\times$.

(h) *If $A \in M_{m,n}(\mathbb{R})$, then $A^+ \in M_{n,m}(\mathbb{R})$.*

Proof. We've already proved the first two. The proof of the remaining properties is left as Exercise 4.3.3. □

Exercises.

4.3.1. Suppose that B is a middle inverse for A. Prove that B^T is a middle inverse for A^T.

4.3.2. A square matrix M is *idempotent* if $M^2 = M$. Show that, if B is a middle inverse for A, then AB and BA are both idempotent matrices.

4.3.3. Complete the proof of Proposition 4.3.9.

Additional recommended exercises from [Nic, §8.6]: 8.6.1, 8.6.17.

4.4 Jordan canonical form

One of most fundamental results in linear algebra is the classification of matrices up to similarity. It turns out that every matrix is similar to a matrix of a special form, called *Jordan canonical form*. In this section we discuss this form. We will omit some of the technical parts of the proof, referring to [Nic, §§11.1, 11.2] for details. Throughout this section we work over the field $\mathbb{C}$ of complex numbers.

By Schur's theorem (Theorem 3.4.2), every matrix is unitarily similar to an upper triangular matrix. The following theorem shows that, in fact, every matrix is similar to a special type of upper triangular matrix.

Proposition 4.4.1 (Block triangulation)**.** *Suppose $A \in M_{n,n}(\mathbb{C})$ has characteristic polynomial*

$$c_A(x) = (x - \lambda_1)^{m_1}(x - \lambda_2)^{m_2} \cdots (x - \lambda_k)^{m_k}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the distinct eigenvalues of A. Then there exists an invertible matrix P such that

$$P^{-1}AP = \begin{bmatrix} U_1 & 0 & 0 & \cdots & 0 \\ 0 & U_2 & 0 & \cdots & 0 \\ 0 & 0 & U_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & U_k \end{bmatrix}$$

where each U_i is an $m_i \times m_i$ upper triangular matrix with all entries on the main diagonal entry equal to λ_i.

Proof. The proof proceeds by induction on n. See [Nic, Th. 11.1.1] for details. □

Recall that if λ is an eigenvalue of A, then the associated eigenspace is

$$E_\lambda = \text{null}(\lambda I - A) = \{\mathbf{x} : (\lambda I - A)\mathbf{x} = 0\} = \{\mathbf{x} : A\mathbf{x} = \lambda\mathbf{x}\}.$$

Definition 4.4.2 (Generalized eigenspace)**.** If λ is an eigenvalue of the matrix A, then the associated *generalized eigenspace* is

$$G_\lambda = G_\lambda(A) := \text{null}(\lambda I - A)^{m_\lambda},$$

where m_λ is the algebraic multiplicity of λ. (We use the notation $G_\lambda(A)$ if we want to emphasize the matrix.)

Note that we always have

$$E_\lambda(A) \subseteq G_\lambda(A).$$

Example 4.4.3. Consider the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in M_{2,2}(\mathbb{C}).$$

The only eigenvalue of A is $\lambda = 0$. The associated eigenspace is

$$E_0(A) = \text{Span}\{(1, 0)\}.$$

However, since $A^2 = 0$, we have

$$G_0 = \mathbb{C}^2.$$

Recall that the *geometric multiplicity* of the eigenvalue λ of A is $\dim E_\lambda$ and we have $\dim E_\lambda \leq m_\lambda$. This inequality can be strict, as in Example 4.4.3. (In fact, it is strict for some eigenvalue precisely when the matrix A is not diagonalizable.)

Lemma 4.4.4. *If λ is an eigenvalue of A, then $\dim G_\lambda(A) = m_\lambda$, where m_λ is the algebraic multiplicity of λ.*

Proof. Choose P as in Proposition 4.4.1 and choose an eigenvalue λ_i. To simplify notation, let $B = (\lambda_i I - A)^{m_i}$. We have an isomorphism

$$G_\lambda(A) = \operatorname{null} B \to \operatorname{null}(P^{-1}BP), \quad \mathbf{x} \mapsto P^{-1}\mathbf{x} \tag{4.12}$$

(Exercise 4.4.1). Thus, it suffices to show that $\dim \operatorname{null}(P^{-1}BP) = m_i$. Using the notation of Proposition 4.4.1, we have $\lambda = \lambda_i$ for some i and

$$\begin{aligned} P^{-1}BP &= (\lambda_i I - P^{-1}AP)^{m_i} \\ &= \begin{bmatrix} \lambda_i I - U_1 & 0 & \cdots & 0 \\ 0 & \lambda_i I - U_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_i I - U_k \end{bmatrix}^{m_i} \\ &= \begin{bmatrix} (\lambda_i I - U_1)^{m_i} & 0 & \cdots & 0 \\ 0 & (\lambda_i I - U_2)^{m_i} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (\lambda_i I - U_k)^{m_i} \end{bmatrix}. \end{aligned}$$

The matrix $\lambda_i I - U_j$ is upper triangular with main diagonal entries equal to $\lambda_i - \lambda_j$; so $(\lambda_i I - U_j)^{m_i}$ it is invertible when $i \neq j$ and the zero matrix when $i = j$ (see Exercise 3.4.3(b)). Therefore $m_i = \dim \operatorname{null}(P^{-1}BP)$, as desired. □

Definition 4.4.5 (Jordan block)**.** For $n \in \mathbb{Z}_{>0}$ and $\lambda \in \mathbb{C}$, the *Jordan block* $J(n, \lambda)$ is the $n \times n$ matrix with λ's on the main diagonal, $1's$ on the diagonal above, and 0's elsewhere. By convention, we set $J(1, \lambda) = \begin{bmatrix} \lambda \end{bmatrix}$.

We have

$$J(1,\lambda) = \begin{bmatrix} \lambda \end{bmatrix}, \; J(2,\lambda) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \; J(3,\lambda) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}, \; J(4,\lambda) = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}, \text{ etc.}$$

Our goal is to show that Proposition 4.4.1 holds with each block U_i replaced by a Jordan block. The key is to show the result for $\lambda = 0$. We say a linear operator (or matrix) T is *nilpotent* if $T^m = 0$ for some $m \geq 1$. Every eigenvalue of a nilpotent linear operator or matrix is equal to zero (Exercise 4.4.2). The converse also holds by Proposition 4.4.1 and Exercise 3.4.3(b).

Lemma 4.4.6. *If $A \in M_{n,n}(\mathbb{C})$ is nilpotent, then there exists an invertible matrix P such that*

$$P^{-1}AP = \operatorname{diag}(J_1, J_2, \ldots, J_k),$$

where each J_i is a Jordan block $J(m, 0)$ for some m.

Proof. The proof proceeds by induction on n. See [Nic, Lem. 11.2.1] for details. □

Theorem 4.4.7 (Jordan canonical form)**.** *Suppose $A \in M_{n,n}(\mathbb{C})$ has distinct (i.e. non-repeated) eigenvalues $\lambda_1, \ldots, \lambda_k$. For $1 \leq i \leq k$, let m_i be the algebraic multiplicity of λ_i. Then there exists an invertible matrix P such that*

$$P^{-1}AP = \operatorname{diag}(J_1, \ldots, J_m), \tag{4.13}$$

where each J_ℓ is a Jordan block corresponding to some eigenvalue λ_i. Furthermore, the sum of the sizes of the Jordan block corresponding to λ_i is equal to m_i. The form (4.13) is called a Jordan canonical form *of A.*

Proof. By Proposition 4.4.1, there exists an invertible matrix Q such that

$$Q^{-1}AQ = \begin{bmatrix} U_1 & 0 & 0 & \cdots & 0 \\ 0 & U_2 & 0 & \cdots & 0 \\ 0 & 0 & U_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & U_k \end{bmatrix},$$

where each U_i is upper triangular with entries on the main diagonal equal to λ_i. Suppose that for each U_i we can find an invertible matrix P_i as in the statement of the theorem. Then

$$\begin{bmatrix} P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2 & 0 & \cdots & 0 \\ 0 & 0 & P_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & P_k \end{bmatrix}^{-1} Q^{-1}AQ \begin{bmatrix} P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2 & 0 & \cdots & 0 \\ 0 & 0 & P_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & P_k \end{bmatrix}$$

$$= \begin{bmatrix} P_1^{-1}U_1P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2^{-1}U_2P_1 & 0 & \cdots & 0 \\ 0 & 0 & P_3^{-1}U_3P_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & P_k^{-1}U_kP_k \end{bmatrix}$$

would be in Jordan canonical form.

So it suffices to prove the theorem in the case that A is an upper triangular matrix with diagonal entries all equal to some $\lambda \in \mathbb{C}$. Then $A - \lambda I$ is strictly upper triangular (upper triangular with zeros on the main diagonal), and hence nilpotent by Exercise 3.4.3(b). Therefore, by Lemma 4.4.6, there exists an invertible matrix P such that

$$P^{-1}(A - \lambda I)P = \operatorname{diag}(J_1, \ldots, J_\ell),$$

where each J_i is a Jordan block $J(m, 0)$ for some m. Then

$$P^{-1}AP = P^{-1}\big((A - \lambda I) + \lambda I\big)P = P^{-1}(A - \lambda I)P + \lambda I = \operatorname{diag}(J_1 + \lambda I, \ldots, J_\ell + \lambda I),$$

and each $J_i + \lambda I$ is a Jordan block $J(m, \lambda)$ for some m. □

Remarks 4.4.8. (a) Suppose that $T\colon V \to V$ is a linear operator on a finite-dimensional vector space V. If we choose a basis B of V, then we can find the matrix of T relative to this basis. It follows from Theorem 4.4.7 that we can choose the basis B so that this matrix is in Jordan canonical form.

(b) The Jordan canonical form of a matrix (or linear operator) is uniquely determined up to reordering the Jordan blocks. In other words, for each eigenvalue λ, the number and size of the Jordan blocks corresponding to λ are uniquely determined. This is most easily proved using more advanced techniques from the theory of modules. It follows that two matrices are similar if and only if they have the same Jordan canonical form (up to re-ordering of the Jordan blocks).

(c) A matrix is diagonalizable if and only if, in its Jordan canonical form, all Jordan blocks have size 1.

Exercises.

4.4.1. Show that (4.12) is an isomorphism.

4.4.2. Show that if $T\colon V \to V$ is a nilpotent linear operator, then zero is the only eigenvalue of T.

Additional exercises from [Nic, §11.2]: 11.2.1, 11.2.2.

4.5 The matrix exponential

As discussed in Section 3.4, we can substitute a matrix $A \in M_{n,n}(\mathbb{C})$ into any polynomial. In this section, we'd like to make sense of the expression e^A. Recall from calculus that, for $x \in \mathbb{C}$, we have the power series

$$e^x = 1 + x + \frac{x}{2} + \frac{x^3}{3!} + \cdots = \sum_{m=0}^{\infty} \frac{x^m}{m!}.$$

(You probably only saw this for $x \in \mathbb{R}$, but it holds for complex values of x as well.) So we might naively try to define

$$e^A = \sum_{m=0}^{\infty} \frac{1}{m!} A^m.$$

In fact, this *is* the correct definition. However, we need to justify that this infinite sum makes sense (i.e. that in converges) and figure out how to compute it. To do this, we use the Jordan canonical form.

As a warm up, let's suppose $A = \operatorname{diag}(a_1, \ldots, a_n)$ is diagonal. Then

$$e^A = \sum_{m=0}^{\infty} \frac{1}{m!} A^m = \sum_{m=0}^{\infty} \frac{1}{m!} \operatorname{diag}(a_1^m, \ldots, a_n^m) = \operatorname{diag}(e^{a_1}, \ldots, e^{a_m}). \tag{4.14}$$

So exponentials of diagonal matrices exist and are easy to compute.

Now suppose A is arbitrary. By Theorem 4.4.7, there exists an invertible matrix P such that $B = P^{-1}AP$ is in Jordan canonical form. For each m, we have

$$\frac{1}{m!} A^m = \frac{1}{m!} \left(PBP^{-1}\right)^m = P \left(\frac{1}{m!} B^m\right) P^{-1}.$$

Summing over m, it follows that

$$e^A = P e^B P^{-1}$$

provided that the infinite sum for e^B converges. If A is diagonalizable, then its Jordan canonical form B is diagonal, and we can then compute e^B as in (4.14).

The above discussion shows that we can restrict our attention to matrices in Jordan canonical form. However, since the Jordan canonical form might not be diagonal, we have some more work to do. Note that if A is in Jordan canonical form, we have

$$A = \operatorname{diag}(J_1, \ldots, J_k)$$

for some Jordan blocks J_i. But then, for all m, we have

$$A^m = \operatorname{diag}(J_1^m, \ldots, J_k^m).$$

Thus

$$e^A = \operatorname{diag}(e^{J_1}, \ldots, e^{J_k})$$

provided that the sum for each e^{J_i} converges. So it suffices to consider the case where A is a *single* Jordan block.

Let's first consider the case of a Jordan block $J = J(n, 0)$ corresponding to eigenvalue zero. Then we know from Exercise 3.4.3(b) that $J^n = 0$. Hence

$$e^J = \sum_{m=0}^{\infty} \frac{1}{m!} J^m = \sum_{m=0}^{n-1} \frac{1}{m!} J^m. \tag{4.15}$$

Since this sum is finite, there are no convergence issues.

Now consider an arbitrary Jordan block $J(n, \lambda)$, $\lambda \in \mathbb{C}$. We have

$$J(n, \lambda) = \lambda I + J(n, 0)$$

Now, if $A, B \in M_{n,n}(\mathbb{C})$ commute (i.e. $AB = BA$) and then the usual argument shows that

$$e^{A+B} = e^A e^B.$$

(It is crucial that A and B commute; see Exercise 4.5.2.) Thus

$$e^{J(n,\lambda)} = e^{\lambda I + J(n,0)} = e^{\lambda I} e^{J(n,0)} = e^{\lambda} I e^{J(n,0)} = e^{\lambda} e^{J(n,0)}.$$

So we can compute the exponential of any Jordan block. Hence, by our above discussion, we can compute the exponential of *any* matrix. We sometimes write $\exp(A)$ for e^A.

Example 4.5.1. Suppose

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

We have

$$\exp\left(\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}\right) = e^2 \exp\left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\right) = e^2 \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and

$$\exp\left(\begin{bmatrix} -1 \end{bmatrix}\right) = \begin{bmatrix} e^{-1} \end{bmatrix}.$$

Hence

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

Proposition 4.5.2. *For $A \in M_{n,n}(\mathbb{C})$ we have $\det e^A = e^{\operatorname{tr} A}$.*

Proof. We first prove the result for a Jordan block $J = J(n, \lambda)$. We have

$$\det e^J = \det\left(e^\lambda e^{J(n,0)}\right) = \left(e^\lambda\right)^n \det e^{J(n,0)} = e^{n\lambda} \cdot 1 = e^{\operatorname{tr} J},$$

where we have used the fact (which follows from (4.15)) that $e^{J(n,0)}$ is upper triangular with all entries on the main diagonal equal to 1 (hence $\det e^{J(n,0)} = 1$). Now suppose A is arbitrary. Since similar matrices have the same determinant and trace (see Exercises 3.4.1 and 3.4.2), we can use Theorem 4.4.7 to assume that A is in Jordan canonical form. So

$$A = \operatorname{diag}(J_1, \ldots, J_k)$$

for some Jordan blocks J_i. Then

$$\begin{aligned} \det e^A &= \det\left(\operatorname{diag}(e^{J_1}, e^{J_2}, \ldots, e^{J_k})\right) \\ &= (\det e^{J_1})(\det e^{J_2}) \cdots (\det e^{J_k}) && \text{(by (1.2))} \\ &= e^{\operatorname{tr} J_1} e^{\operatorname{tr} J_2} \cdots e^{\operatorname{tr} J_k} && \text{(by the above)} \\ &= e^{\operatorname{tr} J_1 + \operatorname{tr} J_2 + \cdots + \operatorname{tr} J_k} && \text{(since } e^{a+b} = e^a e^b \text{ for } a, b \in \mathbb{C}) \\ &= e^{\operatorname{tr}(\operatorname{diag}(J_1, J_2, \ldots, J_k))} \\ &= e^{\operatorname{tr} A}. \end{aligned}$$

□

The matrix exponential can be used to solve certain initial value problems.. You probably learned in calculus that the solution to the differential equation

$$x'(t) = ax(t), \quad x(t_0) = x_0, \quad a, t_0, x_0 \in \mathbb{R}$$

is

$$x(t) = e^{(t-t_0)a} x_0.$$

In fact, this continues to hold more generally. If

$$\mathbf{x}(t) = (x_1(t), \ldots, x_n(t)), \quad \mathbf{x}'(t) = (x_1'(t), \ldots, x_n'(t)), \quad \mathbf{x}_0 \in \mathbb{R}^n, \quad A = M_{n,n}(\mathbb{C}),$$

then the solution to the initial value problem

$$\mathbf{x}'(t) = A\mathbf{x}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0,$$

is

$$\mathbf{x}(t) = e^{(t-t_0)A}\mathbf{x}_0.$$

In addition to the matrix exponential, it is also possible to compute other functions of matrices ($\sin A$, $\cos A$, etc.) using Taylor series for these functions. See [Usm87, Ch. 6] for further details.

Exercises.

4.5.1. Prove that e^A is invertible for any $A \in M_{n,n}(\mathbb{C})$.

4.5.2. Let

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Show that $e^A e^B \neq e^{A+B}$.

4.5.3. If

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

compute e^A.

4.5.4. If

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

compute e^A by summing the power series.

4.5.5. Solve the initial value problem $\mathbf{x}'(t) = A\mathbf{x}(t)$, $\mathbf{x}(1) = (1, -2, 3)$, where A is the matrix of Example 4.5.1.

Chapter 5

Quadratic forms

In your courses in linear algebra you have spent a great deal of time studying linear functions. The next level of complexity involves quadratic functions. These have a large number of applications, including in the theory of conic sections, optimization, physics, and statistics. Matrix methods allow for a unified treatment of quadratic functions. Conversely, quadratic functions provide useful insights into the theory of eigenvectors and eigenvalues. Good references for the material in this chapter are [Tre, Ch. 7] and [ND77, Ch. 10].

5.1 Definitions

A *quadratic form* on $\mathbb{R}^n$ is a homogenous polynomial of degree two in the variables $x_1, \ldots, x_n$. In other words, a quadratic form is a polynomial $Q(\mathbf{x}) = Q(x_1, \ldots, x_n)$ having only terms of degree two. So only terms that are scalar multiples of x_k^2 and $x_j x_k$ are allowed.

Every quadratic form on $\mathbb{R}^n$ can be written in the form $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^T A \mathbf{x}$ for some matrix $A \in M_{n,n}(\mathbb{R})$. In general, the matrix A is not unique. For instance, the quadratic form

$$Q(\mathbf{x}) = 2x_1^2 - x_2^2 + 6x_1x_2$$

can be written as $\langle \mathbf{x}, A\mathbf{x} \rangle$ where A is any of the matrices

$$\begin{bmatrix} 2 & 6 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 6 & -1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 3 \\ 3 & -1 \end{bmatrix}.$$

In fact, we can choose any matrix of the form

$$\begin{bmatrix} 2 & 6-a \\ a & -1 \end{bmatrix}, \quad a \in \mathbb{F}.$$

Note that only the choice of $a = 3$ yields a *symmetric* matrix.

Lemma 5.1.1. *For any quadratic form $Q(\mathbf{x})$ on $\mathbb{R}^n$, there is a* unique *symmetric matrix $A \in M_{n,n}(\mathbb{R})$ such that $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$.*

Proof. Suppose

$$Q(\mathbf{x}) = \sum_{i \le j} c_{ij} x_i x_j.$$

If $A = [a_{ij}]$ is symmetric, we have

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \sum_{i,j=1}^{n} a_{ij} x_i x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i<j} (a_{ij} + a_{ji}) x_i x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i<j} 2a_{ij} x_i x_j.$$

Thus, we have $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$ if and only if

$$a_{ii} = c_{ii} \text{ for } 1 \leq i \leq n, \quad a_{ij} = a_{ji} = \frac{1}{2} c_{ij} \text{ for } 1 \leq i < j \leq n. \qquad \square$$

Example 5.1.2. If

$$Q(\mathbf{x}) = x_1^2 - 3x_2^2 + 4x_3^2 + 3x_1 x_2 - 12x_1 x_3 + 8x_2 x_3,$$

then the corresponding symmetric matrix is

$$\begin{bmatrix} 1 & 1.5 & -6 \\ 1.5 & -3 & 4 \\ -6 & 4 & 2 \end{bmatrix}.$$

We can also consider quadratic forms on $\mathbb{C}^n$. Typically, we still want the quadratic form to take *real* values. Before we consider such quadratic forms, we state an important identity.

Lemma 5.1.3 (Polarization identity)**.** *Suppose $A \in M_{n,n}(\mathbb{F})$.*

(a) *If $\mathbb{F} = \mathbb{C}$, then*

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \frac{1}{4} \sum_{\alpha = \pm 1, \pm i} \alpha \langle \mathbf{x} + \alpha \mathbf{y}, A(\mathbf{x} + \alpha \mathbf{y}) \rangle.$$

(b) *If $\mathbb{F} = \mathbb{R}$ and $A = A^T$, then*

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \frac{1}{4} \Big(\langle \mathbf{x} + \mathbf{y}, A(\mathbf{x} + \mathbf{y}) \rangle - \langle \mathbf{x} - \mathbf{y}, A(\mathbf{x} - \mathbf{y}) \rangle \Big).$$

Proof. We leave the proof as Exercise 5.1.1. $\square$

The importance of the polarization identity is that the values $\langle \mathbf{x}, A\mathbf{y} \rangle$ are completely determined by the values of the corresponding quadratic form (the case where $\mathbf{x} = \mathbf{y}$).

Lemma 5.1.4. *Suppose $A \in M_{n,n}(\mathbb{C})$. Then*

$$\langle \mathbf{x}, A\mathbf{x} \rangle \in \mathbb{R} \text{ for all } \mathbf{x} \in \mathbb{C}^n$$

if and only if A is hermitian.

Proof. First suppose that A is hermitian. Then, for all $\mathbf{x} \in \mathbb{C}^n$,

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, A^H \mathbf{x} \rangle = \langle A\mathbf{x}, \mathbf{x} \rangle \overset{\text{(IP1)}}{=} \overline{\langle \mathbf{x}, A\mathbf{x} \rangle}.$$

Thus $\langle \mathbf{x}, A\mathbf{x} \rangle \in \mathbb{R}$.

Now suppose that $\langle \mathbf{x}, A\mathbf{x}\rangle \in \mathbb{R}$ for all $\mathbf{x} \in \mathbb{C}^n$, and $A = [a_{ij}]$. Using the polarization identity (Lemma 5.1.3) we have, for all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,

$$\begin{aligned}\langle \mathbf{x}, A\mathbf{y}\rangle &= \frac{1}{4}\Big(\langle \mathbf{x}+\mathbf{y}, A(\mathbf{x}+\mathbf{y})\rangle - \langle \mathbf{x}-\mathbf{y}, A(\mathbf{x}-\mathbf{y})\rangle + i\langle \mathbf{x}+i\mathbf{y}, A(\mathbf{x}+i\mathbf{y})\rangle - i\langle \mathbf{x}-i\mathbf{y}, A(\mathbf{x}-i\mathbf{y})\rangle\Big)\\ &\overset{\text{(IP2)}}{\underset{\text{(IP3)}}{=}} \frac{1}{4}\Big(\langle \mathbf{x}+\mathbf{y}, A(\mathbf{x}+\mathbf{y})\rangle - \langle \mathbf{y}-\mathbf{x}, A(\mathbf{y}-\mathbf{x})\rangle + i\langle \mathbf{y}-i\mathbf{x}, A(\mathbf{y}-i\mathbf{x})\rangle - i\langle \mathbf{y}+i\mathbf{x}, A(\mathbf{y}+i\mathbf{x})\rangle\Big)\\ &= \overline{\langle \mathbf{y}, A\mathbf{x}\rangle} \overset{\text{(IP1)}}{=} \langle A\mathbf{x}, \mathbf{y}\rangle\\ &= \langle \mathbf{x}, A^H\mathbf{y}\rangle.\end{aligned}$$

(In the third equality, we used the fact that all the inner products appearing in the expression are real.) It follows that $A = A^H$ (e.g. take $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$). □

In light of Lemma 5.1.4, we define a *quadratic form* on $\mathbb{C}^n$ to be a function of the form

$$Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x}\rangle \quad \text{for some hermitian } A \in M_{n,n}(\mathbb{C}).$$

Thus a quadratic form on $\mathbb{C}^n$ is a function $\mathbb{C}^n \to \mathbb{R}$.

Exercises.

5.1.1. Prove Lemma 5.1.3.

5.1.2. Consider the quadratic form

$$Q(x_1, x_2, x_3) = x_1^2 - x_2^2 + 4x_3^3 + x_1x_2 - 3x_1x_3 + 5x_2x_3.$$

Find *all* matrices $A \in M_{3,3}(\mathbb{R})$ such that $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x}\rangle$. Which one is symmetric?

5.2 Diagonalization of quadratic forms

Suppose $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x}\rangle$ is a quadratic form. If A is diagonal, then the quadratic form is particularly easy to understand. For instance, if $n = 2$, then it is of the form

$$Q(\mathbf{x}) = a_1x_1^2 + a_2x_2^2.$$

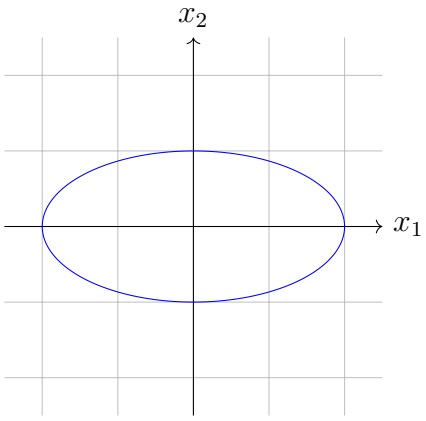
For fixed $c \in \mathbb{R}$, the *level set* consisting of all points $\mathbf{x}$ satisfying

$$Q(\mathbf{x}) = c$$

is an ellipse or a hyperbola (or parallel lines, which is a kind of degenerate hyperbola, or the empty set) whose axes are parallel to the x-axis and/or y-axis. For example, if $\mathbb{F} = \mathbb{R}$, the set defined by

$$x_1^2 + 4x_2^2 = 4$$

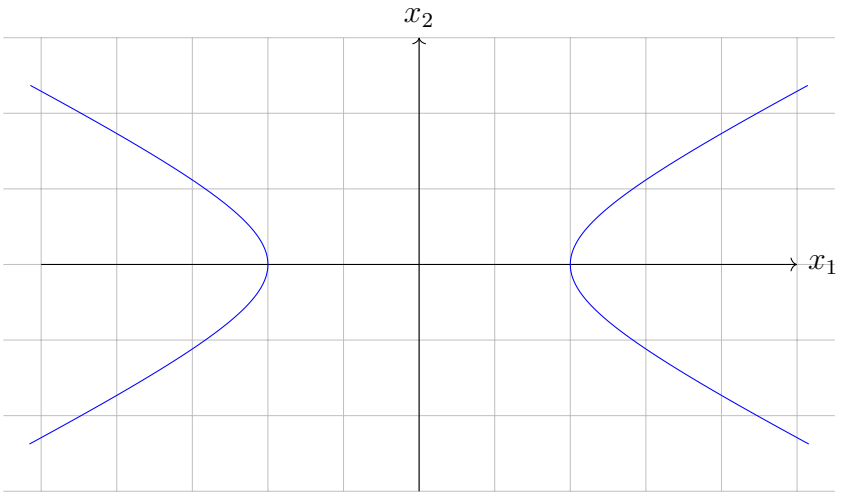
is the ellipse



while the set defined by

$$x_1^2 - 4x_2^2 = 4$$

is the following hyperbola:



If $n = 3$, then the level sets are ellipsoids or hyperboloids.

Given an arbitrary quadratic form, we would like to make it diagonal by changing variables. Suppose we introduce some new variables

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = S^{-1}\mathbf{x},$$

where $S \in M_{n,n}(\mathbb{F})$ is some invertible matrix, so that $\mathbf{x} = S\mathbf{y}$. Then we have

$$Q(\mathbf{x}) = Q(S\mathbf{y}) = \langle S\mathbf{y}, AS\mathbf{y}\rangle = \langle \mathbf{y}, S^H AS\mathbf{y}\rangle.$$

Therefore, in the new variables $\mathbf{y}$, the quadratic form has matrix $S^H AS$.

So we would like to find an invertible matrix S so that $S^H AS$ is diagonal. Note that is a bit different than usual diagonalization, where you want $S^{-1}AS$ to be diagonal. However, if S is unitary, then $S^H = S^{-1}$. Fortunately, we know from the spectral theorem (Theorem 3.4.4) that we can unitarily diagonalize a hermitian matrix. So we can find a unitary matrix U such that

$$U^H AU = U^{-1}AU$$

is diagonal. The columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ of U form an orthonormal basis B of $\mathbb{F}^n$. Then U is the change of coordinate matrix from this basis to the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. We have

$$\mathbf{y} = U^{-1}\mathbf{x},$$

so $y_1, y_2, \ldots, y_n$ are the coordinates of the vector $\mathbf{x}$ in the new basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$.

Example 5.2.1. Let $\mathbb{F} = \mathbb{R}$ and consider the quadratic form

$$Q(x_1, x_2) = 5x_1^2 + 5x_2^2 - 2x_1x_2.$$

Let's describe the level set given by $Q(x_1, x_2) = 1$. The matrix of Q is

$$A = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}.$$

We can orthogonally diagonalizes this matrix (exercise) as

$$D = \begin{bmatrix} 4 & 0 \\ 0 & 6 \end{bmatrix} = U^T AU, \quad \text{with} \quad U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

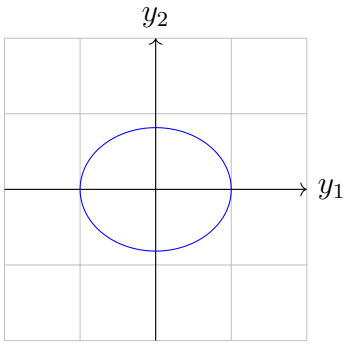
Thus, we introduce the new variables y_1, y_2 given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = U \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} y_1 + y_2 \\ y_2 - y_1 \end{bmatrix}.$$

In these new variables, the quadratic form is

$$Q(y_1, y_2) = 4y_1^2 + 6y_2^2.$$

So the level set given by $Q(y_1, y_2) = 1$ is the ellipse with half-axes $1/2$ and $1/\sqrt{6}$:



The set $\{\mathbf{x} : \langle \mathbf{x}, A\mathbf{x} \rangle = 1\}$ is the same ellipse, but in the basis $(1/\sqrt{2}, -1/\sqrt{2}), (1/\sqrt{2}, 1/\sqrt{2})$. In other words, it is the same ellipse, rotated $-\pi/4$.

Unitary diagonalization involves computing eigenvalues and eigenvectors, and so it can be difficult to do for large n. However, one can non-unitarily diagonalize the quadratic form associated to A, i.e. find an invertible S (without requiring that S be unitary) such that $S^H AS$ is diagonal. This is much easier computationally. It can be done by completion of squares or by using row/column operations. See [Tre, §2.2] for details.

Proposition 5.2.2 (Sylvester's law of inertia)**.** *Suppose $A \in M_{n,n}(\mathbb{C})$ is hermitian. There exists an invertible matrix S such that $S^H AS$ is of the form*

$$\operatorname{diag}(1, \dots, 1, -1, \dots, -1, 0, \dots, 0),$$

where

- *the number of 1's is equal to the number of positive eigenvalues of A (with multiplicity),*
- *the number of -1's is equal to the number of negative eigenvalues of A (with multiplicity),*
- *the number of 0's is equal to the multiplicity of zero as an eigenvalue of A.*

Proof. We leave the proof as Exercise 5.2.3. □

The sequence $(1, \dots, 1, -1, \dots, -1, 0, \dots, 0)$ appearing in Proposition 5.2.2 is called the *signature* of the hermitian matrix A.

Exercises.

5.2.1 ([Tre, Ex. 7.2.2]). For the matrix

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

unitarily diagonalize the corresponding quadratic form. That is, find a diagonal matrix D and a unitary matrix U such that $D = U^H AU$.

5.2.2 ([ND77, Ex. 10.2.4]). For each quadratic form $Q(\mathbf{x})$ below, find new variables $\mathbf{y}$ in which the form is diagonal. Then graph the level curves $Q(\mathbf{y}) = t$ (in the coordinates $\mathbf{y}$) for various values of t.

(a) $x_1^2 + 4x_1x_2 - 2x_x^2$
(b) $x_1^2 + 12x_1x_2 + 4x_2^2$
(c) $2x_1^2 + 4x_1x_2 + 4x_2^2$
(d) $-5x_1^2 - 8x_1x_2 - 5x_2^2$
(e) $11x_1^2 + 2x_1x_2 + 3x_2^2$

5.2.3. Prove Proposition 5.2.2.

5.3 Rayleigh's principle and the min-max theorem

In many applications (e.g. in certain problems in statistics), one wants to maximize a quadratic form $Q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle$ subject to the condition $\|\mathbf{x}\| = 1$. (Recall that A is hermitian.) This is equivalent to maximizing the *Rayleigh quotient*

$$R_A(\mathbf{x}) := \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

subject to the condition $\mathbf{x} \neq 0$. (The proof that this equivalent is similar to the proof of Proposition 2.3.2. We leave it as Exercise 5.3.1.)

It turns out that it is fairly easy to maximize the Rayleigh quotient when A is diagonal. As we saw in Section 5.2 we can always unitarily diagonalize A, and hence the associated quadratic form. Let U be a unitary matrix such that

$$U^H A U = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

and consider the new variables

$$\mathbf{y} = U^{-1}\mathbf{x} = U^H \mathbf{x} \quad \text{so that} \quad \mathbf{x} = U\mathbf{y}.$$

Then

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^H A \mathbf{x} = (U\mathbf{y})^H A U \mathbf{y} = \mathbf{y}^H U^H A U \mathbf{y} = \mathbf{y}^H D \mathbf{y} = \langle \mathbf{y}, D\mathbf{y} \rangle$$

and

$$\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^H \mathbf{x} = (U\mathbf{y})^H U \mathbf{y} = \mathbf{y}^H U^H U \mathbf{y} = \mathbf{y}^H \mathbf{y} = \langle \mathbf{y}, \mathbf{y} \rangle.$$

Thus

$$R_A(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = \frac{\langle \mathbf{y}, D\mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} = R_D(\mathbf{y}). \tag{5.1}$$

So we can always reduce the problem of maximizing a Rayleigh quotient to one of maximizing a *diagonalized Rayleigh quotient*.

Theorem 5.3.1. *Suppose $A \in M_{n,n}(\mathbb{C})$ is hermitian with eigenvalues*

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

and corresponding orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Then

(a) *$\lambda_1 \leq R_A(\mathbf{x}) \leq \lambda_n$ for all $\mathbf{x} \neq 0$;*

(b) *λ_1 is the minimum value of $R_A(\mathbf{x})$ for $\mathbf{x} \neq 0$, and $R_A(\mathbf{x}) = \lambda_1$ if and only if $\mathbf{x}$ is an eigenvector of eigenvalue λ_1;*

(c) *λ_n is the maximum value of $R_A(\mathbf{x})$ for $\mathbf{x} \neq 0$, and $R_A(\mathbf{x}) = \lambda_n$ if and only if $\mathbf{x}$ is an eigenvector of eigenvalue λ_n.*

Proof. By (5.1) and the fact that

$$\mathbf{y} = \mathbf{e}_i \iff \mathbf{x} = U\mathbf{e}_i = \mathbf{v}_i,$$

it suffices to prove the theorem for the diagonal matrix $D = (\lambda_1, \lambda_2, \ldots, \lambda_n)$.

(a) This follows immediately from (b) and (c).

(b) We clearly have

$$R_D(\mathbf{e}_1) = \frac{\langle \mathbf{e}_1, D\mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} = \lambda_1 \frac{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle}{\langle \mathbf{e}_1, \mathbf{e}_1 \rangle} = \lambda_1.$$

Also, for any $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$, we have

$$\begin{aligned} R_D(\mathbf{x}) &= \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \\ &= \frac{\langle \mathbf{x}, A(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \\ &= \frac{\langle x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n, x_1\lambda_1\mathbf{e}_1 + \cdots x_n\lambda_n\mathbf{e}_n \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \\ &= \frac{\lambda_1|x_1|^2 + \cdots + \lambda_n|x_n|^2}{\langle \mathbf{x}, \mathbf{x} \rangle} \\ &\geq \frac{\lambda_1|x_1|^2 + \cdots + \lambda_n|x_n|^2}{|x_1|^2 + \cdots + |x_n|^2} = \lambda_1, \end{aligned}$$

with equality holding if and only if $x_i = 0$ for all i with $\lambda_i > \lambda_1$; that is, if and only if $D\mathbf{x} = \lambda_1\mathbf{x}$.

(c) This follows from applying (b) to $-A$. □

Example 5.3.2. Consider the hermitian matrix

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

We have

$$R_A(1,0,0) = \frac{\langle (1,0,0), (1,2,1) \rangle}{1} = 1 \quad \text{and} \quad R_A(0,1,0) = \frac{\langle (0,1,0), (2,-1,0) \rangle}{1} = -1.$$

Thus, by Theorem 5.3.1, we have

$$\lambda_1 \leq -1 \quad \text{and} \quad 1 \leq \lambda_3.$$

Additional choices of $\mathbf{x}$ can give us improved bounds. In fact, $\lambda_1 = -\sqrt{6}$, $\lambda_2 = -1$, and $\lambda_3 = \sqrt{6}$. Note that $R_A(0,1,0) = -1$ is an eigenvalue, even though $(0,1,0)$ is not an eigenvector. This does not contradict Theorem 5.3.1 since -1 is neither the minimum nor the maximum eigenvalue.

We studied matrix norms in Section 2.3. In Theorem 2.3.5 we saw that there are easy ways to compute the 1-norm $\|A\|_1$ and the ∞-norm $\|A\|_\infty$. But we only obtained an inequality for the 2-norm $\|A\|_2$. We can now say something more precise.

Corollary 5.3.3 (Matrix 2-norm)**.** *For any $A \in M_{m,n}(\mathbb{C})$, the matrix 2-norm $\|A\|_2$ is equal to σ_1, the largest singular value of A.*

Proof. We have

$$\begin{aligned}
\|A\|_2^2 &= \left(\max \left\{ \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \mathbf{x} \neq 0 \right\} \right)^2 \\
&= \max \left\{ \frac{\|A\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} : \mathbf{x} \neq 0 \right\} \\
&= \max \left\{ \frac{\langle A\mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} : \mathbf{x} \neq 0 \right\} \\
&= \max \left\{ \frac{\langle \mathbf{x}, A^H A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} : \mathbf{x} \neq 0 \right\} \\
&= \lambda,
\end{aligned}$$

where λ is the largest eigenvalue of the hermitian matrix $A^H A$ (by Theorem 5.3.1). Since $\sigma_1 = \sqrt{\lambda}$, we are done. □

Theorem 5.3.1 relates the maximum (resp. minimum) eigenvalue of a hermitian matrix A to the maximum (resp. minimum) value of the Rayleigh quotient $R_A(\mathbf{x})$. But what if we are interested in the intermediate eigenvalues?

Theorem 5.3.4 (Rayleigh's principle)**.** *Suppose $A \in M_{n,n}(\mathbb{C})$ is hermitian, with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and an orthonormal set of associated eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. For $1 \leq j \leq n$, let*

- *S_j be the set of all $\mathbf{x} \neq 0$ that are orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j$.*
- *T_j be the set of all $\mathbf{x} \neq 0$ that are orthogonal to $\mathbf{v}_n, \mathbf{v}_{n-1}, \cdots, \mathbf{v}_{n-j+1}$,*

Then:

(a) *$R_A(\mathbf{x}) \geq \lambda_{j+1}$ for all $\mathbf{x} \in S_j$, and $R_A(\mathbf{x}) = \lambda_{j+1}$ for $\mathbf{x} \in S_j$ if and only if $\mathbf{x}$ is an eigenvector associated to λ_{j+1}.*

(b) *$R_A(\mathbf{x}) \leq \lambda_{n-j}$ for all $\mathbf{x} \in T_j$, and $R_A(\mathbf{x}) = \lambda_{n-j}$ for $\mathbf{x} \in T_j$ if and only if $\mathbf{x}$ is an eigenvector associated to λ_{n-j}.*

Proof. The proof is very similar to that of Theorem 5.3.1, so we will omit it. □

Rayleigh's principle (Theorem 5.3.4) characterizes each eigenvector and eigenvalue of an $n \times n$ hermitian matrix in terms of an extremum (i.e. maximization or minimization) problem. However, this characterization of the eigenvectors/eigenvalues other than the largest and smallest requires knowledge of eigenvectors other than the one being characterized. In particular, to characterize or estimate λ_j for $1 < j < n$, we need the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}$ or $\mathbf{v}_{j+1}, \ldots, \mathbf{v}_n$.

The following result remedies the aforementioned issue by giving a characterization of each eigenvalue that is independent of the other eigenvalues.

Theorem 5.3.5 (Min-max theorem)**.** *Suppose $A \in M_{n,n}(\mathbb{C})$ is hermitian with eigenvalues*

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Then

$$\lambda_k = \min_{U:\dim U=k} \{\max\{R_A(\mathbf{x}) : \mathbf{x} \in U,\ \mathbf{x} \neq 0\}\}$$

and

$$\lambda_k = \max_{U:\dim U=n-k+1} \{\min\{R_A(\mathbf{x}) : \mathbf{x} \in U,\ \mathbf{x} \neq 0\}\},$$

where the first min/max in each expression is over subspaces $U \subseteq \mathbb{C}^n$ of the given dimension.

Proof. We prove the first assertion, since the proof of the second is similar. Since A is hermitian, it is unitarily diagonalizable by the spectral theorem (Theorem 3.4.4). So we can choose an orthonormal basis

$$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$$

of eigenvectors, with $\mathbf{u}_i$ an eigenvector of eigenvalue λ_i.

Suppose $U \subseteq \mathbb{C}^n$ with $\dim U = k$. Then

$$\dim U + \dim \operatorname{Span}\{\mathbf{u}_k, \ldots, \mathbf{u}_n\} = k + (n - k + 1) = n + 1 > n.$$

Hence

$$U \cap \operatorname{Span}\{\mathbf{u}_k, \ldots, \mathbf{u}_n\} \neq \{0\}.$$

So we can choose a nonzero vector

$$\mathbf{v} = \sum_{i=k}^{n} a_i \mathbf{u}_i \in U$$

and

$$R_A(\mathbf{v}) = \frac{\langle \mathbf{v}, A\mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} = \frac{\sum_{i=k}^{n} \lambda_i |a_i|^2}{\sum_{i=k}^{n} |a_i|^2} \geq \frac{\sum_{i=k}^{n} \lambda_k |a_i|^2}{\sum_{i=k}^{n} |a_i|^2} = \lambda_k.$$

Since this is true for all U, we have

$$\min_{U:\dim U=k} \{\max\{R_A(\mathbf{x}) : \mathbf{x} \in U,\ \mathbf{x} \neq 0\}\} \geq \lambda_k.$$

To prove the reverse inequality, choose the particular subspace

$$V = \operatorname{Span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}.$$

Then

$$\max\{R_A(\mathbf{x}) : \mathbf{x} \in V,\ \mathbf{x} \neq 0\} \leq \lambda_k,$$

since λ_k is the largest eigenvalue in V. Thus we also have

$$\min_{U:\dim U=k} \{\max\{R_A(x) : x \in U,\ x \neq 0\}\} \leq \lambda_k. \qquad \square$$

Note that, when $k = 1$ or $k = n$, Theorem 5.3.5 recovers Theorem 5.3.1.

Example 5.3.6. Consider the hermitian matrix

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

of Example 5.3.2. Let's use the min-max theorem to get some bounds on λ_2. So $n = 3$ and $k = 2$. Take

$$V = \text{Span}\{(1, 0, 0), (0, 0, 1)\}.$$

Then $\dim V = 2$. The nonzero element of V are the vectors

$$(x_1, 0, x_3) \neq (0, 0, 0).$$

We have

$$R_A(x_1, 0, x_3) = \frac{\langle (x_1, 0, x_3), (x_1, 2x_1 + 5x_3, -x_3) \rangle}{|x_1|^2 + |x_3|^2} = \frac{|x_1|^2 - |x_3|^2}{|x_1|^2 + |x_3|^2}.$$

This attains its maximum value of 1 at any scalar multiple of $(1, 0, 0)$ and its minimum value of -1 at any scalar multiple of $(0, 0, 1)$. Thus, by the min-max theorem (Theorem 5.3.5), we have

$$\lambda_2 = \min_{U:\dim U=2} \{\max\{R_A(\mathbf{x}) : \mathbf{x} \in U,\ \mathbf{x} \neq 0\}\} \leq \max\{R_A(\mathbf{x}) : \mathbf{x} \in V,\ \mathbf{x} \neq 0\} = 1$$

and

$$\lambda_2 = \max_{U:\dim U=2} \{\min\{R_A(\mathbf{x}) : \mathbf{x} \in U,\ \mathbf{x} \neq 0\}\} \geq \min\{R_A(\mathbf{x}) : \mathbf{x} \in V,\ \mathbf{x} \neq 0\} = -1.$$

So we know that $-1 \leq \lambda_2 \leq 1$. In fact, $\lambda_2 \approx 0.69385$.

Example 5.3.7. Consider the *non-hermitian* matrix

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

The only eigenvalue of A is zero. However, for $\mathbf{x} \in \mathbb{R}^2$, the Rayleigh quotient

$$R_A(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = \frac{\langle (x_1, x_2), (x_2, 0) \rangle}{\langle (x_1, x_2), (x_1, x_2) \rangle} = \frac{x_1 x_2}{x_1^2 + x_2^2}$$

has maximum value $\frac{1}{2}$, when $\mathbf{x}$ is a scalar multiple of $(1, 1)$. So it is crucial that A be hermitian in the theorems of this section.

Exercises.

5.3.1. Prove that $\mathbf{x}_0$ maximizes $R_A(\mathbf{x})$ subject to $\mathbf{x} \neq 0$ and yields the maximum value $M = R_A(\mathbf{x}_0)$ if and only if $\mathbf{x}_1 = \mathbf{x}_0/\|\mathbf{x}_0\|$ maximizes $\langle \mathbf{x}, A\mathbf{x}\rangle$ subject to $\|\mathbf{x}\| = 1$ and yields $\langle \mathbf{x}_1, A\mathbf{x}_1\rangle = M$. *Hint*: The proof is similar to that of Proposition 2.3.2.

5.3.2 ([ND77, Ex. 10.4.1]). Use the Rayleigh quotient to find lower bounds for the largest eigenvalue and upper bounds for the smallest eigenvalue of

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

5.3.3 ([ND77, Ex. 10.4.2]). An eigenvector associated with the lowest eigenvalue of the matrix below has the form $\mathbf{x}_a = (1, a, 1)$. Find the exact value of a by defining the function $f(a) = R_A(\mathbf{x}_a)$ and using calculus to minimize $f(a)$. What is the lowest eigenvalue of A?

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

5.3.4 ([ND77, Ex. 10.4.5]). For each matrix A below, use R_A to obtain lower bounds on the greatest eigenvalue and upper bounds on the least eigenvalue.

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

(b) $\begin{bmatrix} 7 & -16 & -8 \\ -16 & 7 & 8 \\ -8 & 8 & -5 \end{bmatrix}$

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

5.3.5 ([ND77, Ex. 10.4.6]). Using $\mathbf{v}_3 = (1, -1, -1)$ as an eigenvector associated with the largest eigenvalue λ_3 of the matrix A of Exercise 5.3.2, use R_A to obtain lower bounds on the second largest eigenvalue λ_2.

5.3.6 ([ND77, Ex. 10.5.3, 10.5.4]). Consider the matrix

$$A = \begin{bmatrix} 0.4 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.2 \\ 0.1 & 0.2 & 0.3 \end{bmatrix}.$$

(a) Use $U = \text{Span}\{(1, 1, 1)\}^\perp$ in the min-max theorem (Theorem 5.3.5) to obtain an upper bound on the second largest eigenvalue of A.

(b) Repeat with $U = \text{Span}\{(1, 2, 3)\}^\perp$.

Index

1-norm, 38
2-norm, 38
∞-norm, 39

A^+, 24
absolute error, 46
absolute value, 37
algebraic multiplicity, 53

back substitution, 26
block form, 10
block triangulation, 98

$\mathbb{C}$, 5
Cauchy–Schwarz inequality, 44
characteristic polynomial, 53
Cholesky factorization, 68
coefficient matrix, 15
column space, 12
commute, 10
complex conjugate, 37
condition number, 44
conjugate transpose, 56

$\operatorname{diag}(a_1, \ldots, a_n)$, 6
diagonalizable, 53
diagonalization, 53
diagonalized Rayleigh quotient, 112
Dimension Theoreom, 12
dominant eigenvalue, 77
dominant eigenvector, 77
dot product, 8, 48

Eckart–Young–Mirsky Theorem, 94
eigenspace, 53
eigenvalue, 53
eigenvector, 53
elementary matrix, 13
elementary permutation matrix, 32
elementary row operations, 13
Euclidean norm, 38

$\mathbb{F}$, 5
$\mathbb{F}^\times$, 5
forward substitution, 26
free variable, 16
Frobenius norm, 42
fundamental subspaces, 91

Gauss–Jordan elimination, 15
gaussian algorithm, 14
Gaussian elimination, 13
generalized eigenspace, 99
generalized inverse, 24
geometric multiplicity, 53
Gershgorin circle theorem, 81
Gershgorin discs, 81
$\mathrm{GL}(n, \mathbb{F})$, 5
Gram matrix, 23, 86
Gram–Schmidt othogonalization algorithm, 49

hermitian conjugate, 56
hermitian matrix, 57

idempotent, 98
identity matrix, 9
ill-conditioned, 36, 46
initial value problem, 104
inner product, 48
inverse, 22
inverse power method, 79
invertible, 22

Jordan block, 100

kernel, 12

$\mathcal{L}(V, W)$, 11
leading 1, 14

leading variable, 16
left inverse, 17
left-invertible, 17
level set, 108
linear transformation, 11
lower reduced, 27
lower triangular, 26
lower unitriangular, 70
LU decomposition, 27
LU factorization, 27

$M_{m,n}(\mathbb{F})$, 5
magnitude, 37
main diagonal, 25
Markov chains, 79
matrix
    addition, 6
    arithmetic, 6
    difference, 7
    multiplication, 8
    negative, 6
    scalar multiple, 7
    size, 5
    square, 7
    subtraction, 7
    symmetric, 7
matrix exponential, 102
matrix norm, 40
matrix product, 8
matrix-vector product, 8
maximum norm, 39
$M_i(a)$, 13
middle inverse, 94
Moore–Penrose inverse, 24, 95

$\mathbb{N}$, 5
negative definite, 66
negative semi-definite, 66
nilpotent, 100
non-leading variable, 16
nonsingular, 22
norm, 37
normal equations, 76
normal matrix, 63
normed vector space, 37
null space, 12
nullity, 12

operator norm, 40
orthogonal, 49
    basis, 49
orthogonal complement, 51
orthogonal matrix, 58
orthogonal projection, 51
orthogonally diagonalizable, 62
orthonormal, 49
    basis, 49
overdetermined, 15

p-norm, 39
$PA = LU$ factorization, 33
permutation matrix, 32
$P_{i,j}$, 13
PLU factorization, 33
polarization identity, 107
positive definite, 66
positive semi-definite, 66
power method, 77
Principal axes theorem, 62
principal submatrices, 67
principle axes, 62
principle components, 94
pseudoinverse, 24, 95

QR factorization, 71
QR method, 80
quadratic form, 106, 108

$\mathbb{R}$, 5
$\mathbb{R}_{>0}$, 65
rank, 12, 15
Rank-Nullity Theorem, 12
Rayleigh quotient, 112
reduced QR factorization, 72
reduced row-echelon form, 14
reduced row-echelon matrix, 14
regular stochastic matrix, 79
relative error, 46
right inverse, 20
right-invertible, 20
row space, 12

row-echelon form, 14
row-echelon matrix, 14

scalar, 5
scalar multiplication, 7
Schur decomposition, 61
Schur's theorem, 60
signature, 111
similar matrices, 53
singular, 22
singular matrix, 87
singular value decomposition, 85
singular values, 87
size of a matrix, 5
spectral theorem, 61
spectrum, 62
square
    linear system, 15
    matrix, 7
standard basis vector, 6
steady state vector, 79
stochastic matrix, 79
SVD, 85
Sylvester's law of inertia, 111
symmetic matrix, 7
symmetric, 57

tall matrix, 15
thin QR factorization, 72
trace, 61
transpose, 7
triangular, 26
triangule inequality, 37
two-sided inverse, 22

underdetermined, 15
unitarily diagonalizable, 60
unitary matrix, 58
unknown, 15
upper triangular, 25

variable, 15
vector, 6
vector norm, 37
vector space, 7

well-conditioned, 46
wide matrix, 15

$\mathbb{Z}$, 5
zero matrix, 7

Bibliography

[AK08] Grégoire Allaire and Sidi Mahmoud Kaber. *Numerical linear algebra*, volume 55 of *Texts in Applied Mathematics*. Springer, New York, 2008. Translated from the 2002 French original by Karim Trabelsi. URL: https://doi.org/10.1007/978-0-387-68918-0.

[BV18] S. Boyd and L. Vandenberghe. *Introduction to Linear Algebra – Vectors, Matrices, and Least Squares*. Cambridge University Press, Cambridge, 2018. URL: http://vmls-book.stanford.edu/.

[Mey00] Carl Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. With 1 CD-ROM (Windows, Macintosh and UNIX) and a solutions manual (iv+171 pp.). URL: https://doi.org/10.1137/1.9780898719512.

[ND77] Ben Noble and James W. Daniel. *Applied linear algebra*. Prentice-Hall, Inc., Englewood Cliffs, N.J., second edition, 1977.

[Nic] W. K. Nicholson. *Linear Algebra With Applications*. URL: https://lyryx.com/products/mathematics/linear-algebra-applications/.

[Pen55] R. Penrose. A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.*, 51:406–413, 1955.

[Tre] S. Treil. *Linear Algebra Done Wrong*. URL: http://www.math.brown.edu/~treil/papers/LADW/LADW.html.

[Usm87] Riaz A. Usmani. *Applied linear algebra*, volume 105 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1987.